Body Appreciation Scale (BAS-2): Measurement Invariance across Genders and Item Response Theory Examination.

Daniel Zarate
Victoria University
Joshua Marmara (joshua.marmara@live.vu.edu.au)
Victoria University
Camilla Potoczny
Victoria University
Warwick Hosking
Victoria University
Vasielios Stavropoulos
Victoria University

Research Article

Keywords: Body Appreciation, Measurement Invariance, Item Response Theory, Psychometric Properties, Positive Psychology, Gender

Posted Date: March 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-265909/v1

Version of Record: A version of this preprint was published at BMC Psychology on July 30th, 2021. See the published version at https://doi.org/10.1186/s40359-021-00609-3.

Abstract

Background: The present study considers a measure of positive body image, the Body Appreciation Scale-2, which assesses acceptance and/or favourable opinions towards the body (BAS-2[29]). Differential functioning of the scale across the two genders, as well as its items, has not been excluded. The present study contributes to this area of knowledge via the employment of gender Measurement Invariance (MI) and Item Response Theory (IRT) analyses.

Methods: A group of 386 adults from the community were assessed (N = 394, 54.8% men, 43.1% women, M_{age} = 27.48; SD = 5.57).

Results: MI analysis observed invariance across males and females at the configural level, and non-invariance at the metric level. Further, the two-parameter logistic model employed to observe IRT properties indicated that all items demonstrated, although variable, strong discrimination capacity.

Conclusions: The items showed increased reliability for latent levels of ∓ 2 SD from the mean level of Body Appreciation. The implications and interpretations of the findings for clinical practice are discussed.

Introduction

Body image is a multidimensional construct that represents one's cognitions, behaviours, perceptions and affective responses towards their body [7]. Contemporary literature has predominantly focused on negative body image and its relationship with poor mental health [2, 8]. The studies appear to focus on a uni-dimensional component of body image by emphasizing a negative connotation, primarily related to mental health treatment seekers [27]. Such conceptual biases have been challenged by literature suggesting placing emphasis on the whole spectrum of body image variations, ranging from negative to positive [5, 6]. In this context, one's body appreciation (BA) is linked with one's positive body image. BA is depicted as "accepting and holding favourable opinions towards the body, while rejecting mainstream ideals of stereotypical human beauty" [29]. To measure BA, Avalos et al. (2005) pioneered the body appreciation scale (BAS). The use of the BAS has demonstrated positive ties between BA and one's psychological well-being (e.g., self-esteem, optimism, positive affect; [1, 23], and negative links with body surveillance, body shame, and body dissatisfaction. These findings underpinned the introduction of an upgraded tool––the Body Appreciation Scale 2 (BAS-2, [29]) to eliminate sex-specific and body dissatisfaction-based language. Interestingly, this 10-item scale was originally devised to measure BA exclusively among women and was later modified to include men [29]. Thus, one could assume that the scale may operate differently between male and female respondents. This would mean that any BAS-based comparisons between the two genders could be confounded by the psychometric properties of the scale. This possibility highlights the importance of establishing the psychometric properties of the BAS-2 across traditional binary forms of gender (men and women). To address this aim, one would need to utilise Measurement Invariance (MI[18]).

Measurement Invariance (MI)

Establishing MI across observed groups, such as genders, is of paramount importance to claim significance over inferred comparative observations [21]. MI can be considered as a test of heterogeneity, that evaluates whether the measurement properties of a construct remain stable across groups, thus securing meaningful comparisons. A comprehensive method to evaluate whether MI exists across groups is Multigroup Confirmatory Factor Analysis

(MCFA, [12]). This method involves evaluating whether significant differences in variance across groups exist at different/successive levels of the construct. These entail configural (i.e., factorial structure), metric (i.e., factor loadings), scalar (i.e., intercepts or thresholds) and strict (i.e., residuals) MI. In this case, confirming configural invariance would imply that the number of factors and pattern of item-factor loadings within the BAS-2 are similar for men and women. In that line, achieving support for BAS-2 metric invariance would suggest that the item-factor loading relationship is being measured with the same metric scale for both groups. Last, achieving support for BAS-2 scalar invariance would suggest that item intercept values are equal across groups. Thus, males and females would be expected to rate each item similarly when experiencing the same level of BA. It is noted that testing for equality of error/residual variance across the groups, as an additional layer of invariance, is often disregarded. Due to residuals being expected to be random, testing their intergroup equality would result in excessively stringent, and thus likely unnecessary and un-informative models [3].

Indeed, Avalos and colleagues' (2005) invitation for further investigation of the BAS/BAS-2 [29] equivalence of psychometric properties across the two genders has been evaluated via USA [28], Spanish [27], Polish [19], French [14], Danish, Portuguese, Swedish [15] and Chinese samples [24]. These studies concluded that gender MI was consistently achieved at the configural and metric levels, and usually (although not always) achieved at the scalar level (with Chinese and Danish samples observing non-invariance; [24, 15]). However, the applied criteria oscillated between a 'more relaxed approach' guided by difference between models in CFI and RMSEA values (as suggested by [9]) and a 'more stringent approach' guided by difference in χ^2 values [3]. Evaluating MI with a 'more relaxed approach' resulted in support for invariance (at all three levels) in the Polish sample and support for partial scalar invariance in the Portuguese, Danish and Swedish samples [3, 15, 19]. Evaluating MI with a 'more stringent approach' resulted in support for invariance (at all three levels) in the Chinese, French and Spanish sample and support for partial scalar invariance for the USA sample, after freeing items 1 and 9 to vary across gender groups (with original BAS comprising 13 items; [3, 14, 24]). Nonetheless, correlated residuals were required in the French (6-10, 6-7, and 1-5 items) and Spanish (1-5, 2-9 and 8-10 items) samples to achieve acceptable fit indices and proceed with the analysis [14, 24, 25]. Taken together, these indicate that items may not only operate differently across genders, but also across different national samples. Given the importance of the construct, as well as its wide use in both clinical and community populations, further thorough examination of the BAS-2 psychometric properties is imperative.

Item Response Theory (IRT)

Item Response Theory (IRT) projects as a superior way of assessing the psychometric properties of a scale at the item level. IRT is assumed to outperform Classical Test Theory's (CTT) psychometric estimation, such as the BAS/BAS-2 MI analyses previously implemented, in a twofold manner [11]. First, while CTT explains relationships between latent variables and items, IRT aims to explain how both the latent trait and item properties are related to individual responses to an item. Second, IRT can generate reliability indices and standard errors for each item rather than just an overall reliability index. It does so while accounting for assessment at different levels of the trait and specific item characteristics. Thus, while CFA (CTT) aims to explain relationships between BAS-2 items and BA (as the latent construct), IRT models evaluate distinct relationships between items and participants' responses to those items, taking into consideration the level of the latent trait that these participants present with [3]. The item-participant relationship is represented by the probability that participants with a certain level of the latent trait (in this case BA) will endorse a particular item [11]. This is graphically represented by the item response function (IRF, [11]). IRF is expressed via a nonlinear (logit) regression line, the location of which on the axis is

determined by item difficulty (β) and discrimination (α) parameters. Difficulty (β) indicates the level of the latent trait where there is a .5 probability that a participant will endorse a specific criterion or item. For instance, 'easier' items have lower β values and their IRF is represented closer to the horizontal axis. For clarification purposes, those that endorse the easier items are said to have lower body appreciation. Conversely, those who endorse the difficult items are said to have higher body appreciation [11]. Discrimination (α) describes how steeply the rate of success (positive response) of an individual varies according to their level of experience of the latent trait. Thus, items more strongly related to the latent variable present steeper IRF functions. IRT models vary according to the number of parameters with a one-parameter logistic model (1PL, including β), two parameter-logistic model (2PL, including β and α) and three parameter-logistic model (3PL) including additionally pseudo-guessing (i.e., lower horizontal asymptote at which even individuals with low levels of the trait behaviour will have a high rate of success, [11]). To maximise information obtained employing IRT, and considering the BAS was measured utilising a 5-point Likert scale, three polytomous (graded, generalised partial and nominal model) IRT model fit were applied.

The Present Study

Prompted by the above literature, the present study aims to contribute to the available knowledge related to the psychometric properties of the BAS-2 in two significant ways: a) it aims to expand gender MI findings via the use of a different national sample and the employment of more stringent research methods; and b) it will be the first to examine the differential item functioning of the BAS-2 items for participants with different levels of BA. Such knowledge is significant in at least three important ways: a) it will add clarity considering the gender comparability of the BAS-2 scores in both research and clinical practice; b) it will allow ranking of the BAS-2 items based on their psychometric performance (i.e., item priority ranking); and c) it will inform how the different BAS-2 items may provide reliable and/or less reliable information among participants with higher and lower levels of BA.

Methods

Participants

After receiving ethics approval from the Victoria University Ethics Committee, participants were recruited online via a crowd sourcing platform called Prolific.co and were awarded $2.50 for their time each. As part of a larger study, 394 participants completed an online survey including the BAS-2. These included 216 men, 170 women, and 8 participants identified as non-binary. These eight participants were excluded in the present analyses targeting binary gender differences. The remaining participants' (N = 386) age ranged from 18 to 39 years (M = 27.54, SD = 5.58). Only the 386 full responses were utilised for statistical analyses resulting in a maximum random sampling error of .089 for a 95% confidence interval and .117 for a 99% confidence interval. Most participants worked full-time (44.3%), had an undergraduate degree (40.4%), were heterosexual (80.5%), were from the outer metropolitan suburbs (41.7%), reported Caucasian ethnicity (57.8%) and lived in the United States of America (USA; 54.9%).

Measures

The 10-item BAS-2 [29] uses a 5-point Likert scale with responses ranging from 1 (*Never*) to 5 (*Always*). Higher scores indicate higher BA. To calculate one's final BA score, item responses are summed, resulting in a score between 5 and 50. Table 1 presents a description of the items and descriptive statistics for the current sample.

The internal consistency of the BAS-2 in the present study was excellent (Cronbach's α = .954, McDonald's ω = .956). Considering past assessed psychometric properties of the scale, Tylka and Wood-Barcalow (2015) found a unidimensional factor structure, along with strong internal consistency (Cronbach's α = .97), construct validity and test-retest reliability (r = .90) in community and college samples of men and women.

Statistical analysis

To address the outlined aims, a series of statistical processes were employed: (i) a sequential multigroup CFA to observe MI across men and women; and (ii) psychometric examination at the scale and item level via IRT.

First, following in the methodology applied in [22], we conducted a multigroup CFA to test for Measurement Invariance across gender groups (males and females). This process involves a stepwise model comparison with progressively restrictive parameters to test for ill-fitting models and subsequently observe sources of non-invariance [3]. More specifically, a configural model (factor loadings and intercepts free to vary) is compared with a metric model (factor loadings constrained to be equal across groups and intercepts free), and a scalar model (equal factor loadings and intercepts), respectively. Considering the stringent nature of χ^2 comparisons between the configural, the metric and the scalar levels ($\Delta \chi^2 < .05$), we also evaluated incremental fit differences in CFI and RMSEA values (ΔCFI > .010, ΔRMSEA > .015, [12, 18]). It should be noted that if full measurement invariance is not attained when comparing models, partial invariance can be explored to determine the source of non-invariance by sequentially freeing constrained parameters across different items, until non-significant difference across models is achieved [21].

A combination of statistical processes was applied to determine the source of non-invariance across the different levels. The Satorra-Bentler test of scaled χ^2 difference for factor loadings and intercepts was employed, as it has been identified as an appropriate test to obtain significant differences between nested and comparison models [20]. Subsequently, modification indices were calculated with *RStudio* for all contemplated parameters at the metric and scalar levels (factor loadings, λ; and item intercept, α). Finally, we applied the Benjamini-Hochberg (BH) procedure for controlling false discovery rate in multiple comparisons. The BH procedure has demonstrated superior power of detection when compared with other correction methods (Bonferroni, Hommel, Hochberg; [26]).

Third, BAS psychometric properties were assessed within the IRT context applying a sequence of polytomous models. IRT assumptions of uni-dimensionality and local independence were met given that all BAS items loaded on a single factor and all item residual correlations were below the accepted threshold (< 0.1, [10]). IRT models employed included the unidimensional graded, generalised partial credit and nominal models [4]. The graded response model deals with ordered polytomous categories and is the preferred method for assessing questionnaires with Likert scales. The generalised partial credit model (GPCM) estimates partial credit points for correctly endorsing some aspects of the item [16]. The nominal response model addresses responses to items with two or more nominal categories and employs polichoric matrices [17]. Considering the stringent nature that χ^2 values demonstrates with samples over 200 participants, criteria for assessing best fitting model was determined by (i) the loglikelihood index of fit [10], (ii) RMSEA < .05 as criteria for sufficient fit [13], and (iii) Bayesian and Akaike Information Criterion (BIC and AIC respectively; with smaller values demonstrating a better model fit, [10]). Subsequently, item parameter characteristics were assessed with the Item Characteristic Curve (ICC) and Item Information Function (IIF); while test characteristics were assessed with the Test Information

Function (TIF) and the Test Characteristic Curve (TCC[10]).

Results

Measurement Invariance

First, the BAS unidimensional factorial structure across gender was assessed. Both groups demonstrated acceptable fit according to acceptance criteria for RMSEA, TLI and CFI suggested by Hu and Bentler (1999) (males: χ^2 = 80.044, *df* = 35, *p* < .001, CFI = .972, TLI = .963, RMSEA = .077; females: χ^2 = 59.404, *df* = 35, *p* = .006, CFI = .980, TLI = .975, RMSEA = .064). Unstandardised item loadings for men ranged from 1 to 1.45 (*Figure 1*) and for women ranged from 0.84 to 1.61 (*Figure 2*). Both groups demonstrated good internal reliability coefficients (males Cronbach's α = .955, and McDonald's ω = .957; females α = .943, ω = .946).

Second, MI for men and women scoring on the BAS was conducted. The unidimensional BAS configural model showed acceptable fit for the sample (χ^2 = 161.20, *p* < .001, CFI = .974, TLI = .967, RMSEA = .072) with a statistically significant decrease in absolute fit (Satorra-Bentler scaled Δ χ^2 = 19.38, *p* = .022) and non-significant change in incremental fit (Satorra-Bentler scaled ΔCFI = .004; ΔRMSEA = .001) at the metric level (Table 2). Given the significant decrease in absolute fit between configural and metric models, no meaningful observations could be inferred between metric and scalar model comparison. Therefore, we proceeded to identify non-invariant parameters by evaluating modification indices and utilising the Benjamini-Hochberg procedure. As presented in Table 3, parameters that produced a significant Satorra-Bentler scaled Δ χ^2 were λ 2, λ 8, λ 9, α 1 and α 9. After calculating Benjamini-Hochberg adjusted *p* values it was determined that all 5 parameters presented a "*p* < BH *p*" condition, thus remaining significant for partial invariance purposes. Indeed, free estimation of factor loading 2, 8 and 9, and intercepts 1 and 9 achieved a not significant decrease when compared to the configural model (Satorra-Bentler scaled Δ χ^2 = 10.61, *p* = .056).

Psychometric IRT properties

Comparisons across the graded model (GM), generalised partial credit model (PCM), and the nominal model (NM) were conducted. The GM demonstrated better fit to data ($\chi^{2Loglikelihood}$ = 8111.38; RMSEA = .06; BIC = 8410.20; AIC = 8211.38) when compared to the PCM ($\chi^{2Loglikelihood}$ = 8182.21; BIC = 8481.02; AIC = 8282.21), and the NM ($\chi^{2Loglikelihood}$ = 8134.35; BIC = 8612.46; AIC = 8294.35). Specifically, when discrimination parameters (i.e., α) where constrained to be equal across models, a significant decrease in fit indices was observed ($\chi^{2Loglikelihood}$ = 11414.61; BIC = 11653.66; AIC = 11494.61).

. Discrimination parameters for all ten items fell within the very high range (0 = non discriminative; 0.01–0.34 = very low; 0.35–0.64 = low; 0.65–1.34 = moderate; 1.35–1.69 = high; >1.70 = very high; Baker, 2001) between 1.87 (α item 5) and 5.19 (α item 4). The descending sequence of the items' discrimination power (*α*) is 4, 6, 2, 3, 9, 10, 8, 7, 1 and 5 (see Table 4). Furthermore, the item difficulty parameters (β), demonstrated a considerable level of fluctuations between the different thresholds across the 10 items. Indicatively, for the first threshold the ascending item sequence of difficulty was 6, 10, 9, 8, 2, 4, 3, 1 and 5. Considering the fourth threshold, this alternated to 3, 4, 10, 9, 1, 6, 7, 5, 8 and 2. Nevertheless, the threshold difficulty parameters progressively increased between the first and the last threshold across all items (see Table 4 and Figure 3). Conclusively, IRT analyses indicated that: (i)

while increasing item scores correctly described increasing levels of body appreciation behaviours across all items, the rate of these increases is different across the items, and (ii) different thresholds perform differently across items considering their level of difficulty.

Considering the items' reliability across the different levels of the latent trait, controlling concurrently for the different levels of items' difficulty, meaningful variations were confirmed. Indicatively, the IIF of item 4 provided the highest level of information/reliability in the ranges between 2 and 1 and a half SD below and over the mean and the area around half SD below and over the mean. The IIFs of items 2, 6, 9 and 10 showed better performance in the range between 2 SDs above and below the mean (although with some variability of less than 1 point). Items 1 and 5 showed a rather low and undifferentiated level of reliability in the area between minus 3 SDs below the mean and 2 SDs above the mean with significant drop for behaviours exceeding 2 SDs above the mean. Finally, item 7 showed average reliability for the area between 2 SDs above and below the mean and significant drop for score around 3 SDs higher or below the mean (see Figure 4).

The performance of the scale as whole is visualized by the Test Characteristic Curve (TCC) and the Test Information Function (TIF). The TCC graph illustrates that the trait of BA inclined steeply, as the total score reported increased (from 4 to 49; see Figure 5). Considering the information provided by the scale, improved information (TIF) scores were around -2 SDs below the mean, up to about +2 SDs above the mean (see Figure 5).

These results suggest that the scale (as a whole) provides a sufficient and reliable psychometric measure for assessing individuals with high and low levels of the BA behaviours in the range between 2 SDs below and over the mean. Nevertheless, it may not be an ideal measure for extremely low and high BA in the areas around 3 SDs above and below the mean. The BA behaviour at the levels of 2 SDs below and above the mean trait level correspond with raw scores of 4 and 39 respectively, and based on these, they could be suggested as conditional (before clinical assessment confirmation) diagnostic cut-off points.

Discussion

The present study is the first of this type to combine CTT and IRT procedures to assess BAS-2 psychometric properties at both the scale and the item level for an English-speaking sample. Considering MI, the loadings of items 2, 8 and 9, and the intercepts of items 1 and 9 were shown to be non-invariant across males and females, when strict (χ^2) comparisons were applied. Considering the IRT evaluation, and although all items presented with high discrimination capacity, this fluctuated according to the following descending sequence of items 4, 6, 2, 3, 9, 10, 8, 7, 1 and 5. Similarly, items' difficulty parameters differed across the different item thresholds. Finally, although the scale as a whole seems to perform sufficiently and reliably when examining BA levels that lie $\mp$ 2 SD beyond the mean, it is not ideal for extremely low and high levels of BA that lie $\mp$ 3 SD beyond the mean.

Uni-dimensionality and Measurement Invariance across Genders

In line with past studies, BAS-2 demonstrated an appropriate unidimensional factorial structure with all items loading saliently and significantly on a single latent variable [14, 19, 25, 29]. When dividing the sample into men and women, BAS-2 maintained an appropriate unidimensional factorial structure with all items loading significantly and acceptable model fit indices for both groups. Further, when using a 'relaxed' approach (i.e., changes in CFI and RMSEA, [3]) to establish invariance across gender groups, BAS-2 demonstrated support for invariance at configural, metric and scalar levels. However, when contemplating a more 'stringent' approach

(Satorra-Bentler Scaled $\Delta \chi^2$), non-invariance at the metric and scalar levels were observed. Lack of full MI has been similarly observed in a USA sample [28]. Thus, one could say that although BA is perceived in the same unidimensional way across binary genders, cautious comparisons need to be attempted due to different gender response patterns across the different items.

Specifically, support for partial invariance revealed that the degree of relationship between BA and items 1, 3, 4, 5, 6, 7 and 10 is equivalent for males and females. Nevertheless, the BA relationship with items 2 (I feel good about my body), 8 (My behaviour reveals my positive attitude toward my body) and 9 (I am comfortable in my body) is unequally associated with males and females due to different response styles. Thus, the metric utilised for BA measurement is non-equivalent across the two gender groups and thus comparisons based on the responses of these items need to be avoided, or carefully interpreted.

Further, the observed support for partial invariance suggested that sources of non-invariance across gender groups were also present in item intercepts. While items 2, 3, 4, 5, 6, 7 and 8 were invariant, items 1 (*"I respect my body"*) and 9 (*"I am comfortable in my body"*). demonstrated unequal intercepts between men and women. That is, while women are expected to score higher ratings of item 1 (*"I respect my body"*) at all levels of the latent variable, men are expected to score higher ratings of item 9 *("I am comfortable in my body"*). Interestingly, Tylka (2013) found a similar source of non-invariance for item 1 in a USA sample. This may suggest that males and females who experience the same level of BA may provide unequal responses for this particular item (i.e., gender specific item scaling).

Scale and Item Discrimination, difficulty, and reliability

IRT findings identified variability across BAS-2 items when considering different levels of BA within participants. Considering that IRT principles relate to the identification of most appropriate items for the evaluation of a specific level of a latent trait, items were evaluated and ranked in relation to their discrimination, difficulty, and reliability [11]. The descending order of items' discrimination power was 4, 6, 2, 3, 9, 10, 8, 7, 1 and 5, suggesting that items invoking positive feelings (item 6 "*I feel love for my body*", item 2 "*I feel good about my body*", and item 3 "*I feel that my body …*") and clear statements reflecting dispositional attitude (item 4 "*I take a positive attitude toward my body*") are able to capture body appreciation levels more effectively as the criterion increases in the individual. Further, while the level of difficulty of endorsing an item increased between the first (*never*) and last options (*always*) of the Likert scale, the sequence of item difficulty varied across thresholds. That is, the ascending order of endorsed items between the first (*never*) and second (*seldom*) options of the Likert scale was 6, 10, 9, 8, 2, 4, 3, 1, and 5. However, the ascending order of endorsed items between the fourth ("*often*") and last ("*always*") options of the Likert scale was 3, 4, 10, 9, 1, 6, 7, 5, 8, and 2. This suggests that participants felt more inclined to endorse "*never*" or "*seldom*" loving their body or feeling beautiful than respecting their body or being attentive to their body needs. Alternatively, participants felt more inclined to endorse *"often*" or *"always*" seeing good qualities and taking a positive attitude towards their bodies than feeling good about their bodies. Therefore, it is proposed that items should be interpreted differently when conducting clinical assessment of BA.

Considering the scale (TIF), improved information performance was observed in the range between 2 SDs below and above the mean. However, considerable variation was observed in relation to the level of information precision provided by each criterion. More specifically, findings demonstrated that item 4 (*"I take a positive attitude toward my body"*) provided the highest level of information/reliability between 2 SD below and 1.5 SD above the mean. Items 2 (*"I feel good about my body"*), 6 (*"I feel love for my body"*), 9 (*"I am comfortable in my*

body") and 10 (*"I feel like I am beautiful even if I am different from media images of attractive people"*) provided a considerable amount of information/reliability between 2 SDs below and above the mean. Finally, items 1(*"I respect my body"*) and 5 (*"I am attentive to my body's needs"*) provided a consistently low amount of information/reliability between 3 SDs below and above the mean. However, these two items along with item 3 (*"I feel that my body has at least some good qualities"*) provided the most information between 2 and 3 SDs below the mean. This indicates that the following three-item sequence should be prioritised when attempting to identify participants with significantly low BA: (i) "*I feel that my body has at least some good qualities*", (ii) "*I respect my body*", and (iii) "*I am attentive to my body needs*". Lastly, the Test Characteristic Curve (TCC) demonstrated an appropriate steepness indicating that BAS-2 clearly identifies increments in body appreciation as the overall score increases. This favours BAS-2 as a sufficient psychometric measure for the assessment of individuals with high and low levels of body appreciation. Nonetheless, the instruments performance significantly decreases to differentiate very low (-3 SD) and very high (+3 SD) BA levels.

Conclusion, Limitations And Further Research

Overall, the present findings suggest that BA comparisons across gender based on BAS-2 should be cautiously interpreted due to response pattern differences affecting the metric and the scale properties of the instrument. Furthermore, the instrument may not perform well for clinically low and high BA levels and thus its use should be accompanied by clinical interviews for formal assessment. Last, items differ considering their suitability to discriminate participants with different levels of the latent trait with certain items.

Despite the unique and innovative contribution this study makes to the evaluation of BAS-2 psychometric properties, a number of limitations should be highlighted. The employed sample encompassed adult English speakers from developed countries. That is, findings observed in the current study might lack a wide generalisability of application to samples involving youth, and non-English speakers. In addition, considering that a community sample of healthy adults was employed, reported IRT properties might not accurately reflect those suffering from pathological body dissatisfaction. Future studies may wish to address these limitations to improve assessment procedures informed by BAS-2.

Declarations

Ethical approval and consent to participate: Ethics approval granted by the Victoria University Ethics Committee. The current study only involved adult subjects (+18 years old) and informed consent was obtained in all cases.

Consent for publication: All authors of the manuscript have read and agreed to its content and are accountable for all aspects of the accuracy and integrity of the manuscript in accordance with ICMJE criteria. All methods were carried out in accordance with relevant guidelines and regulations.

Availability of data and materials: All relevant data is available and accompanying this manuscript (i.e., tables, figures, syntax). Utilised data is exclusively owned by the authors is not publicly available due to ethical considerations.

Competing Interests: Dr Vasielios Stavropoulos is an associate editor of BMC. All other authors have not competing interests.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Authors' contributions: DZ contributed to the article's conceptualization, data curation, formal analysis, methodology, project administration, and writing of the original draft. JM contributed to data curation, writing of the original draft, review, editing the final draft and project administration. VS contributed to the article's conceptualization, data curation, formal analysis, methodology, project administration, and writing of the original draft. CP and WH contributed to writing, reviewing, and editing the final draft.

Acknowledgments: The authors would like to thank Dr Stavropoulos for his unconditional support and guidance.

References

[1] Avalos, L.C., Tylka, T.L., & Wood-Barcalow, N. (2005). The Body Appreciation Scale: Development and psychometric evaluation: Scale Development and psychometric evaluation. *Body Image, 2,* 285-297. http://doi.org/10.1016/j.bodyim.2005.06.002

[2] Barnes, M., Abhyankar, P., Dimova, E., & Best, C. (2020). Associations between body dissatisfaction and self-reported anxiety and depression in otherwise healthy men: A systematic review and meta-analysis. *PLoS ONE, 15*(2).

[3] Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications. https://www.guilford.com/books/Confirmatory-Factor-Analysis-for-Applied-Research/Timothy-Brown/9781462515363

[4] Cai, L., Yang, J.S., & Hansen, M. (2011). Generalized Full-Information Item Bifactor Analysis. *Psychological Methods, 16*(3), 221-248. http://doi.org/10.1037/a002333550

[5] Cash, T.F. (2002). Cognitive-behavioural perspectives on body image. In T.F. Cash & Pruzinsky (Eds.), *Body image: A handbook of theory, research, and clinical practice* (pp. 38-46). Guildford Press.

[6] Cash, T. F., & Pruzinsky, T. E. (1990). Body images: Development, deviance, and change. Guilford Press.

[7] Cash, T. F., & Smolak, L. (Eds.). (2011). Body image: A handbook of science, practice, and prevention. Guilford Press.

[8] Chan, C. Y., Lee, A. M., Koh, Y. W., Lam, S. K., Lee, C. P., Leung, K. Y., & Tang, C. S. K. (2020). Associations of body dissatisfaction with anxiety and depression in the pregnancy and postpartum periods: A longitudinal study. Journal of Affective Disorders, 263, 582-592.

[9] Cheung, G.W., & Rensvold, R.B. (2002). Evaluating Goodness of Fit Indexes for Testing Measurement Invariance. *Structural Equation Modelling, 9*(2), 233-255. http://doi.org/10.1207/S15328007SEM0902_5

[10] De Ayala, R.J. (2008). *The Theory and Practice of Item Response Theory.* Guilford Press.

[11] Embretson, S., & Reise, S. (2013). *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates.

[12] Gomez, R., Vance, A., & Stavropoulos, V. (2018b). Test-retest measurement invariance of clinic referred children's ADHD symptoms. *Journal of Psychopathology and Behavioral Assessment, 40*(2), 194-205.

http://doi.org/10.1007/s10862-017-9636-4

[13] Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexed in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling, 6*(1), 1-55. http://doi.org/10.1080/107055199909540118

[14] Kertechian, S., & Swami, V. (2017). An examination of the factor structure and sex invariance of a French translation of the Body Appreciation Scale-2 in university students. *Body Image, 21*, 26-29. http://doi.org/10.1016/j.bodyim.2017.02.005

[15] Lemoine, J.E., Konradsen, H., Lunde Jensen, A., Roland-Levy, C., Ny, P., Khalaf, A., & Torres, S. (2018). Factor structure and psychometric properties of the Body Appreciation Scale-2 among adolescents and young adults in Danish, Portuguese, and Swedish. *Body Image, 26*, 1-9. https://doi.org/10.1016/j.bodyim.2018.04.004

[16] Muraki, E. (1997). A Generalized Partial Credit Model. In W.J. Van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. Springer.

[17] Preston, K., Reise, S., Cai, L., & Hays R.D. (2011). Using the Nominal Response Model to Evaluate Response Category Discrimination in the PROMIS Emotional Distress Item Pools. *Educational and Psychological Measurement, 71*(3), 523-550. http://doi.org/10.1177/0013164410382250

[18] Putnick, D.L., & Bornstein, M.H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Dev Rev, 41*, 71-90. http://doi.org/10.1016/j.dr.2016.06.004

[19] Razmus, M., & Razmus, W. (2017). Evaluating the psychometric properties of the Polish version of the Body Appreciation Scale-2. *Body Image, 23,* 45-49. http://doi.org/10.1016/j.bodyim.2017.07.004

[20] Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75,* 243-248. http://doi.org/10.1007/s11336-009-9135-y

[21] Stavropoulos, V., Bamford, L., Beard, C., Gomez, R., & Griffiths, M.D. (2019). Test-retest Measurement Invariance of the Nine-Item Internet Gaming Disorder Scale in Two Countries: A Preliminary Longitudinal Study. *International Journal of Mental Health and Addiction.* http://doi.org/10.1007/s11469-019-000999-w

[22] Stavropoulos, V., Beard, C., Griffiths, M.D., Burleigh, T., Gomez, R., & Pontes, H.M. (2018). Measurement Invariance of the Internet Gaming Disorder Scale- Short-Form (IGDS9-SF) Between Australia, the USA, and the UK. *International Journal of Mental Health and Addiction, 16*, 377-392. http://doi.org/10.1007/s11469-017-9786-3

[23] Swami, V., Stieger, S., Haubner, T., & Voracek, M. (2008). German Translation and psychometric evaluation of the Body Appreciation Scale. *Body Image, 5*, 122-127. http://doi.org/10.1016/j.bodyim.2007.10.002

[24] Swami, V., Ng, S., & Barron, D. (2016). Translation and psychometric evaluation of a Standard Chinese version of the Body Appreciation Scale-2. *Body Image, 18,* 23-26. http://doi.org/10.1016/j.bodyim.2016.04.005

[25] Swami, V., Alias Garcia, A., & Barron, D. (2017). Factor structure and psychometric properties of a Spanish translation of the Body Appreciation Scale-2 (BAS-2). *Body Image, 22,* 13-17. http://doi.org/10.1016/j.boyim.2017.05.002

[26] Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83. http://doi.org/10.3102/10769986027001077

[27] Tylka, T. L. (2011). Positive psychology perspectives on body image. In Body image: A handbook of science, practice, and prevention, 2nd ed. (pp. 56-64). Guilford Press.

[28] Tylka, T.L. (2013). Evidence for the Body Appreciation Scale's measurement equivalence/invariance between U.S. college women and men. *Body Image, 10*, 415-418. http://doi.org/10.1016/j.bodyim.2013.02.006

[29] Tylka, T.L., & Wood-Barcalow, N.L. (2015). The Body Appreciation Scale-2: Item refinement and psychometric evaluation. *Body Image, 12,* 53-67. http://doi.org/10.106/j.bodyim.2014.09.006

Tables

Table 1.

Descriptive statistics for BAS-2 10 items (N = 386)

	Overall				Men	Women
	M	*SD*	Skewness	Kurtosis	*M*	*M*
1.I respect my body	3.53	.94	-.32	-.19	3.52	3.55
2.I feel good about my body	3.04	.99	-.13	-.38	3.12	2.95
3.I feel that my body has at least some good qualities	3.60	1.02	-.43	-.30	3.65	3.54
4.I take a positive attitude toward my body	3.21	1.06	-.18	-.54	3.30	3.09
5.I am attentive to my body's needs	3.41	.95	-.18	-.33	3.44	3.38
6.I feel love for my body	2.93	1.13	.03	-.72	3.05	2.78
7.I appreciate the different and unique characteristics of my body	3.07	1.12	-.07	-.73	3.14	2.98
8.My behaviour reveals my positive attitude toward my body	3.02	1.09	-.01	-.68	3.13	2.89
9.I am comfortable in my body	3.18	1.13	-.23	-.70	3.33	2.99
10.I feel like I am beautiful even if I am different from media images of attractive people	3.05	1.20	-.06	-.85	3.13	2.95

Note. M = mean; SD = Standard Deviation; Min = Minimum; Max = Maximum

*Note: * = Statistically significant p < .05. Partial invariance achieved by freeing factor loadings 2,8 and 9, and intercept 2 and 9*

Table 2.

Test of invariance BAS questionnaire

	Df	Δ Df	χ^2	Δ χ^2	*p*	CFI	ΔCFI	RMSEA	Δ RMSEA	AIC	BIC
Configural – Model 1 (free loadings, free intercepts)	70		161.20			.974		.072		8249.7	8487.0
Metric – Model 2 (equal loadings, free intercepts)	79	9	177.56	19.38	.022*	.971	.004	.073	.001	8248.0	8449.8
Scalar – Model 3 (equal loadings, equal intercepts)	88	9	193.85	16.18	.063	.968	.003	.072	.001	8246.3	8412.5
Partial Invariance	82	12	172.17	10.61	.056	.975	.001	.071	.001	8236.6	8426.5

Table 3.

Benjamini-Hochberg procedure: testing item intercept and factor loadings for BAS invariance between men and women.

Model	Parameter Relaxed	df	χ^2	*P* value	BH adj *p* value	Sig
M_0		88	193.846			
M_1	λ1	86	193.658	.4120	.0112	
M_2	λ 2	86	190.588	**.0142**	.0212	*****
M_3	λ 3	86	193.663	.8657	.0037	
M_4	λ 4	86	193.845	.9999	.0012	
M_5	λ 5	86	190.325	.0530	.0187	
M_6	λ 6	86	192.848	.2388	.0125	
M_7	λ 7	86	193.700	.8041	.0062	
M_8	λ 8	86	191.137	**.0162**	.0200	*
M_9	λ 9	86	186.770	**.0001**	.0025	*
M_{10}	λ 10	86	192.557	.1378	.0162	
M_{11}	α 1	86	187.529	**.0017**	.0237	*****
M_{12}	α 2	86	193.821	.1051	.0175	
M_{13}	α 3	86	191.871	.1416	.0150	
M_{14}	α 4	86	193.837	.9769	.0025	
M_{15}	α 5	86	192.915	.4328	.0100	
M_{16}	α 6	86	192.061	.1817	.0137	
M_{17}	α 7	86	193.605	.7888	.0075	
M_{18}	α 8	86	193.189	.5242	.0087	
M_{19}	α 9	86	188.323	**.0040**	.0225	*****
M_{20}	α 10	86	193.641	.8427	.0050	

Note. df = degrees of freedom; *BH = adj p value* Benjamini Hochberg adjusted *p* value. *Sig = Significance* is determined by *p* value smaller than BH adj *p* value. Parameter relaxed denotes which parameter has been relaxed in comparison to M_0

Table 4.

BAS 2PL IRT Properties.

Item	α	β_1	β_2	β_3	β_4
1	2.09	-2.73	-1.43	-0.05	1.34
2	3.71	-1.70	-0.61	0.50	1.71
3	3.54	-2.12	-1.22	-0.11	0.89
4	5.19	-1.68	-0.66	0.27	1.23
5	1.87	-2.76	-1.33	0.11	1.54
6	4.08	-1.31	-0.33	0.52	1.41
7	2.90	-1.63	-0.51	0.40	1.41
8	2.96	-1.61	-0.46	0.48	1.54
9	3.04	-1.55	-0.61	0.27	1.32
10	3.01	-1.38	-0.46	0.41	1.27

Note = α defines the capacity of an item to discriminate between varying levels of body appreciation (θ).

β represents the level of latent trait observed to endorse each item at a specific threshold.

Figures

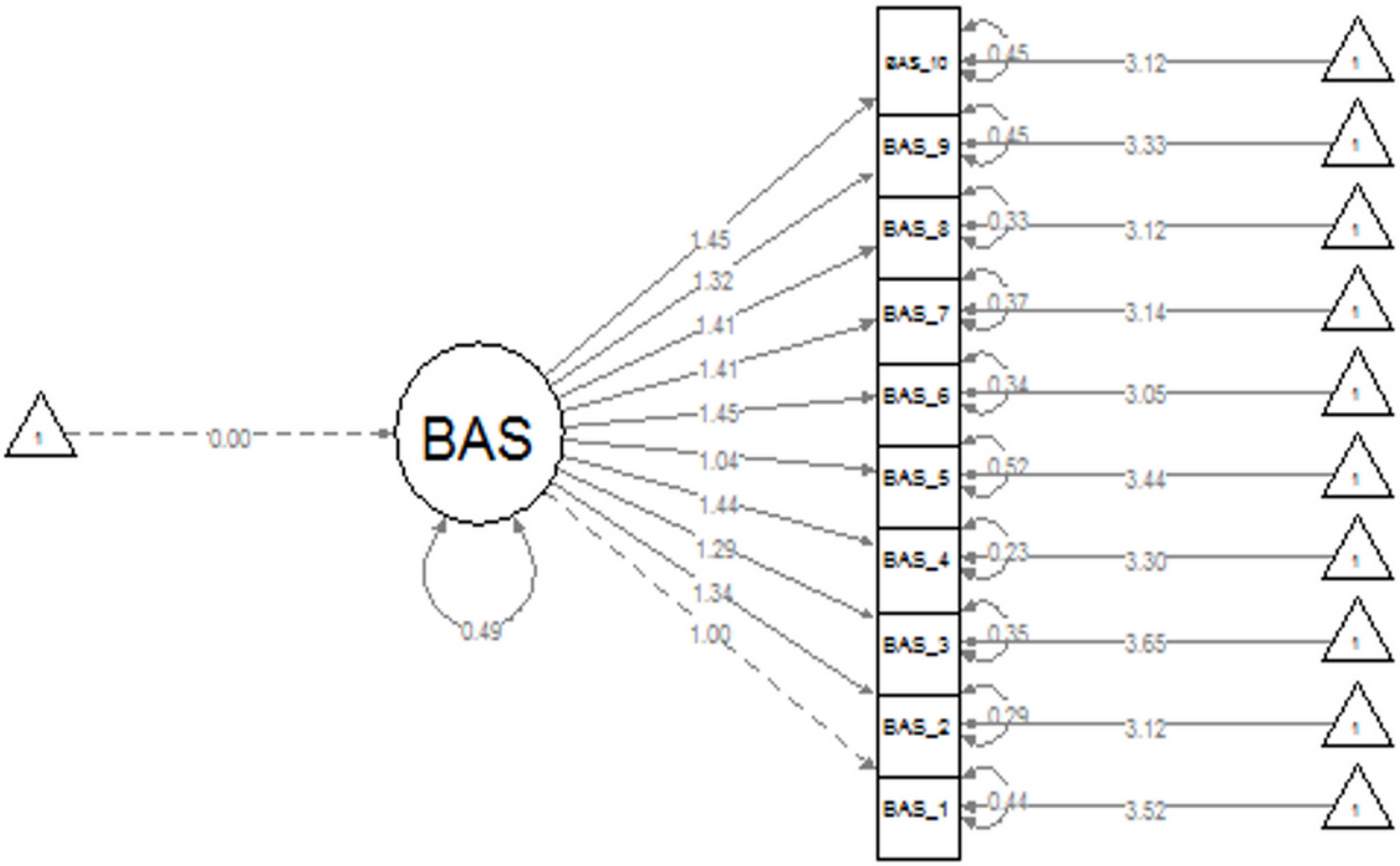

Figure 1

Body appreciation scale unstandardised item loading for Men.

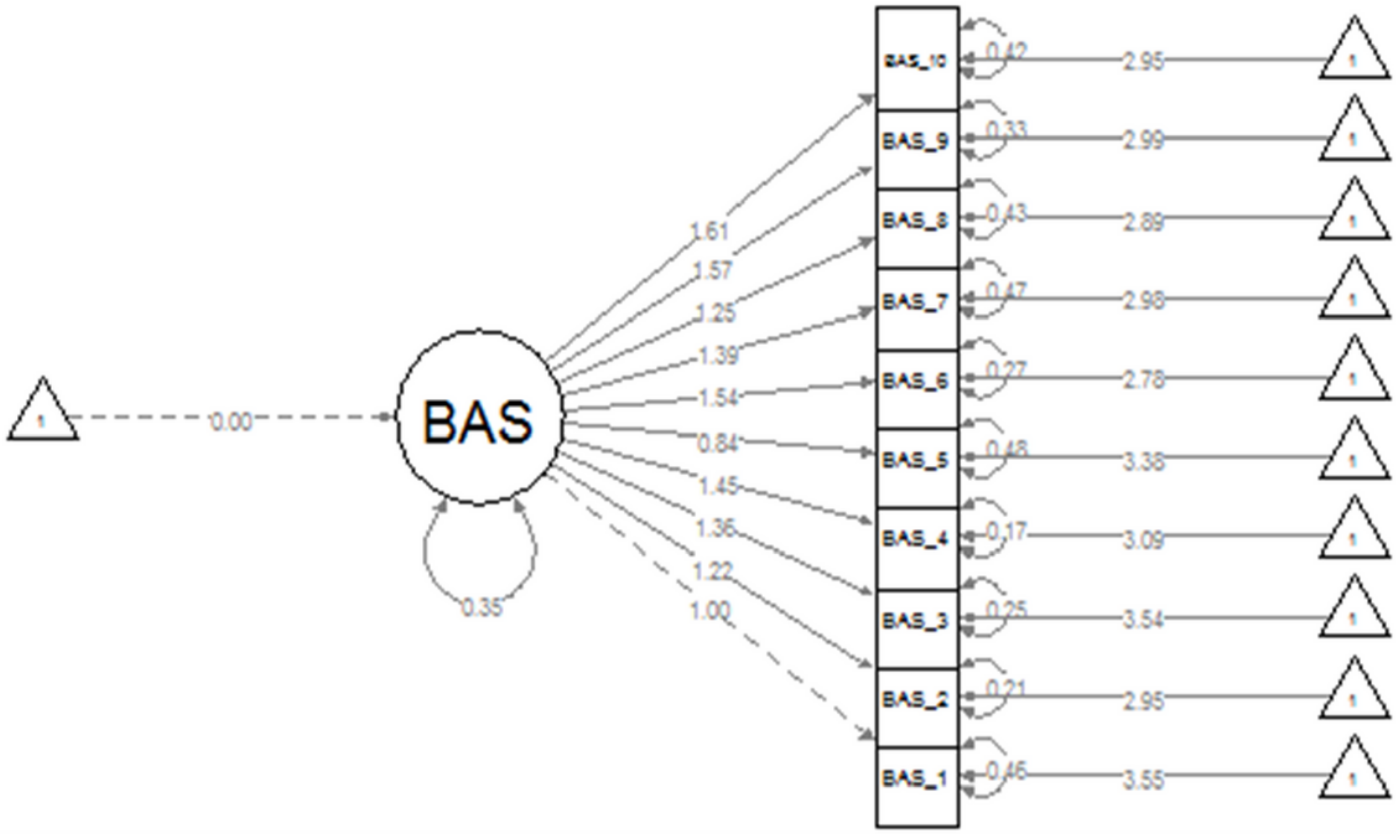

Figure 2

Body appreciation scale unstandardised item loading for Women.

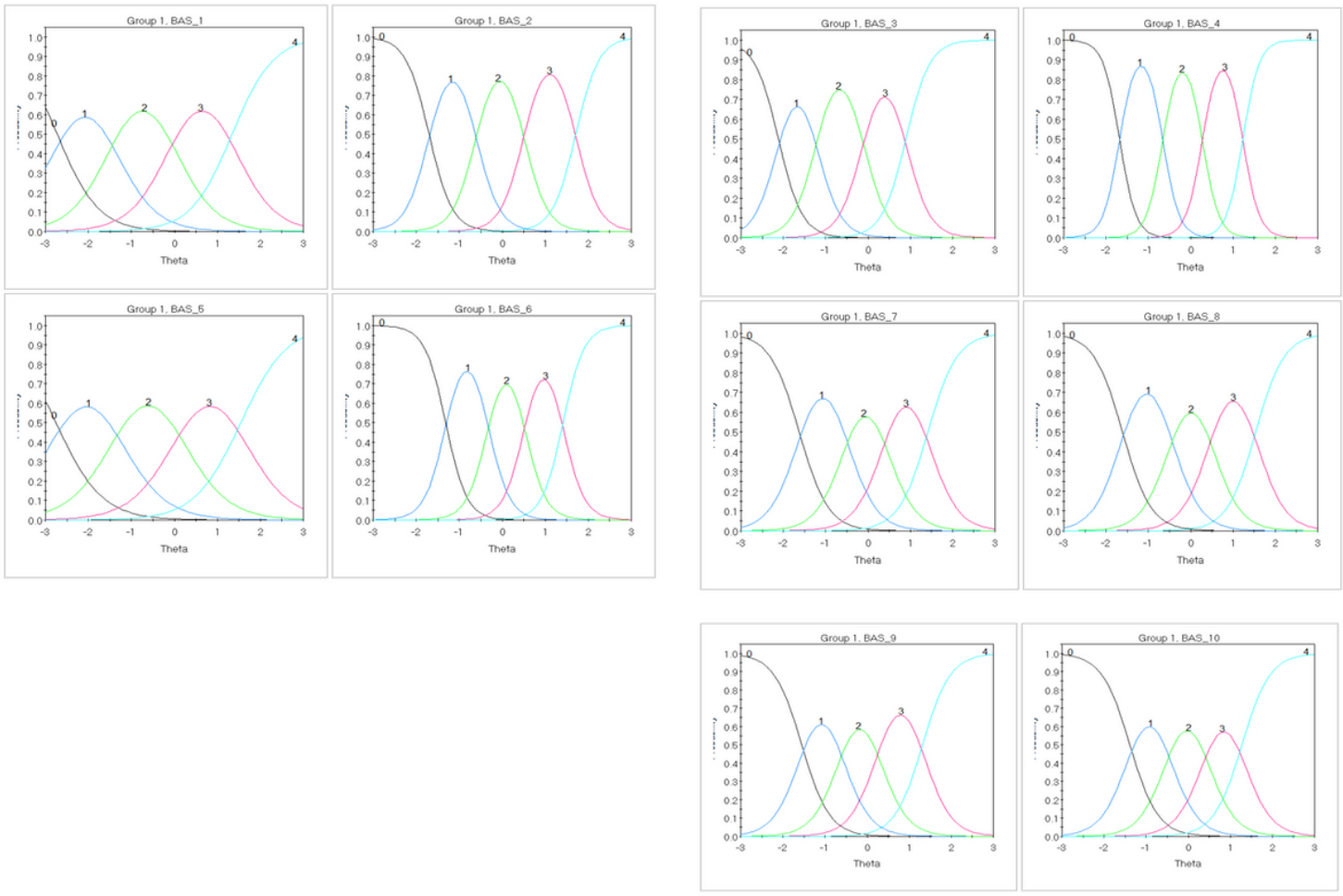

Figure 3

BAS Items' Characteristic Curves (ICC).

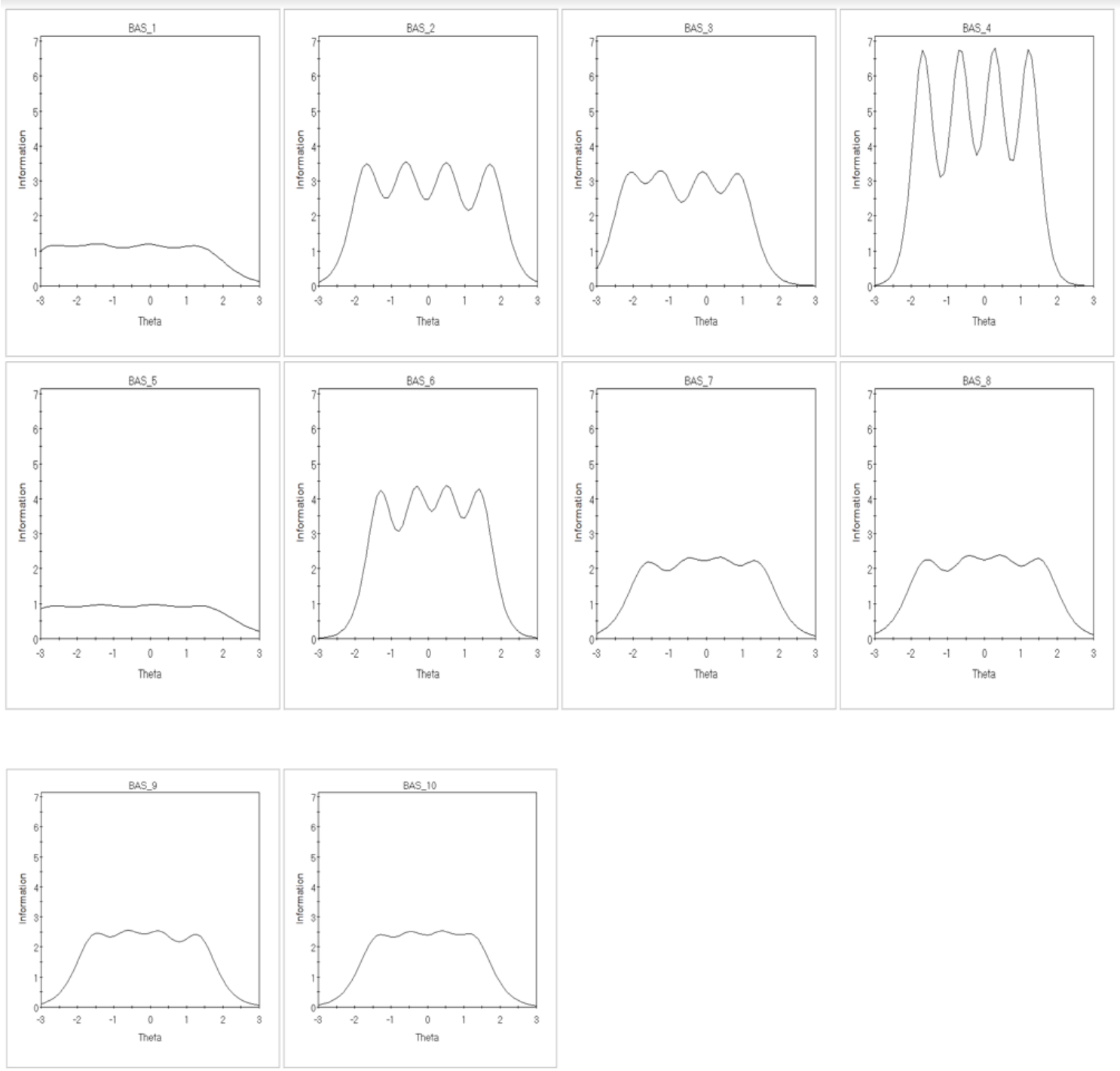

Figure 4

BAS Item Information Function (IIF).

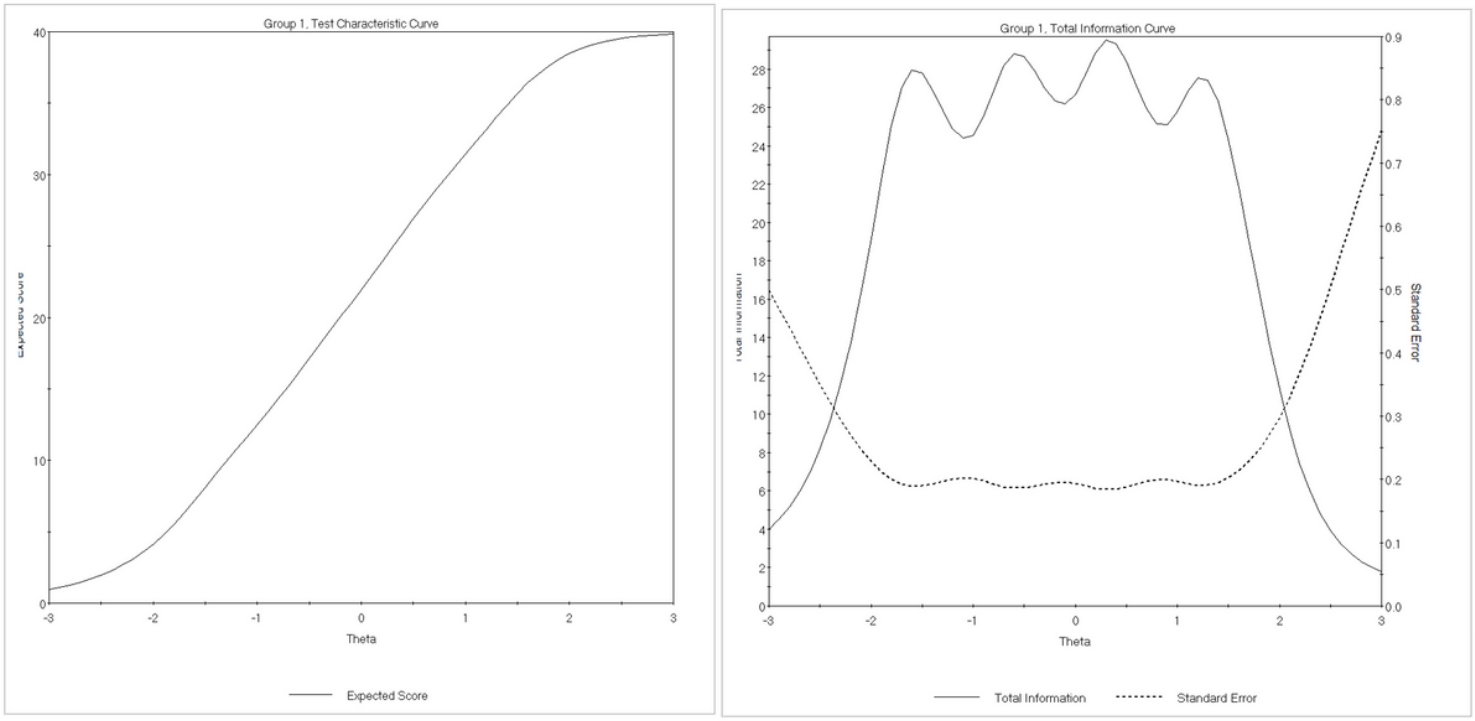

Figure 5

BAS Test Characteristic Curve (TCC) and Test Information Function (TIF).