Original Article

Gender Differences or Gender Bias?

Examination of the Assessment of Sadistic Personality Using Item Response Theory and Differential Item Functioning

Rachel A. Plouffe[1,2], Christopher Marcin Kowalski[3], Paul F. Tremblay[3], Donald H. Saklofske[3], Radosław Rogoza[4], Rossella Di Pierro[5], and Saad Chahine[6]

[1]Department of Psychiatry, Schulich School of Medicine & Dentistry, University of Western Ontario, London, ON, Canada
[2]The MacDonald Franklin OSI Research Centre, Lawson Health Research Institute, London, ON, Canada
[3]Department of Psychology, University of Western Ontario, London, ON, Canada
[4]Department of Psychology, Cardinal Stefan Wyszyński University, Warsaw, Poland
[5]Department of Psychology, University of Milano-Bicocca, Milan, Italy
[6]Faculty of Education, Queen's University, Kingston, ON, Canada

Abstract: Sadism, defined by the infliction of pain and suffering on others for pleasure or subjugation, has recently garnered substantial attention in the psychological research literature. The Assessment of Sadistic Personality (ASP) was developed to measure levels of everyday sadism and has been shown to possess excellent reliability and validity using classical test theory methods. However, it is not known how well ASP items discriminate between respondents of different trait levels, or which Likert categories are endorsed by persons of various trait levels. Additionally, individual items should be evaluated to ensure that men and women of similar levels of sadism have an equal probability of response endorsement. The purpose of this research was to apply item response theory (IRT) and differential item functioning (DIF) to investigate item properties of the ASP across its three translations: English, Polish, and Italian. Overall, the results of the IRT analysis showed that with the exception of Item 9, the ASP demonstrated sound item properties. The DIF rate analyses identified two items from each questionnaire that were of practical significance across gender. Implications of these results are discussed.

Keywords: sadism, item response theory, differential item functioning, personality, assessment

The psychological study of human evil has gained immense traction among researchers, practitioners, and laymen alike. The widespread interest in understanding human malevolence is not surprising considering the genocidal, brutal, and murderous events that have occurred throughout human history. A catalyst for this recent development in psychological research has been the introduction of subclinical sadism to the personality literature, defined by the infliction of pain and suffering on others for pleasure or subjugation (Chabrol et al., 2009; Plouffe et al., 2017).

Since sadism's introduction to the personality research literature, several measures have been developed to assess individual levels of the construct, including the Assessment of Sadistic Personality (ASP; Plouffe et al., 2017), the Comprehensive Assessment of Sadistic Tendencies (CAST; Buckels & Paulhus, 2014), the Varieties of Sadistic Tendencies (VAST; Paulhus & Jones, 2015), and the Short Sadistic Impulse Scale (SSIS; O'Meara et al., 2011). The SSIS evaluates the hurting nature of the sadistic individual, whereas the CAST and VAST demonstrate broad content coverage, reflecting verbal, physical, and vicarious elements of sadism. The concise 9-item ASP evaluates the sadist's subjugation, pleasure-seeking through antagonistic behaviors, and lack of empathy. Overall, these scales have been shown to possess excellent reliability and validity based on classical test theory (CTT) methods (e.g., O'Meara et al., 2011; Plouffe et al., 2019). A recent study conducted by Kowalski et al. (2020) assessed the validity of translated versions of the ASP. Overall, they found evidence for convergent and discriminant validity. Configural and partial metric invariance were also satisfied in this study, and following the implementation of alignment optimization, latent mean differences could be calculated between countries.

Despite evidence for construct validity and reliability across translations of the ASP, it is not known how well the items discriminate between respondents of different trait levels, or which Likert categories are endorsed by individuals varying in levels of sadism across its translations. When assessing antagonistic traits, despite having

https://doi.org/10.1027/1015-5759/a000634

continuous 5-point response scales, individuals frequently endorse the lowest response categories (e.g., Persson et al., 2017). While this may reflect either characteristic of the samples surveyed or the actual frequency of these traits, supplementing CTT with further item-level analyses across translations is both important and required. Furthermore, individual items should be evaluated to ensure that subgroups (i.e., men and women) who have equal levels on ASP sadism also have an equal probability of response endorsement.

The overarching purpose of this research is to apply item response theory (IRT) and differential item functioning (DIF) to examine item properties of the ASP across gender and its three translations: English, Polish, and Italian. This research will have important implications across research, clinical, and vocational domains, as trait measures of sadism are frequently used to predict important criterion variables including, for example, workplace deviance (Min et al., 2019), intimate partner violence (Plouffe et al., 2020), and overall aggressive behavior (Chester et al., 2019). Importantly, to ensure the accuracy of trait measurement and relevant predictions, researchers must employ item analysis procedures designed to maximize the reliability and validity of measures across genders and nations.

Item Response Theory: Discrimination and Trait Thresholds in the Assessment of Sadistic Personality

Item response theory is a modeling technique employed to test the relationships between latent variables and their manifestations, or item responses (DeMars, 2010; Embretson & Reise, 2000). There are several advantages to implementing IRT procedures over CTT. For example, item parameters are not sample specific, and measurement error (and reliability) is not constant across trait levels. Therefore, implementing an IRT framework will allow for a more accurate cross-national evaluation of the ASP.

For this research, the Graded Response Model (GRM; Samejima, 1969, 1996) was used to model these associations for the ASP's polytomous (Likert scale) items. The model demonstrates response probability as a function of latent subclinical sadism (θ). The GRM tests two types of parameters: four thresholds (b_1–b_4) and item discrimination (*a*). The *b* parameters reflect the thresholds at which participants have a 50% probability of selecting the next category or higher, and are measured on the same scale as θ. The *a*-parameter indicates the degree to which items differentiate between participants with varying levels of θ. Together, these parameters can be used to plot item category characteristic curves (CCCs) to assess the probability of response endorsement in a given category, and a test information function (TIF) to indicate the amount of information the ASP provides as a function of θ. More detailed descriptions of IRT are described by Embretson and Reise (2000) and Hambleton and Swaminathan (2013).

Only one study to date has assessed the ASP using IRT methodology (Dinić et al., 2020). Results of the IRT analysis showed that ASP precision was higher when individuals' levels of latent sadism were also high. However, this study included only a Serbian translation of the ASP and did not evaluate item discrimination or threshold parameters. Therefore, it is important to test whether these results hold across additional ASP translations and to extend findings by Dinić et al. (2020) by including additional parameters.

Differential Item Functioning: Measurement Equivalence in the Assessment of Sadistic Personality

Establishing measurement equivalence is crucial in ensuring that a questionnaire assesses a construct the same way across groups (Embretson & Reise, 2000). For example, several studies have found that on average, men score higher than women on sadism (e.g., Plouffe et al., 2017, 2019). However, it remains unclear whether these findings represent true gender differences or measurement differences across men and women.

Differential item functioning can be used to evaluate whether questionnaire items operate the same way for men and women. Specifically, DIF occurs if groups of men and women with the same level of latent subclinical sadism differ in probabilities of endorsing the ASP items, resulting in biased measurements. These gender differences may be due to, for example, different interpretations of item content, different motivations for endorsing item responses, or differences in relevance (e.g., Edelen et al., 2009). If, however, measurement equivalence is found, then men and women can be compared in terms of their position on the ASP, and its validity will be further established.

Objective

The aim of this study is to comprehensively evaluate the item properties of the ASP across three translations (i.e., English, Polish, and Italian) to ensure item validity. We will investigate both relationships between latent trait sadism and ASP item responses, as well as measurement equivalence of the ASP for men and women. This is important to determine the psychometric soundness of the ASP for the assessment of sadism within and between different countries.

Method

Participants and Procedure

Community and university student samples were recruited from Italy, Poland, and Canada for the present research. These data were drawn from a larger personality study that evaluated the cross-national invariance of the ASP (Kowalski et al., 2020). Study procedures were approved by the respective institutional ethical review boards. To achieve accurate parameter estimates, sample sizes of $n = 500$ were required for this study (Reeve & Fayers, 2005).

The Italian sample comprised 568 participants (340 women, 228 men) between 18 and 30 years of age (M = 23.57, SD = 2.55). The sample comprised mostly university students (64.30%, n = 365), followed by workers (31.70%, n = 180), and unemployed participants (4.00%, n = 23). Participants in the Polish sample were 556 individuals (411 women, 144 men, 1 other) ranging in age from 16 to 70 years (M = 23.48, SD = 4.60). Again, most participants were students (54.50%, n = 303), 38.70% of participants were workers (n = 215), 6.50% of participants were unemployed (n = 36), and 0.40% of participants (n = 2) were retired. Italian and Polish participants were invited to partake in the online study through announcements on the social networking website, Facebook. They were compensated with a small monetary reward (approximately US $0.70).

The Canadian sample included 638 undergraduate students (456 women, 181 men, 1 unspecified) ranging in age from 17 to 43 years (M = 18.50, SD = 2.10). Canadian participants completed the ASP and a series of other personality questionnaires online. They received course credit for their participation.

Measures

The Assessment of Sadistic Personality (ASP; Plouffe et al., 2017)

The ASP is a 9-item self-report measure of subclinical sadism. Participants respond to items on a 5-point Likert rating scale (1 = *strongly disagree*, 5 = *strongly agree*). Italian and Polish versions of the ASP were created using back-translation procedures (Kowalski et al., 2020). Two of the original ASP authors were involved in this process to ensure that item meaning was upheld. Past studies support the reliability and validity of the ASP (Plouffe et al., 2017, 2019).

Data Analytic Strategy

To achieve our objective of investigating relationships between latent trait sadism and ASP item responses across Italian, Polish, and English translations, we calibrated ASP items separately for each sample using Samejima's (1969, 1996) unconstrained GRM in R Version 3.5.2 (R Development Core Team, 2019) with the ltm package (Rizopoulos, 2006). For each item, one a and four b parameters were produced, in addition to item CCCs and TIFs. We used three methods to evaluate the assumption of scale unidimensionality: Horn's (1965) parallel analysis, the minimum average partial test (MAP; Velicer, 1976), and the broken-stick method (Jackson, 1993).

To determine whether measurement equivalence exists for men and women across ASP translations, DIF was assessed using the non-parametric Mantel-Haenszel (MH) method (Cochran, 1954; Holland & Thayer, 1988; Mantel & Haenszel, 1959) in jMetrik Version 4.1.1 (Meyer, 2018). Using the MH method, reference and focal groups (i.e., gender groups) are split into subgroups representing matched observed ASP scores. ETS classification levels were used to categorize the magnitude of DIF (Zwick, Thayer, & Mazzeo, 1997). Rather than solely relying on statistical significance, the ETS classification is dependent on practical significance (Meyer, 2014). This classification relies on an sP-DIF* (i.e., the sP-DIF divided by the item score range; Meyer, 2014) as an effect size. According to the classification, an item that is classified as A has an sP-DIF* value that is strictly less than .05, indicating no DIF of practical significance. An item classified as B has an sP-DIF* value that is greater than and inclusive of .05, but less than .10, and indicates the presence of DIF of moderate practical significance. Finally, a C classified item has an sP-DIF* value that is .10 or greater and has a high amount of DIF, practically speaking, and is of greatest concern.

Results

Descriptive Statistics

Descriptive statistics for men and women are presented in Table 1. Response endorsement proportions for the ASP across samples are presented in Table 2. Overall, response endorsement proportions spread as expected across the different category options, with smaller endorsement proportions for the *Strongly Agree* option.

Graded Response Model

Before item response theory models were conducted, the unidimensionality of the ASP was assessed using Horn's (1965) parallel analysis, Velicer's (1976) MAP test, and the broken-stick method (Jackson, 1993). For the Canadian and Polish samples, the parallel analysis indicated that there should be three factors/one component retained.

Table 1. Descriptive statistics for all samples

Scale	α	ω	Women *M (SD)*	Men *M (SD)*
ASP Canada	.87	.92	15.60 (6.09)	20.54 (6.00)
ASP Poland	.83	.88	17.93 (6.40)	22.58 (7.69)
ASP Italy	.86	.90	14.83 (5.57)	19.01 (7.14)

Note. Men scored significantly higher than women across all samples.

Table 2. Item response proportions for Assessment of Sadistic Personality across samples

Item	Strongly Disagree (1)	Disagree (2)	Neither Agree nor Disagree (3)	Agree (4)	Strongly Agree (5)
Canada ASP1	.42	.35	.11	.11	.01
Canada ASP2	.51	.34	.10	.04	.01
Canada ASP3	.56	.28	.12	.03	.004
Canada ASP4	.57	.27	.09	.06	.01
Canada ASP5	.54	.24	.13	.08	.01
Canada ASP6	.56	.28	.11	.04	.01
Canada ASP7	.43	.25	.16	.14	.02
Canada ASP8	.40	.27	.15	.15	.03
Canada ASP9	.29	.33	.14	.14	.11
Poland ASP1	.47	.26	.14	.10	.03
Poland ASP2	.47	.26	.14	.10	.03
Poland ASP3	.52	.24	.13	.07	.04
Poland ASP4	.49	.21	.14	.11	.05
Poland ASP5	.27	.21	.21	.20	.12
Poland ASP6	.73	.17	.05	.04	.01
Poland ASP7	.62	.15	.08	.10	.05
Poland ASP8	.33	.24	.16	.18	.09
Poland ASP9	.22	.19	.20	.23	.16
Italy ASP1	.63	.21	.07	.07	.01
Italy ASP2	.53	.24	.10	.10	.02
Italy ASP3	.74	.15	.05	.04	.01
Italy ASP4	.53	.22	.14	.11	.01
Italy ASP5	.58	.18	.12	.11	.02
Italy ASP6	.73	.19	.05	.03	.01
Italy ASP7	.53	.22	.14	.08	.02
Italy ASP8	.36	.27	.17	.15	.05
Italy ASP9	.30	.28	.17	.16	.10

In the Italian sample, the parallel analysis indicated that two factors/one component should be retained. However, parallel analysis can be influenced by large sample sizes, such that eigenvalues associated with the random factors approach 1.00 (Revelle, 2016). Across all samples, the Velicer MAP test and the broken-stick method indicated that one factor should be retained (see Electronic Supplementary Material, ESM 1). Based on these findings, the ASP satisfied the unidimensionality assumption. We also examined the correlations between residuals to test for local independence. In our samples, one residual correlation of 36 (0.03%) in Italy, two of 36 (0.06%) in Canada, and three of 36 (0.08%) in Poland exceeded the typical .20 cut-off value (Chen & Thissen, 1997). However, residual correlations largely depend on number of items, response categories, and sample size (Christensen et al., 2017). Marais (2013) further indicated that for scales comprising below 20 items, it is not accurate to directly interpret residual correlations. Thus, it is reasonable to interpret tests of local independence with caution.

Item parameters and standard errors for the samples are displayed in Table 3 and item CCCs are shown in

Table 3. Item response parameters for Assessment of Sadistic Personality across samples

Item	a	b_1	b_2	b_3	b_4
Canada ASP1	2.57	−0.29	1.36	1.76	3.21
Canada ASP2	2.85	0.12	1.68	2.45	3.04
Canada ASP3	3.03	0.22	0.94	1.80	2.35
Canada ASP4	3.78	0.18	0.86	1.29	1.86
Canada ASP5	3.13	0.22	0.82	1.50	2.42
Canada ASP6	3.32	0.20	0.87	1.58	2.10
Canada ASP7	1.39	−0.10	0.93	1.77	4.05
Canada ASP8	1.21	−0.29	0.77	1.60	3.76
Canada ASP9	0.22	−4.25	1.74	4.82	9.56
Poland ASP1	2.00	−0.11	0.78	1.47	2.61
Poland ASP2	1.91	−0.09	0.83	1.53	2.61
Poland ASP3	2.38	0.06	0.84	1.47	2.21
Poland ASP4	2.35	−0.04	0.63	1.19	2.06
Poland ASP5	2.02	−0.82	−0.10	0.57	1.53
Poland ASP6	2.45	0.72	1.51	2.04	3.09
Poland ASP7	0.99	0.55	1.36	1.98	3.33
Poland ASP8	0.96	−0.93	0.31	1.19	2.71
Poland ASP9	0.56	−2.37	−0.61	0.93	3.23
Italy ASP1	2.67	0.45	1.26	1.70	2.66
Italy ASP2	2.05	0.15	1.05	1.55	2.65
Italy ASP3	2.96	0.81	1.51	1.95	2.75
Italy ASP4	2.19	0.13	0.90	1.55	2.99
Italy ASP5	3.28	0.27	0.85	1.34	2.40
Italy ASP6	3.12	0.75	1.63	2.13	2.89
Italy ASP7	1.27	0.15	1.16	2.13	3.46
Italy ASP8	1.61	−0.50	0.50	1.26	2.55
Italy ASP9	0.76	−1.20	0.56	1.65	3.31

Note. *a* represents item discrimination or degree to which items differentiate between participants with varying levels of θ. *b1–b4* represents item thresholds in which participants have a 50% probability of selecting each category or higher.

Figures 1–3. Across items, the CCCs generally indicate that there is good discrimination between different response options. Notably, across all samples, discrimination was lowest for Item 9 ("I would not purposely hurt anybody, even if I didn't like them"). This is consistent with the finding that compared to the other items, similar proportions of individuals selected each response option for Item 9, which is the only negatively-worded item (see Table 2). In the Polish sample, six of nine items had b_1 values below 0, whereas in the other two samples, most b_1 were above 0, indicating that higher levels of θ were required to endorse even the lowest response category in the Canadian and Italian samples. Based on these results and the results of past studies (i.e., Plouffe et al., 2019), we examined the item properties using IRT with Item 9 excluded (see ESM 1). Excluding Item 9 did not result in notable changes to the other items' properties.

Test information functions are shown in Figure 4 for each translation. Overall, they indicated that the ASP provides the most information when levels of θ are between approximately 0 and 4. In other words, the most reliable information is provided when sadism levels are above the mean. However, even at 1 *SD* below the mean, reliability is still high (given r_{xx} = information/[1 + information]).

Differential Item Functioning

When evaluating DIF levels using ETS classification in the Canadian sample, both items 7 ("Watching people get into fights excites me") and 9 ("I would not purposely hurt anybody, even if I didn't like them") were classified as B, indicating that their sP-DIF* value was greater than .05, but less than .10 (moderate practical significance). Specifically, men more readily endorsed Item 7 and women more readily endorsed Item 9 when they were matched on ASP total scores. All other items were classified as A items, indicating that their sP-DIF* value was less than .05 (practically speaking, nonsignificant DIF; Meyer, 2014; Zwick et al., 1997; Table 4).

In the Italian sample, Item 4 ("When I mock someone, it is funny to see them get upset") and Item 9 were classified

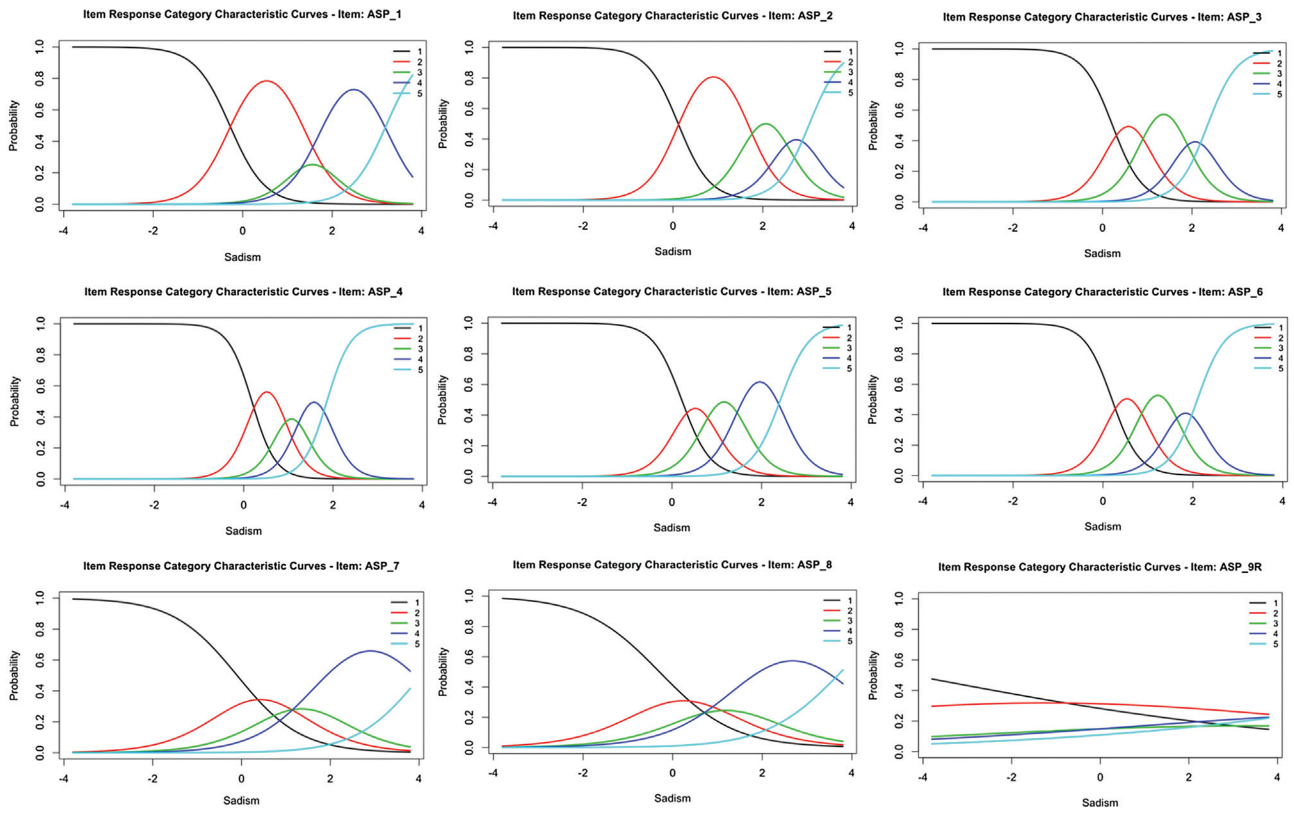

Figure 1. Item response category characteristic curves – Canada.

as B items, indicating that the level of DIF was moderately practically significant. Item 4 favored men, and similar to the Canadian sample, Item 9 favored women.

Finally, in the Polish-speaking sample, Item 5 ("Being mean to others can be exciting") and Item 7 were classified as B items with moderate DIF levels. Item 5 favored women, and like the Canadian sample, Item 7 favored men.[1] We also examined DIF with Item 9 excluded (see ESM 1). Excluding Item 9 did not result in changes to other items' DIF levels (Table 2 in ESM 1).

Discussion

When assessing the validity of a psychological measure, it is of great importance to investigate relationships between latent variables and their manifestations, as well as to ensure that the measure is invariant across different groups. Overall, based on the IRT analyses, the ASP demonstrated sound psychometric properties across its translations. Specifically, the ASP items discriminated adequately, and category thresholds spread well across varying levels of latent sadism. However, Item 9 ("I would not purposely hurt anybody, even if I didn't like them") had the smallest discrimination values across each sample. Additionally, according to the response endorsement proportions, participants were more likely to endorse the positive end of Item 9 than the remaining items. One possible explanation for these findings pertains to the item's negative wording. Although negatively-worded items may reduce response biases by encouraging respondents to engage in more controlled cognitive processing (Podsakoff et al., 2003), research has shown that these items tend to be less reliable and valid than short and concise items because they are more likely to measure multiple constructs (Holden et al., 1985). Alternatively, this finding can be explained by item content. The item "I would not purposely hurt anybody, even if I didn't like them" is innocuous in comparison to the remaining ASP items. Additionally, aside from sadism, there are several individual difference variables that

[1] The authors conducted gender invariance testing separately across countries to supplement DIF analyses. Partial scalar invariance was achieved when the intercept for Item 7 was freed in Canada and Poland. Scalar invariance was achieved in Italy. Across all samples, men scored higher on latent sadism than women (see ESM 1 for more information).

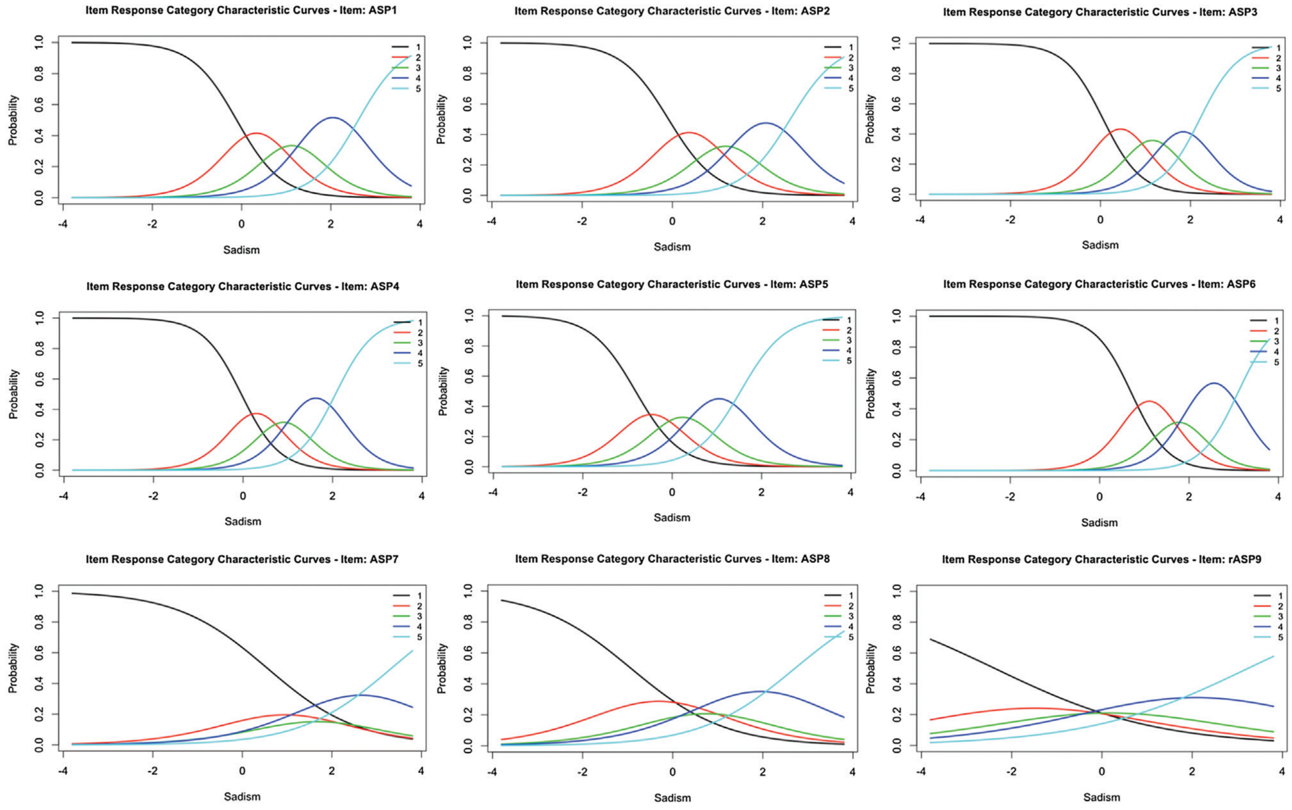

Figure 2. Item response category characteristic curves – Poland.

demonstrate significant relationships with revenge-seeking (Mullet et al., 2005). Therefore, it is probable that individuals low on latent sadism but high on these other individual difference variables may hurt another individual when they believe that they are wronged, which would lead them to endorse this item. Based on these and previous findings that Item 9 has the weakest loading on latent sadism (Plouffe et al., 2017, 2019), we reran the IRT analyses without Item 9 for each sample (see ESM 1). We found negligible changes in the IRT results, indicating that the scale functions well at the item level without Item 9 included. Therefore, we propose the use of an 8-item ASP (ASP8).

Results of the DIF analyses revealed that the ASP items were largely unbiased across genders. However, two items from each ASP translation had DIF rates that were of moderate practical significance, as indicated by the ETS Classification. Item 9 displayed moderate levels of DIF in both the Canadian and Italian samples and in both cases, the DIF favored women. This could reflect the fact that women generally tend to be less aggressive than men (Bettencourt & Miller, 1996). In the Canadian and Polish samples, Item 7 ("Watching people get into fights excites me") also displayed moderate levels of DIF favoring men. This is unsurprising as men often prefer more combative entertainment. For example, Sargent et al. (1998) found that men preferred contact sports such as football, ice hockey, boxing, and Karate, whereas women preferred gymnastics, skiing, diving, and figure skating. In the Italian sample, Item 4 ("When I mock someone, it is funny to see them get upset") displayed moderate DIF. This result is consistent with past research that has shown that men typically score lower than women on such variables as agreeableness across multiple cultures (e.g., Costa et al., 2001). This finding could also reflect gender differences in humor styles, as men tend to use aggressive humor more than women (e.g., Baughman et al., 2012). Finally, in the Polish sample, Item 5 ("Being mean to others can be exciting") displayed moderate DIF favoring women. One possible explanation for this finding is that "being mean", although a general term, may be more strongly associated with relational aggression than physical aggression. Therefore, this finding may reflect sex differences in the propensity for relational aggression, as women tend to engage in relational aggression more relative to men (Archer & Coyne, 2005). Of course, these explanations are post hoc and cannot be assumed based on our results, but rather taken as suggestions requiring further study.

This work has important implications for personality research and its applications across clinical and even

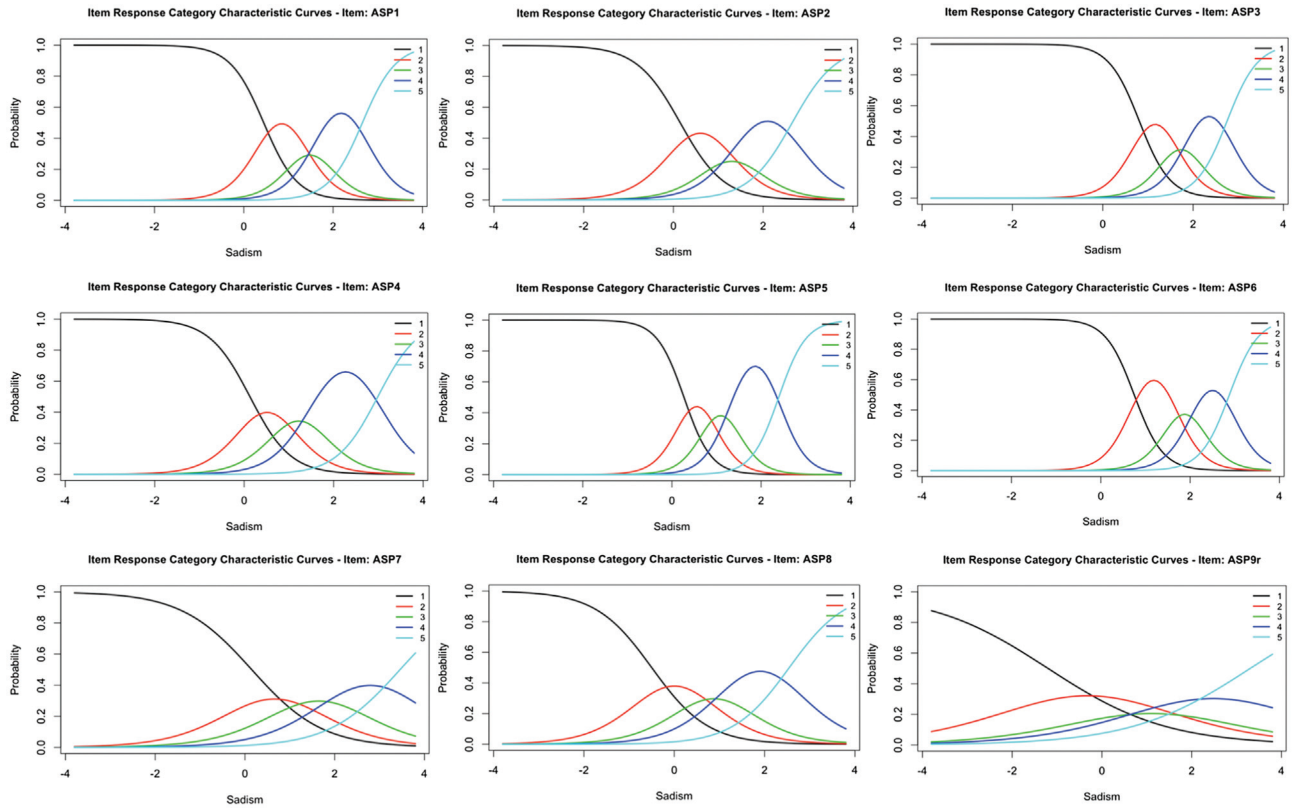

Figure 3. Item response category characteristic curves – Italy.

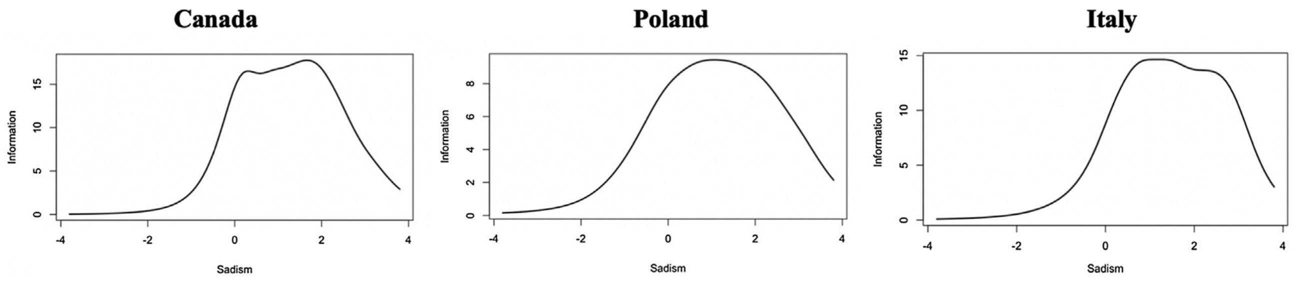

Figure 4. Test information functions for Canada, Poland, and Italy.

vocational domains. To achieve accurate estimates of important outcome variables, such as aggression, workplace deviance, or bullying, among others, it is imperative that the ASP is recognized as a valid and reliable measurement tool for subclinical sadism by using sophisticated analysis techniques (i.e., IRT) with advantages over CTT methods. Our results showed that the ASP items adequately discriminated between individuals with varying trait levels and demonstrated invariance across men and women. However, perhaps the most important takeaway from this research is that Item 9 was ineffective in assessing levels of sadism across three languages. This calls for a modification of the ASP, such that we propose the use of an 8-item ASP (ASP8). Although Item 9 bears little weight in estimating latent sadism scores, we recommend that future research should further examine the viability of the ASP8 for accurate assessment of sadism or modify the item such that it is positively phrased.

Limitations and Future Directions

Our investigation had several limitations. First, our data were self-report in nature. Future research should investigate the properties of the ASP using multimethod and multi-informant study designs to ensure accurate representations of individuals' personality characteristics.

Table 4. Mantel-Haenszel differential item functioning summary across samples

Item	χ^2	*p* value	Effect size (95% CI)
Canada ASP1	0.01	.920	.02 (−.14, .17)
Canada ASP2	0.60	.440	−.04 (−.17, .08)
Canada ASP3	0.08	.780	.01 (−.12, .14)
Canada ASP4	0.30	.580	−.04 (−.17, .09)
Canada ASP5	5.43	.020	−.16 (−.29, −.02)
Canada ASP6	0.03	.870	−.01 (−.13, .12)
Canada ASP7	14.60*	.001	.32 (.14, .50)
Canada ASP8	0.90	.340	.11 (−.09, .31)
Canada ASP9	4.16*	.040	−.21 (−.42, .01)
Poland ASP1	1.89	.170	.03 (−.12, .18)
Poland ASP2	0.22	.640	−.03 (−.19, .13)
Poland ASP3	5.54	.020	.16 (.02, .30)
Poland ASP4	0.77	.380	.05 (−.10, .20)
Poland ASP5	5.78*	.020	.25 (.05, .45)
Poland ASP6	0.12	.730	−.03 (−.13, .06)
Poland ASP7	16.26*	.001	−.37 (−.55, −.19)
Poland ASP8	0.58	.450	−.09 (−.30, .13)
Poland ASP9	0.04	.850	.03 (−.23, .29)
Italy ASP1	0.12	.730	−.01 (−.10, .08)
Italy ASP2	4.07	.040	−.16 (−.28, −.04)
Italy ASP3	0.39	.530	.02 (−.05, .10)
Italy ASP4	18.40*	.001	−.25 (−.36, −.14)
Italy ASP5	0.00	.970	.01 (−.08, .10)
Italy ASP6	0.68	.410	−.04 (−.11, .03)
Italy ASP7	0.00	.950	.00 (−.01, .26)
Italy ASP8	3.02	.080	.13 (−.01, .26)
Italy ASP9	8.56*	.001	.29 (.12, .46)

Note. Items marked with * flagged for DIF. Matching variable = ASP total score. DIF focal group = women; reference group = men.

The unequal distribution of men and women may have had an impact on our results. In addition, each of our samples had a mean age between 18 and 23 years. Thus, it is possible that our findings will not translate to older age groups. Future research should test our hypotheses across more balanced gender samples and different age groups.

As mentioned above, our results also call into question the effectiveness of Item 9 of the ASP. Negatively-worded items purportedly reduce the risk of response bias, but recent research has questioned the overall effectiveness of these items (e.g., Van Sonderen et al., 2013). We thus conducted IRT and DIF analyses without Item 9 and found negligible changes to our results. Therefore, we propose the use of an 8-item ASP (ASP8).

This study was the first to evaluate relationships between latent variables, gender, and response patterns of the ASP items. Our results showed that the ASP accurately represents manifestations of latent sadism with the exception of Item 9. Overall, our results support the ASP (or the ASP8) as an effective self-report measurement tool for assessing subclinical sadism across multiple countries.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000634

ESM 1. Item Response Parameters for Assessment of Sadistic Personality – 8 item – across samples

References

Archer, J., & Coyne, S. M. (2005). An integrated review of indirect, relational, and social aggression. *Personality and Social Psychology Review, 9*(3), 212–230. https://doi.org/10.1207/s15327957pspr0903_2

Baughman, H. M., Giammarco, E. A., Veselka, L., Schermer, J. A., Martin, N. G., Lynskey, M., & Vernon, P. A. (2012). A behavioral genetic study of humor styles in an Australian sample. *Twin Research and Human Genetics, 15*(3), 663–667. https://doi.org/10.1017/thg.2012.23

Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin, 119*(3), 442–447. https://doi.org/10.1037/0033-2909.119.3.422

Buckels, E. E., & Paulhus, D. L. (2014). *Comprehensive Assessment of Sadistic Tendencies* (CAST). (Unpublished instrument) University of British Columbia.

Chabrol, H., Van Leeuwen, N., Rodgers, R., & Séjourné, N. (2009). Contributions of psychopathic, narcissistic, Machiavellian, and sadistic personality traits to juvenile delinquency. *Personality and Individual Differences, 47*(7), 734–739. https://doi.org/10.1016/j.paid.2009.06.020

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Chester, D. S., DeWall, C. N., & Enjaian, B. (2019). Sadism and aggressive behavior: Inflicting pain to feel pleasure. *Personality and Social Psychology Bulletin, 45*(8), 1252–1268. https://doi.org/10.1177/0146167218816327

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10*(4), 417–451. https://doi.org/10.2307/3001616

Costa, P. T. Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dinić, B. M., Bulut Allred, T., Petrović, B., & Wertag, A. (2020). A test of three sadism measures. *Journal of Individual Differences, 41*(4), 219–227. https://doi.org/10.1027/1614-0001/a000319

Edelen, M. O., McCaffrey, D. F., Marshall, G. N., & Jaycox, L. H. (2009). Measurement of teen dating violence attitudes: An item response theory evaluation of differential item functioning according to gender. *Journal of Interpersonal Violence, 24*(8), 1243–1263. https://doi.org/10.1177/0886260508322187

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality, 19*(4), 386–394. https://doi.org/10.1016/0092-6566(85)90007-8

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Erlbaum.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Jackson, D. A. (1993). Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*(8), 2204–2214. https://doi.org/10.2307/1939574

Kowalski, C. M., Di Pierro, R., Plouffe, R. A., Rogoza, R., & Saklofske, D. H. (2020). Enthusiastic acts of evil: The Assessment of Sadistic Personality in Polish and Italian populations. *Journal of Personality Assessment, 102*(6), 770–780. https://doi.org/10.1080/00223891.2019.1673760

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Marais, I. (2013). Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 111–130). Wiley.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge.

Meyer, J. P. (2018). *jMetrik* (Version 4.1.1). http://www.itemanalysis.com

Min, H., Pavisic, I., Howald, N., Highhouse, S., & Zickar, M. J. (2019). A systematic comparison of three sadism measures and their ability to explain workplace mistreatment over and above the Dark Triad. *Journal of Research in Personality, 82*, Article 103862. https://doi.org/10.1016/j.jrp.2019.103862

Mullet, E., Neto, F., & Riviere, S. (2005). Personality and its effects on resentment, revenge, forgiveness, and self-forgiveness. In E. L. Worthington Jr. (Ed.), *Handbook of forgiveness* (pp. 159–181). Routledge.

O'Meara, A., Davies, J., & Hammond, S. (2011). The psychometric properties and utility of the Short Sadistic Impulse Scale (SSIS). *Psychological Assessment, 23*(2), 523–531. https://doi.org/10.1037/a0022400

Paulhus, D. N., & Jones, (2015). Measures of dark personalities. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 562–594). Academic Press.

Persson, B. N., Kajonius, P. J., & Garcia, D. (2017). Testing construct independence in the Short Dark Triad using item response theory. *Personality and Individual Differences, 117*, 74–80. https://doi.org/10.1016/j.paid.2017.05.025

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2017). The Assessment of Sadistic Personality: Preliminary psychometric evidence for a new measure. *Personality and Individual Differences, 104*, 166–171. https://doi.org/10.1016/j.paid.2016.07.043

Plouffe, R. A., Smith, M. M., & Saklofske, D. H. (2019). A psychometric investigation of the Assessment of Sadistic Personality. *Personality and Individual Differences, 140*, 57–60. https://doi.org/10.1016/j.paid.2018.01.002

Plouffe, R. A., Wilson, C. A., & Saklofske, D. H. (2020). The role of dark personality traits in intimate partner violence: A multi-study investigation. *Current Psychology*. Advance online publication. https://doi.org/10.1007/s12144-020-00871-5

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

R Development Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods of practice* (2nd ed., pp. 55–73). Oxford University Press.

Revelle, W. (2016). *An overview of the psych package*. ftp://cran.r-project.org/pub/R/web/packages/psych/vignettes/overview.pdf

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100.

Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer.

Sargent, S. L., Zillman, D., & Weaver, J. B. III (1998). The gender gap in the enjoyment of televised sports. *Journal of Sport and Social Issues, 22*(1), 46–64. https://doi.org/10.1177/019372398022001005

Van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS One, 8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Describing and categorizing DIF in polytomous items. *ETS Research Report Series*, (1), I-52. https://doi.org/10.1002/j.2333-8504.1997.tb01726.x

History
Received September 3, 2020
Revision received November 17, 2020
Accepted November 21, 2020
Published online February 11, 2021
EJPA Section / Category Personality

Funding
The work of Radosław Rogoza was supported by the National Science Centre, Poland [grant number 2015/19/N/HS6/00685].

ORCID
Rachel A. Plouffe
https://orcid.org/0000-0003-0393-9008

Rachel A. Plouffe
The MacDonald Franklin OSI Research Centre
Lawson Health Research Institute
550 Wellington Road
London, ON N6C 0A7
Canada
rplouffe@uwo.ca