Supplemental Material

This online Supplemental Material describes the methods used to analyze: (1) the Mokken model, (2) the reliability analyses, and (3) the criterion validity analyses. Sijtsma and Molenaarand (2002) provide a good introduction on non-parametric Item Response Theory (IRT). Practical applications of Mokken scaling in Stata and R are available in Hardouin, Bonnaud-Antignac, and Sébille (2011) and Van der Ark (2007), respectively. The current study follows these procedures.

In their simpler form (i.e., MHM), Mokken models have three assumptions: unidimensionality, local independence, and monotonicity. Unidimensionality is a basic assumption of measurement theory and means that a set of items all measure just one construct in common (Hattie, 1985). The unidimensionality of the SeSaS was first assessed through Principal Axis Factoring (PAF) with varimax rotation on the tetrachoric correlation matrix, using two criteria: (1) the ratio of the first-to-second-eigenvalues greater than three (Gorsuch, 1983), and (2) the variance accounted for by the first factor higher than 20% (Reckase, 1979). Unidimensionality was also assessed through Loevinger's (1948) scalability coefficients, which are based on the ratio of observed and expected Guttman errors and represent a measure of homogeneity of the items (*Hi*) and the scale as a whole (*H*). The coefficients range between 0 and 1, where values <.3 indicate poor scalability, values between .3 and .4 indicate weak scalability, values between .4 and .5 indicate medium scalability, and values >.5 indicate good scalability (Mokken, 1971).

Local independence means that the responses to the items are independent, conditional to the subject's location on the latent trait, which should explain all relationships between the items (Hardouin et al., 2011). Local independence is related to unidimensionality and was assumed with the Dimensionality Evaluation to Enumerate Contributing Traits index (DETECT; Zhang & Stout, 1999), which is sensible to violations of this assumption. The

DETECT index is a non-parametric procedure that identifies the total number of dominant dimensions underlying a set of items. It is computed by summing covariances between item pairs for all items of the test, conditioned on the observed scores of the remaining items (Wells, Rios, & Faulkner-Bond, 2016). Values multiplied by 100 typically range between 0 and 5, where values ≥1 indicate large multidimensionality, values between 0.4 and 1 indicate moderate to large dimensionality, values between 0.4 and 0.2 indicate moderate to weak multidimensionality, and a value <0.2 indicate unidimensionality (Reckase, 2009).

Monotonicity is a central feature of item response theory models and allows validating the SeSaS score as an ordinal measure of sexual sadism. It assumes that the probability of scoring positive in an item increases monotonically with increasing values of the latent trait (Hardouin et al., 2011). Monotonicity was initially assessed by observing the graphical representations of the trace lines (nonparametric item response functions; IRF). The trace lines describe the relationship between the probability of item responses and the latent trait and should be non-decreasing. Monotonicity was also assessed using indices based on the Guttman errors (Guttman, 1944), which represent the number of individuals who present a given behavior but do not present a more frequent one (less difficult). From these estimations, it is possible to calculate the number of violations (#vi) of the monotonicity assumption and their statistical significance (#zsig). Due to the low relative frequency of most items, the minimal size of the rest-score groups was set at 25 individuals, the minimum size of violations at .03 (minimum decrease in the IRF to be counted as a violation), and their statistical significance at p <.05. When significant violations were found, their seriousness was assessed with the *crit* statistic defined by Molenaar and Sijtsma (2000), which is a goodness-of-fit index that combines the number of assumption violations, their size and significance. Values <40 indicate that the violation can be ascribed to sampling variation, values between 40 and 80 indicate that further analysis must be considered to draw a conclusion, and values >80 cast serious doubts on the assumption (Molenaar & Sijtsma, 2000).

Besides the three former assumptions (i.e., unidimensionality, local independence, and monotonocity), DMMs must also satisfy the non-interception assumption (i.e., invariant item ordering; IIO). This refers to the item trace lines not crossing each other. That is, items that are more difficult remain more difficult than less difficult ones regardless of the positioning of the individual on the latent continuum of sadism. This property increases the validity of comparisons of the scale scores in different data sets (Van Schuur, 2003). Non-intersection was tested with indexes based on the P-matrix. Items satisfy this assumptions if the elements of the P + + matrix are increasing in each row and the elements of the P − − matrix are decreasing in each row (Hardouin et al., 2011). Using a procedure similar to the test of monotonicity (see Molenaar & Sijtsma, 2000), it is also possible to determine the number of violations, their statistical significance, and the *crit* value for the IIO assumption, which are interpreted in the same way as for monotonicity. In addition, we explored if the SeSaS was of a deterministic Guttman type (i.e., items can be ordered in a reproducible hierarchy) by calculating the coefficient of reproducibility (CR). The CR indicates the percentage of original responses that can be reproduced by knowing the subjects score on the scale. It is calculated as 1− (total number of Guttman errors / (number of items × number of subjects)), where values >.90 indicate a reproducible scale (Guttman, 1944).

Regarding the reliability analyses, the Sijtsma and Molenaar (1987) statistic (*Rho*) estimates reliability as the probability of observing the same result twice and is an unbiased estimator given that the DMM assumptions hold (Van der Ark, 2007). It should be >.70 to indicate a reliable scale (Sijtsma & Molenaarand, 2002). Lambda-2 (λ_2) is a more accurate estimate of the lower bound reliability than Cronbach's alpha (α; e.g., Sijtsma, 2009). Values >.70 in these measures of internal consistency are considered acceptable in the social sciences

(George & Mallery, 2010). Inter-rater reliability was calculated using Fleiss kappa (κ; with ordinal weights due to the nature of the SeSaS total score), which is an adaptation of Cohen's kappa for more than two raters. Values >.75 are considered excellent (Cicchetti, 1994).

Regarding the criterion validity analyses, the Area Under the Curve (AUC) represents the probability that a randomly chosen diagnosed sexual sadist would have a higher score on the SeSaS than a randomly chosen non-sadist. Values >.56, .64, and .71 can be considered as small, medium and large in forensic psychology and psychiatry (Rice & Harris, 2005). These AUC values correspond to point-biserial correlations (r_{pb}) of .10, .24, and .37, respectively. However, compared with Pearson product-moment correlation coefficients, point-biserial correlations underestimate the magnitude of an association the further the base-rate deviates from .50 in either direction (Rice & Harris, 2005). The likelihood ratios correspond to the probability that a diagnosed sexual sadist would have a positive test result (LR+) or a non-sadist would have a negative test result (LR-), given a specific cut score on the SeSaS. LR+ >10 and LR- <0.1 can be interpreted as a large and conclusive probability (Reiman, Goode, Hegedus, Cook, & Wright, 2013). The Youden's (*J*; Youden, 1950) index summarizes the performance of the SeSaS in identifying diagnosed sexual sadists at different cut scores (J = sensitivity + specificity – 1) and may be used as a criterion for selecting an optimal threshold value. Its value ranges from 0 to 1, where 1 represents perfect discrimination.

References

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. doi: http://dx.doi.org/10.1037/1040-3590.6.4.284

George, D., & Mallery, P. (2010). *SPSS for Windows step by step: A simple guide and reference 18.0 update*. Upper Saddle River, NJ: Prentice Hall Press.

Gorsuch, R. L. (1983). *Factor analysis: Second edition*. Hillsdale, NJ: Lawrence Erlbaum.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150. doi: http://dx.doi.org/10.2307/2086306

Hardouin, J.-B., Bonnaud-Antignac, A., & Sébille, V. (2011). Nonparametric item response theory using Stata. *Stata Journal, 11*(1), 30-51.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164. doi: https://doi.org/10.1177/014662168500900204

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin, 45*(6), 507-529. doi: http://dx.doi.org/10.1037/h0055827

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague, Netherlands: Mouton.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual for MSP5 for Windows: A program for Mokken scale analysis for polytomous items*. Groningen, the Nederlands: ProGamma.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207-230. doi: https://doi.org/10.3102/10769986004003207

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reiman, M. P., Goode, A. P., Hegedus, E. J., Cook, C. E., & Wright, A. A. (2013). Diagnostic accuracy of clinical tests of the hip: A systematic review with meta-analysis. *British Journal of Sports Medicine, 47*(14), 893-902. doi: 10.1136/bjsports-2012-091035

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615-620. doi: 10.1007/s10979-005-6832-7

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*(1), 79-97. doi: 10.1007/BF02293957

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric Item Response Theory. *Political Analysis, 11*(2), 139-163. doi:https://doi.org/10.1093/pan/mpg002

Wells, C. S., Rios, J., & Faulkner-Bond, M. (2016). Testing assumptions of Item Response Theory models. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future*. New York, NY: The Guilford Press.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35. doi: https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249. doi: 10.1007/BF02294536