The paraphilia of sexual sadism involves sexual fantasies, urges, and behaviors that center around the subjugation and humiliation of another human being. The afflicted individual obtains sexual gratification from exerting power over another person through acts of cruelty or degradation. Although aspects of power and subjugation may be part of consensual sadomasochistic role play, severe sexual sadism refers to the forensically relevant form of the disorder in which someone else is being victimized against his or her own will. This clearly has legal implications, as it may lead to sexual offenses such as rape or sexual homicide.

Given the link between paraphilia and sexual reoffending in general (Hanson & Morton-Bourgon, 2005) and the putatively increased risk for sexual offense relapses among sexual sadists in particular (Berner, Berger, & Hill, 2003; Kingston, Seto, & Bradford, 2009), there is a need to diagnose sexual sadism as accurately as possible within forensic settings. The results on the reliability of the criteria for diagnosing sexual sadism are inconclusive: In a study with 15 experienced clinicians who independently judged 12 case vignettes on whether a diagnosis of sexual sadism was present in each case, Marshall, Kennedy, Yates, and Serran (2002) found an unacceptably low interrater agreement of $\kappa = .14$. Similarly, Levenson (2004) noted an insufficient level of observer agreement of $\kappa = .30$ with regard to sexual sadism. Doren and Elwood (2009), on the other hand, reported rates of agreement in excess of 90% among 34 experienced raters concerning the diagnosis of sexual sadism. Doren and Elwood did not provide any statistics correcting for chance agreement (such as $\kappa$), however.

Consequently, several authors emphasized the potential utility of behavioral indicators instead of clinical self-report data for diagnosis (e.g., Knight & Prentky, 1990; Marshall & Kennedy, 2003; McLawsen, Jackson, Vannoy, Gagliardi, & Scalora, 2008; Yates, Hucker, & Kingston, 2008). The use of behavioral indicators deriving from information on the offense seems pragmatic: It is plausible that sexual sadists are particularly unwilling to disclose violent sexual fantasies and urges that focus on extreme forms of violence, degradation, or humiliation of others, especially within forensic settings (cf., Dietz, Hazelwood, & Warren, 1990).

It needs to be acknowledged, though, that the current Criterion A for sexual sadism in the revised fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM–IV–TR; American Psychiatric Association, 2000)—like its presumption revision in DSM–5 (Krueger, 2010, p. 342)—focuses on “recurrent, intense sexually arousing fantasies or sexual urges” in which the psychological or physical suffering of a nonconsenting victim is experienced as sexually arousing. These urges or fantasies may involve corresponding acts or lead to distress or interpersonal problems in other ways (Criterion B). Furthermore, violent sexual fantasies seem to be particularly common among individuals who committed grievous sex-
ually sadistic offenses (Burgess, Hartman, Ressler, Douglas, & McCormack, 1986; MacCulloch, Snowden, Wood, & Mills, 1983). This may suggest a possibly causal or at least a shaping or maintaining influence of such deviant sexual fantasies on subsequent violent acts (McGuire, Carlisle, & Young, 1965) even though sexual fantasies including dominance are estimated to occur in 13%–54% of men, according to various surveys with participants from the general population (Leitenberg & Henning, 1995). Hence, behavioral indicators should not replace the diagnostic relevance of sadistic fantasies, but they will likely prove useful, particularly within forensic settings where denial of sexually sadistic fantasies may be more common than in clinical work (Leitenberg & Henning, 1995).

Furthermore, it has been suggested to distinguish more clearly between the relevance of a diagnosis of sexual sadism for forensic purposes and consensual sadomasochistic role play in the community within DSM–5 (Krueger, 2010). It should therefore be emphasized that the use of the behavioral indicators of severe sexual sadism described herein is limited to cases in which sexual violence or bodily harm has been inflicted by the testee against a nonconsenting victim. Although some actions within consensual sadomasochistic role play may appear phenomenologically similar by virtue of accepting dominant or submissive roles (e.g., Santtila, Sandnabba, & Nordling, 2006), consensual sadomasochistic role play as such can serve various motives (such as hedonism or escapism; Taylor & Ussher, 2001) and does not imply psychological distress or prior traumatization of the person, according to a recent national survey (Richters, de Visser, Rissel, Grulich, & Smith, 2008).

In a comparison of psychiatric diagnosis with behavioral indicators of sexual sadism, Kingston, Seto, Firestone, and Bradford (2010) found that only the latter yielded significant predictive validity for sexual offense recidivism. Following the findings from Marshall et al. (2002), Marshall and Hucker (2006) put forward a list of 15 dichotomous (yes/no) behavioral indicators of severe sexual sadism. On the basis of a German sample of sexual offenders from a high-secure forensic psychiatric hospital (N = 100), Nitschke, Osterheider, and Mokros (2009) empirically derived an 11-item subset from this list (including the additional item of “Insertion of objects into victim’s bodily orifice[s]”). The resulting Severe Sexual Sadism Scale (SSSS; see Table 2 for items) showed high reliability (r_u = .93) within the construction sample. More specifically, the SSSS fulfilled the statistical properties of a deterministic (Guttman) scale. A sum score of at least 4 points on the SSSS differentiated perfectly between the (clinically assessed) sexual sadists and nonsadists within the construction sample. Furthermore, the sum score on the SSSS correlated significantly with psychopathy as measured with the total score on the Psychopathy Checklist–Revised (Hare, 2003): r = .29 (Mokros, Osterheider, Hucker, & Nitschke, 2011). The average intrarater reliability across the individual item codings on the SSSS, given by Nitschke et al. (2009), was κ = .86 (range: .65–1.00).

Given that analyses of the SSSS were so far limited to the construction sample, it is paramount to test whether the scale properties (in particular, internal consistency, reliability, and one-dimensionality) carry over to other samples. Hence, we analyzed data from another sample of sexual offenders. We hypothesized:

**Hypothesis 1:** Severe sexual sadism as measured with the 11-item set of the SSSS is a unitary (one-dimensional) construct.

**Hypothesis 2:** The SSSS total score is a sufficient statistic for the level of the underlying trait (sexual sadism), as evidenced by a deterministic (Guttman) scale.

### Method

**Participants**

Participants were 105 adult male sexual offenders who had been evaluated between 2002 and 2004 at the Federal Evaluation Centre for Violent and Sexual Offenders (FECVSO) of the Austrian Prison Service (Schilling, Ross, Pfäfflin, & Eher, 2010). The FECVSO is a department of the Austrian Federal Ministry of Justice. Participants were included consecutively if they had a sexual crime (rape, sexual homicide) as the index offense. In brief, since the end of 2001 every sexual offender convicted of an unconditional prison sentence by an Austrian court has to be reported to the FECVSO. After a file-based risk assessment of every offender, a substantial proportion of these reported offenders (about 60%) are routinely seen for risk assessment by experienced forensic psychiatrists and psychologists at the FECVSO. Selection of offenders for clinical forensic assessment is done by one of the following criteria: a total score of more than 5 points in the Static-99 actuarial risk assessment instrument (Hanson & Thorn- ton, 2000), age under 25, a prison sentence of more than 4 years, a conviction for a child sexual abuse offense with a nonrelated victim, and any offender reconvicted for a sexual crime. Consequently, all participants in the present sample were interviewed clinically (Eher, Matthes, Schilling, Haubner-MacLean, & Rettenberger, 2011).

**Design and Procedure**

The criteria of the SSSS (Nitschke et al., 2009; cf. Marshall & Hucker, 2006) were coded based on clinical and court files. Coding was done by an experienced forensic psychologist who had not been involved in the diagnostic assessment and risk assessment procedures for the cases at hand within the FECVSO, and thus was blinded against the clinical diagnoses. Items were initially coded on a 5-point Likert-type scale (with verbal anchoring points of 0 = clearly absent, 1 = possibly present, 2 = present to some extent, 3 = clearly present, and 4 = clearly dominant feature). In order to make the data commensurate with the analyses done by Nitschke et al. (2009), the data were dichotomized by collapsing coding Categories 0 and 1 as not present and coding Categories 2–4 as present.

Data were analyzed with the PASW Statistics program, Version 18.0.0 (SPSS, Chicago, IL), with the exception of factor analyses that were carried out with Mplus, Version 6.11, for Mac (Muthén & Muthén, Los Angeles, CA). The factor analyses used a uniform least squares estimation algorithm (using the unweighted least squares estimator for exploratory and the means-and-variance-adjusted unweighted least squares estimator for confirmatory factor analysis). Finally, Rasch scaling was implemented in Maple, Version 14.01 (Waterloo Maple, Waterloo, Canada), by the first
author, with conditional maximum likelihood estimation of item parameters based on the paired-comparison algorithm described by Ford (1957). Person parameters were estimated with the weighted maximum likelihood method described by Warm (1989). Similarly, kernel density estimation was facilitated with Maple, Version 14.01, by means of a Gaussian normal kernel and a normal bandwidth parameter.

Results

On average, participants in the sample were 33.19 years old (SD = 9.71; range: 15.31–60.28) at the time of the judicial verdict. The length of the determinate prison sentences of 102 participants was 50.44 months (SD = 46.02; range: 6–258) on average; an additional two offenders were sentenced to life imprisonment. Data on length of sentence was not available for one offender.

According to the assessment of the experienced forensic psychiatrists and psychologists at the FECVSO, a subsample of 18 participants (17.1%) were diagnosed as suffering from sexual sadism according to DSM–IV–TR criteria.

A subset of six cases was chosen at random and coded independently by the third and fourth authors. Averaging joint occurrences and disparities across the six cases and focusing on the 11 items of the SSSS as derived by Nitschke et al. (2009) yielded an overall index of interrater agreement of κ = .58 (95% CI [.40, .77]). According to the criteria given by Landis and Koch (1977), this would indicate a moderate level of agreement. Concerning the total score on the SSSS, the consistency variant of the intraclass correlation coefficient (type [2, 1], single measure) calculated to .82.

As Table 1 shows, most of the items of the SSSS were positively correlated. Exceptions were the two least frequent criteria (Item 5 ["Offender mutilates sexual parts of the victim’s body"] and Item 8 ["Offender mutilates nonsexual parts of the victim’s body"]). These two items showed negative correlations with some of the other items. Overall, the mean tetrachoric correlation between the 11 items was .41. In an exploratory factor analysis, a one-factorial solution (variance explained = .472) accounted for two thirds (68%) of the total communality (6.97) as estimated by Thurstone’s method (based on maximum correlation coefficients) and nearly half (43%) of the total variance.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.50**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.37**</td>
<td>.62***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.68***</td>
<td>.61***</td>
<td>.78***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.01</td>
<td>.47*</td>
<td>.14</td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.44**</td>
<td>.48***</td>
<td>.36*</td>
<td>.28</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.61***</td>
<td>.32</td>
<td>.16</td>
<td>.25</td>
<td>.23</td>
<td>.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.18</td>
<td>.11</td>
<td>.06</td>
<td>.31</td>
<td>.77***</td>
<td>.15</td>
<td>.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.41**</td>
<td>.36**</td>
<td>.39*</td>
<td>.60***</td>
<td>.00</td>
<td>.10</td>
<td>.15</td>
<td>.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.69***</td>
<td>.53***</td>
<td>.75***</td>
<td>.79***</td>
<td>.30</td>
<td>.35*</td>
<td>.61***</td>
<td>.46*</td>
<td>.45***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.61***</td>
<td>.32</td>
<td>.56***</td>
<td>.84***</td>
<td>.72***</td>
<td>.40</td>
<td>.28</td>
<td>.60**</td>
<td>.15</td>
<td>.81***</td>
<td></td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001 (all two-sided).
Under Andrich’s (1988) method (i.e., estimating error variance based on the standard error of measurement of person parameters), the reliability of the 11-item Rasch scale was estimated at $r_{tt}$ = .86 in the sample. Internal consistency of the scale was estimated at Cronbach’s $\alpha$ = .75 and Guttman’s $\lambda_2 = .78$.

The test characteristic curve plots the total score on the SSSS as a function of the underlying trait level $\theta$ (see Figure 1). The test characteristic curve had the steepest ascent at a trait level of $-0.05$, indicating that this trait level differentiated best between cases. A trait level of $0.05$ would correspond with the range between the total scores values of 5 ($\theta = -0.32$) and 6 ($\theta = 0.33$).

The score distribution of the sexual sadists was swamped by the score distribution of the nonsadistic sexual offenders due to sexual sadists representing the minority (17%) in the sample (see Figure 2). The two graphs in Figure 2 come closest to intersection at a raw score total of 7.55. The intersection would thus imply a cutoff at a total score of 7. A cutoff value of 7 points would afford a sensitivity (true-positive rate) of 56%, a specificity (true-negative rate) of 90%, and a selection ratio (or posterior probability) of 53% within the present sample. The cutoff of 4 points originally recommended by Nitschke et al. (2009) would yield a sensitivity of 83%, a specificity of 58%, and a selection ratio (posterior probability) of 29%.

Generally, the probability that a sexual sadist, drawn at random from the subsample of sexual sadists, would have a higher score on the SSSS than a randomly chosen nonsadistic sexual offender was 81%. This value represents the area under a receiver operating characteristic curve: $$.81 (p < .001, 95\% \text{ CI } [.72, .91])$. Expressed in terms of standard deviation units, the difference between the arithmetic mean of sexual sadists on the SSSS ($M = 5.94, SD = 1.86, Mdn = 7$) and the corresponding score of the nonsadistic sexual offenders ($M = 3.31, SD = 2.22, Mdn = 3$) calculated to $d = 1.23$, which represents a large effect size.

### Table 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Percent present</th>
<th>Factor loading</th>
<th>$\sigma_i$</th>
<th>Infit</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Offender is sexually aroused by the act.</td>
<td>61.0</td>
<td>.73</td>
<td>$-1.9681$</td>
<td>0.1415</td>
<td>.556</td>
</tr>
<tr>
<td>2. Offender exercises power/control/domination over the victim.</td>
<td>59.0</td>
<td>.56</td>
<td>$-1.7340$</td>
<td>0.5342</td>
<td>.703</td>
</tr>
<tr>
<td>3. Offender humiliates or degrades the victim.</td>
<td>29.5</td>
<td>.70</td>
<td>0.0250</td>
<td>0.3060</td>
<td>.620</td>
</tr>
<tr>
<td>4. Offender tortures the victim or engages in acts of cruelty on the</td>
<td>24.8</td>
<td>.92</td>
<td>$0.4487$</td>
<td>$-1.1040$</td>
<td>.135</td>
</tr>
<tr>
<td>5. Offender mutilates sexual parts of the victim’s body.</td>
<td>4.8</td>
<td>.31</td>
<td>2.0253</td>
<td>$-0.3285$</td>
<td>.371</td>
</tr>
<tr>
<td>6. Offender engages in gratuitous violence toward the victim.</td>
<td>7.90</td>
<td>.44</td>
<td>$-2.9562$</td>
<td>0.9442</td>
<td>.827</td>
</tr>
<tr>
<td>7. Offender keeps trophies of the victim.</td>
<td>9.5</td>
<td>.46</td>
<td>1.6583</td>
<td>0.9157</td>
<td>.820</td>
</tr>
<tr>
<td>8. Offender mutilates nonssexual parts of the victim’s body.</td>
<td>3.8</td>
<td>.40</td>
<td>2.6603</td>
<td>$-0.1077$</td>
<td>.457</td>
</tr>
<tr>
<td>9. Victim is abducted or confined.</td>
<td>60.0</td>
<td>.46</td>
<td>$-1.7947$</td>
<td>2.1623</td>
<td>.985</td>
</tr>
<tr>
<td>10. Evidence of ritualism in the offense.</td>
<td>35.2</td>
<td>.97</td>
<td>$-0.4864$</td>
<td>1.6244</td>
<td>.052</td>
</tr>
<tr>
<td>11. Insertion of objects into victim’s bodily orifices.</td>
<td>9.5</td>
<td>.85</td>
<td>1.9417</td>
<td>$-0.5436$</td>
<td>.293</td>
</tr>
</tbody>
</table>

Note. $\sigma_i$ = item difficulty parameter according to Rasch scaling; infit = weighted mean square standardized residual ($z$ transform); an item fit statistic that denotes significant underfit (i.e., lack of discrimination if infit is greater than 1.96, $p > .975$) or overfit (i.e., an almost deterministic response pattern if infit is less than $-1.96$, $p < .025$; Wright & Masters, 1982).

![Figure 1](https://example.com/figure1.png)  
**Figure 1.** Test characteristic curve of the Severe Sexual Sadism Scale (solid line) describing the relationship between the latent trait ($\theta$) and the total score, with 95% confidence bands (dashed lines).
Discussion

The present study replicated the structure of an 11-item set indicative of severe sexual sadism with a sample of 105 sexual offenders from Austria (18 of whom had been diagnosed as sexual sadists). Confirmatory factor analysis supported the view that the items represent a one-dimensional scale. However, the deterministic (Guttman) properties that were identified in the construction sample (Nitschke et al., 2009) could not be confirmed. Still, the 11 items were commensurate with the one-parameter logistic (Rasch) model. Hence, the item set retained the properties of forming a cumulative scale and, more importantly, of the total score as a sufficient statistic for the underlying trait. Similarly, O’Meara, Davies, and Hammond (2011) recently applied Rasch modeling to a brief self-rating scale from a related domain, the Short Sadistic Impulse Scale, capturing a (nonsexual) sadistic personality trait.

Estimates of reliability as well as sensitivity and specificity were lower than in the construction sample. Still, at \( r_a = 0.86 \), the reliability of the scale within the present sample met the standard for clinical decisions (\( r_a \geq 0.85 \)) suggested by Rosenthal and Rosnow (1991). In the current testing sample, the criterion validity of the SSSS with regard to the clinical DSM–IV–TR diagnosis of sexual sadism was good, with a large effect size for distinguishing participant groups. The cutoff of 4 points suggested by Nitschke et al. (2009) would lead to 42% false positives and 17% false negatives, affording a biased selection ratio in the present sample (with a base rate of 17.1% sexual sadists): Among the individuals above the cutoff, only one in three would meet the clinical criteria for sexual sadism. The latent trait level corresponding with the cutoff of 4 points was too low. For the higher cutoff of 7 points that afforded a more acceptable specificity (90%) and a selection ratio above 50%, the standard error of measurement of the associated trait level (\( \theta = 0.99 \)) was 0.82. The range of \( \theta = 0.99 \pm 1 \) standard error of measurement covers the interval [0.12, 1.81]. Expressed in terms of the total score, this would entail the neighboring values of 6 points (\( \theta = 0.33 \)) and 8 points (\( \theta = 1.64 \)), respectively.

Given the difficulties in ascertaining sexually sadistic fantasies or urges of sexual offenders within forensic settings, the SSSS may therefore become a potentially useful complement to other diagnostic procedures. The use of the scale is limited to forensic cases in which sexual offenses or violent acts were committed against nonconsenting victims. It should be acknowledged, though, that development and replication of the 11-item set were based on data from two German-speaking countries (Austria and Germany). Consequently, it would be necessary to test the scale properties in English-speaking and other countries as well. Furthermore, it should be noted that the construction (\( N = 100 \)) and the present testing sample (\( N = 105 \)) were comparatively small. In particular, the present testing sample did not allow for splitting the sample in order to conduct a test on the homogeneity of person parameters (Andersen, 1973). Therefore, replications with larger data sets are needed. Finally, it would be highly interesting to find out whether the total score on the SSSS is predictive of violent or sexual reoffending among sexual offenders released from custody. A study that addresses this issue and uses a retrospective coding of cases is currently under way.

References


Conference of the Association for the Treatment of Sexual Abusers, Dallas, TX.


Received May 4, 2011
Revision received October 24, 2011
Accepted October 27, 2011