



UvA-DARE (Digital Academic Repository)

Do We Hold Males and Females to the Same Standard? A Measurement Invariance Study on the Psychopathy Checklist-Revised

Klein Haneveld, E.; Molenaar, D.; de Vogel, V.; Smid, W.; Kamphuis, J.H.

DOI

[10.1080/00223891.2021.1947308](https://doi.org/10.1080/00223891.2021.1947308)

Publication date

2022

Document Version

Final published version

Published in

Journal of Personality Assessment

License

Article 25fa Dutch Copyright Act

[Link to publication](#)

Citation for published version (APA):

Klein Haneveld, E., Molenaar, D., de Vogel, V., Smid, W., & Kamphuis, J. H. (2022). Do We Hold Males and Females to the Same Standard? A Measurement Invariance Study on the Psychopathy Checklist-Revised. *Journal of Personality Assessment*, 104(3), 368-379. <https://doi.org/10.1080/00223891.2021.1947308>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).






Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Do We Hold Males and Females to the Same Standard? A Measurement Invariance Study on the Psychopathy Checklist-Revised

Evelyn Klein Haneveld¹ , Dylan Molenaar² , Vivienne de Vogel³ , Wineke Smid³ , and Jan H. Kamphuis² 

¹FPC Oostvaarderskliniek; ²Psychological Methods, University of Amsterdam (UvA); ³FPC Van der Hoeven Kliniek

ABSTRACT

Psychopathy in females has been understudied. Extant data on gender comparisons using the predominant measure of assessment in clinical practice, the Psychopathy Checklist Revised (PCL-R), points to a potential lack of measurement invariance (MI). If indeed the instrument does not perform equally (well) in both genders, straightforward comparison of psychopathy scores in males and females is unwarranted. Using a sample of female and male forensic patients ($N = 110$ and $N = 147$ respectively), we formally tested for MI in a structural equation modeling framework. We found that the PCL-R in its current form does not attain full MI. Four items showed threshold-biases and particularly Factor 2 (the Social Deviance Factor) is gender biased. Based on our findings, it seems reasonable to expect that specific scoring adjustments might go a long way in bringing about more equivalent assessment of psychopathic features in men and women. Only then can we begin to meaningfully compare the genders on the prevalence, structure, and external correlates of psychopathy.

ARTICLE HISTORY

Received 21 November 2020
Accepted 10 June 2021

Psychopathy is considered a highly relevant syndrome in forensic mental health. Its association with violent behavior and criminal recidivism is well established (Hare & Neumann, 2008; Leistico et al., 2008), and it has been found to be predictive of poor treatment response (Ogloff et al., 1990; Wong & Hare, 2005). The most widely used instrument to assess psychopathy in clinical practice is the Psychopathy Checklist-Revised (PCL-R; Hare, 1991, 2003), and a host of studies documented favorable psychometric properties in a variety of samples of predominantly male violent (sexual) offenders. Psychopathy in females was initially substantially understudied, but in the past two decades this situation has improved (Verona & Vitale, 2018). The extant PCL-R research in females yields a picture of possible gender differences in reliability and particularly in validity. A crucial implication of the notion that the instrument may, psychometrically, not perform equally (well) in both genders, is that straightforward comparison of males and females with psychopathy may in fact be unwarranted.

Several studies provided indicative evidence for differences in the psychometric properties of the PCL-R across genders. First, ten years after the publication of the first edition of the PCL-R (Hare, 1991), Vitale and Newman (2001) published a review of studies on the psychometric performance of the PCL-R in females. The authors concluded that the PCL-R appeared to have good inter-rater reliability and internal consistency when used in female populations. The results regarding the base rate of psychopathy among women varied widely across female samples. Of particular

interest was the observation that when restricted to incarcerated women, observed psychopathy prevalences were generally lower than in males, with estimates ranging from 11–23% in women to 15–30% in men. The authors discussed several possible reasons for this finding, one of which was that the PCL-R may not adequately capture psychopathy in women. An alternative explanation was that some of the items of the PCL-R may not have the same sensitivity for females. Moreover, there were indications for gender differences in factor structure. Exploratory factor analysis in male samples originally led to the identification of two underlying factors, a model that was reliably replicated (Hare & Neumann, 2008). Factor 1 contains the personality traits typically associated with psychopathy; i.e., shallow affect, lack of empathy, and a manipulative, arrogant interpersonal style. Factor 2 reflects a chronically impulsive, aggressive and antisocial lifestyle. Preliminary evidence indicated that the factor structure was different in females; that is, not all individual items loaded equally across genders. However, because they used a small sample of mixed ethnicity, Vitale and Newman remained cautious, and recommended the use of factor analysis in larger samples, as well as Item Response Theory (IRT) analysis to compare structure and functioning of items across gender.

Bolt and colleagues took up this challenge and conducted multigroup IRT analyses in several populations, described as “female criminal offenders,” “male criminal offenders,” and “male forensic psychiatric patients” (Bolt et al., 2004). An assumption in IRT is that there is a single latent factor or trait underlying all items of the respective scale. IRT can

detect differences in item-trait functioning across groups. Bolt and colleagues argued that the extant factor models of the PCL-R all have highly intercorrelated factors, and verified that a single-factor model adequately fit their data. Items of the PCL-R that showed differential item functioning across genders were items 5 (Conning/manipulative), 12 (Early behavior problems), 18 (Juvenile delinquency), and 20 (Criminal versatility), with females scoring higher on item 5, and lower on items 12, 18, and 20, respectively. Bolt and colleagues then proceeded their analyses using the original Two-Factor model, as well as the more fine-grained Four-Facet model that Hare introduced in the second edition of the PCL-R (Hare, 2003) to discuss the results. Factor 1 (consisting of the Interpersonal and Affective Facets 1 and 2) did not display large differences between male and female offenders. Evidence for differential item functioning was primarily found in the behavioral items of Factor 2. Of the facets residing under Factor 2, Facet 4 (Antisocial; consisting of items 10, 12, 18, 19, and 20) in particular produced consistently lower scores for females. Bolt and colleagues also computed the information function for the four facets and the Total Score. The information function indicates the precision with which the facets measure the latent traits. In females, Facets 3 (Lifestyle) and 4 (Antisocial) provided notably less information regarding the latent traits than in males. The authors pointed out that not only were the behaviors measured with these facets likely to be less prevalent among females, they were also more weakly related to psychopathy. Nevertheless, based on the information function of the PCL-R Total Score, their conclusion was that the effect on test-functioning was quite modest, especially at the high-end scores. Accordingly, they deemed the PCL-R (Total Score) effective for the task of distinguishing psychopathic from non-psychopathic females.

A more recent systematic review (Beryl et al., 2014) again evaluated the current evidence regarding prevalence and factor structure of psychopathy in female populations in secure settings, defined as both criminal justice and secure inpatient settings. Of 28 studies that were included, 21 used the PCL-R. Heterogeneity in sample characteristics and study procedures prevented quantitative data synthesis, but prevalence in females was again found to be generally lower than in male populations in secure settings. Nine of the reviewed studies investigated factor structure of the PCL-R. Although two studies replicated the Two-Factor/ Four-Facet model that was found in males, Beryl and colleagues concluded that Cooke and Michie's Three-Factor model (2001) had best fit for females. In this model, items related to criminal and antisocial behavior were removed; the remaining three factors were named "Arrogant/Deceitful Interpersonal Style," "Deficient Affective Experience," and "Impulsive/Irresponsible Behavioral Style" (Cooke & Michie, 2001). These three factors are identical to the first three facets of the Four-Facet model, and consist of 13 of the 20 original items of the PCL-R. Cooke and Michie argued that antisocial behavior is a correlate, and possibly better thought of as a possible consequence of psychopathy, rather than a core feature. The debate about whether or not antisocial behavior

should be part of the conceptualization of psychopathy is ongoing. However, considering the IRT-findings by Bolt and colleagues (2004) that illuminated that Facet 4 appears to function less well for females, it is perhaps not surprising that the Three-Factor model yielded best fit.

To our knowledge, there are no other studies comparing the structure of the PCL-R across gender in adult samples. However, two studies utilizing the Psychopathy Checklist: Youth Version (PCL:YV; Forth et al., 2003) may also be informative (Dillard et al., 2013; Tsang et al., 2015). The PCL:YV was modeled on the PCL-R, and although several items were adapted to the specific population of adolescents, it appears to have a similar Two-Factor/Four-Facet structure (Hare et al., 2018). Dillard and colleagues (2013) used IRT to compare functioning of the PCL:YV in adolescent boys ($n=307$) and girls ($n=144$) that had come into contact with the law. For girls, quite a few items (11 of 20: items 1, 3, 4, 6, 7, 11, 12, 13, 16, 18, and 20) demonstrated differential functioning when compared to boys. The information function of all four facets was higher for boys than for girls. However, comparable to Bolt et al. (2004), Factor 1 (incorporating Facet 1 and 2) was less sensitive to gender than Factor 2 (Facet 3 and 4). In general, Factor 1 was more useful in identifying the underlying construct of psychopathy than Factor 2. Tsang et al. (2015) used IRT with Cooke and Michie's Three-Factor model in a larger sample of adolescents involved in the justice system ($N=1007$, 38% female). Of the 13 items, 4 showed differential functioning across gender (2, 13, 14, and 16), partly overlapping with the results of Dillard et al. (2013). In general, Cooke and Michie's Factor 1 and 2 (identical to Facet 1 and 2) were again found to be better at discriminating high levels of the underlying trait of psychopathy, than Factor 3.

Taken together, the observed gender differences in prevalence, factor structure, and item functioning, albeit rather inconsistent across studies, point to the possibility that the PCL-R does not possess the same psychometric properties across gender. This is problematic, as this would imply that the scores on the PCL-R cannot be meaningfully compared across males and females. To be able to meaningfully compare male and female scores, the PCL-R should adhere to measurement invariance across gender (Mellenbergh, 1989; Meredith, 1993). Measurement invariance implies that a male and a female with the same position on a factor underlying the PCL-R (e.g., "lack of empathy") should have the same expected score on the item(s) measuring that factor (see for example Eigenhuis et al., 2017). If measurement invariance is violated, males and females that are (for instance) equally empathic will nevertheless display differences in their empathy scores. As a result, differences in the item scores between males and females do not necessarily indicate that there are "real" differences in the underlying pathology. Therefore, establishing measurement invariance is important to guarantee the comparability of male and female scores of the PCL-R. Additionally, it is a prerequisite to be able to compare predictive validity across gender.

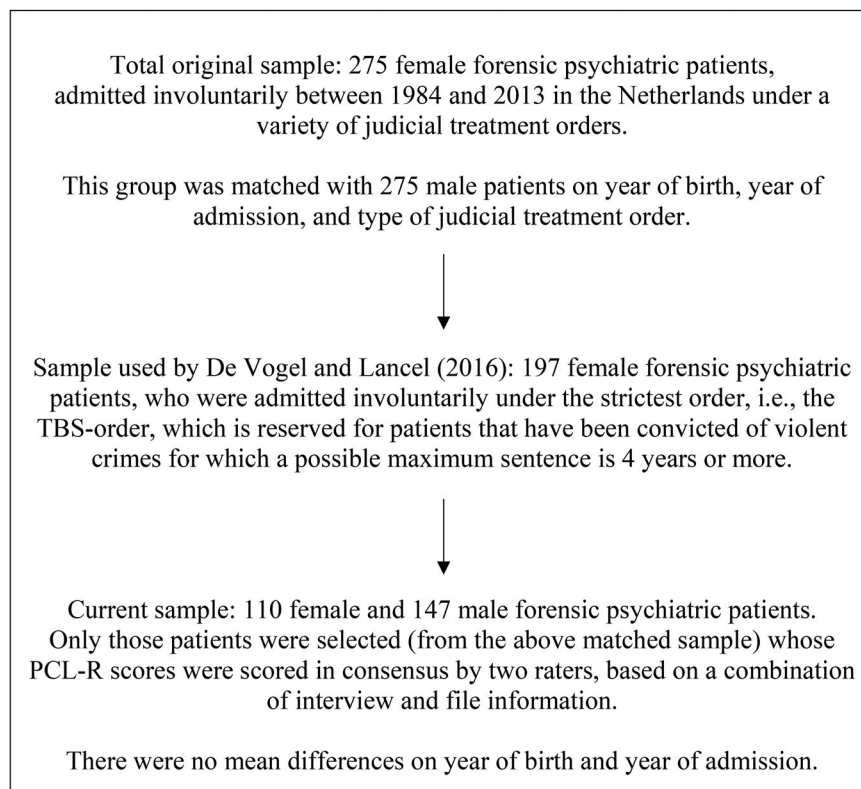


Figure 1. Flowchart of sample composition.

In conclusion, there appears to be only one study testing measurement invariance of the PCL-R across gender in an adult sample (Bolt et al., 2004). Of course, replication is needed in different samples and settings to be able to come to more robust conclusions. As Bolt et al. (2004) only considered female criminal offenders, male criminal offenders, and male forensic psychiatric patients, the aim of the present study is to test for measurement invariance with respect to gender in a forensic psychiatric population of males and females. Key objective is to clarify possible gender biases in the functioning of the PCL-R in forensic psychiatric samples, with the overall aim to inform and improve current clinical practice.

Method

Data

In 2012, a multicenter research project was started in the Netherlands on gender differences in forensic psychiatry (for more information, see De Vogel et al., 2016). The project and all studies connected to it (including the current study) were conducted with official permission from the directors of the hospitals involved, in compliance with the (then) applicable ethical standards. The patients were recruited from four different forensic psychiatric hospitals in The Netherlands. The total female sample ($N=275$) represented nearly all female forensic psychiatric patients in the Netherlands admitted involuntarily between 1984 and 2013 with a variety of judicial treatment orders. The female

sample was matched with a male sample based on year of birth, year of admittance, and type of judicial treatment order.

De Vogel and Lancel (2016) used a subgroup of 197 female and 197 male patients to study gender differences in the assessment and manifestation of psychopathy. All patients were admitted involuntarily under the same treatment order, the so-called TBS-order (“ter beschikking stelling”). This type of TBS-order is reserved for patients who have been convicted of violent crimes for which a possible maximum sentence is 4 years in prison. For a more elaborate description of the TBS-system, see Klein Haneveld et al. (2021). The data from the respective hospitals were aggregated, as allocation to the different hospitals had been random and there were no mean differences between the patients from the different hospitals on nationality and IQ. The full range of possible PCL-R scores was included. However, the sample was not matched on administration and scoring of the PCL-R. For 87 females and 50 males the rating was not done in consensus, and partly based on file information only. See De Vogel and Lancel (2016) for more details about the data. For the current study, see our flowchart of sample composition (Figure 1). We used a sub-group of the matched sample described above, with 110 female and 147 male patients. Only those patients were selected whose PCL-R scores were scored in consensus by two raters, based on a combination of interview and file information. In this sub-group, there were no mean differences on year of birth and year of admission.

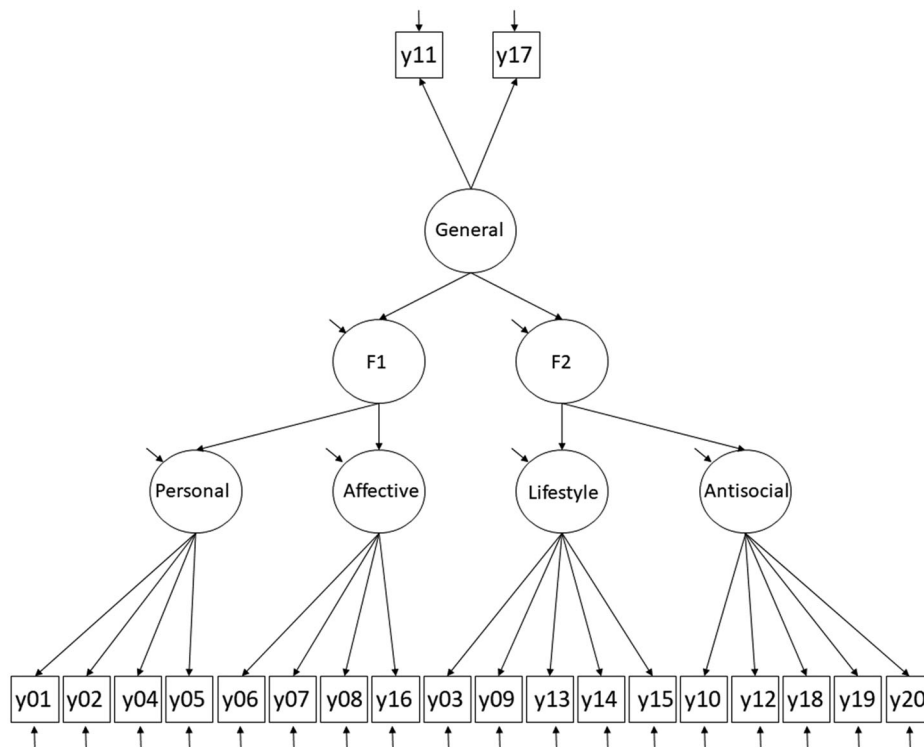


Figure 2. Graphical representation of the 4F-2S-1T model.

Note. "Personal" denotes the Interpersonal Factor, "F1" denotes the Interpersonal/Affective Factor, "F2" denotes the Social Deviance Factor, and "General" denotes the General Psychopathy Factor.

Instrument

The PCL-R (Hare, 1991, 2003) consists of 20 items, which can be scored 0 (*definitely does not apply*), 1 (*may apply or partly applies*), or 2 (*definitely applies*), leading to a maximum Total Score of 40. The rating is based on a review of extensive file information, preferably in combination with a semi-structured interview. Psychometric properties have been documented in the current manual (Hare, 2003). In previous research at one of the hospitals involved in this study, significant inter-rater reliability was established (single measure ICC = .88 for Total Score) for a partially overlapping sample (Hildebrand et al., 2002).

Analytic strategy

Factor structure of psychopathy

In this study, we rely on confirmatory factor analysis of the PCL-R which is, like the IRT approach by Bolt et al. (2004), a suitable approach to test for measurement invariance across gender (Putnick & Bornstein, 2016). To this end, a theory about the underlying PCL-R factor structure is required. At this point we note that in addition to a factor structure, psychopathy can also be represented as a network of symptoms (e.g., Verschuere et al., 2018). We consider a network perspective valuable as it can be used to clarify what the central features of psychopathy are. However, here we focus on quantifying individual differences using the PCL-R. Therefore, similar to, for example, Bolt et al. (2004) and Hare and Neumann (2008), we consider factor analysis a suitable approach.

Since the introduction of the PCL-R several factor structures have been proposed. In the one-factor model (Bolt et al., 2004), referred to as "1F", the only source of individual differences is assumed to be a General Psychopathy Factor. On the contrary, the Two-Factor model (Hare, 1991), referred to as "2F", identifies an Interpersonal/Affective Factor (Factor 1) and a Social Deviance Factor (Factor 2). In addition, in the Four-Facet model (e.g., Hare, 2003), referred to as "4F", a distinction is made between an Interpersonal, an Affective, a Lifestyle, and an Antisocial Facet. A hierarchical model has been proposed which combines the 4F with the 2F at the higher level. That is, the higher-order model contains four first-order facets and two second-order factors (Hare & Neumann, 2005) and is referred to as "4F-2S". In this model, the interpretation of the first- and second-order factors is the same as in the 4F and the 2F above. The idea of the General Psychopathy Factor in the 1F has also been added to the 4F, which results in a model with four first-order facets and a single second-order factor, referred to as "4F-1S" (Neumann et al., 2007). Again, the interpretation of these factors is the same as in the 4F and 1F.

The PCL-R contains two items (items 11 and 17) that do not explicitly measure one of the lower-order factors in the models above. However, to enable analyses of these two items we follow the idea presented by Bolt et al. (2004) and include the two items in an overarching higher-order factor. This factor thus includes all lower-order factors and items 11 and 17 as indicators. For model 1F, this simply means that the two items are included as indicators of the single

(and only) factor. For the 2F model, this implies that a higher-order factor is added that includes the 2 factors and items 11 and 17 as indicators. The resulting model thus coincides with the 2F-1S model. For the 4F-2S model, we added a third-order factor (making the model a 4F-2S-1T model) that includes the 2 second-order factors and items 11 and 17 as indicators. Thus, the models that we considered are the 1F, the 2F-1S, the 4F-1S, and the 4F-2S-1T model. For the 4F-2S-1T, a graphical representation of this model is depicted in Figure 2.

Note that we have chosen to use the full PCL-R, and have not included Cooke and Michie's Three-Factor model as a separate model. Testing the MI of the PCL-R according to the Cooke and Michie model would mean leaving out 7 of the 20 items. The seven items that fall outside the model are considered very relevant for clinical forensic practice, as they mostly address different forms of antisocial behavior. Furthermore, the three Cooke and Michie factors were incorporated in identical form into the more fine-grained Four-Facet model of the PCL-R. The present results using the four facets are informative about the three factors of Cooke and Michie. However, at the suggestion of a reviewer, we did verify our results in the hierarchical Three-Factor model; please see footnotes 3 and 4.

Identification of the models

To enable identification, we included a number of constraints. Specifically, in the 2F-1S model, the two second-order loadings (of the two first-order factors on the second-order factor) were constrained to be equal. In the 4F-2S-1T model, the first-order factors that load on the same second-order loadings were constrained to be equal (i.e., the second-order loadings of the first-order Facets 1 and 2 were constrained to be equal, and the second-order loadings of the first-order Facets 3 and 4 were constrained to be equal). In addition, the third-order loadings of second-order factor 1 and 2 on the third-order factor were also constrained to be equal. All models were identified by fixing the factor means to 0 and the factor variances to 1 within the male and female sample. For the model that fits the data best, we tested for measurement invariance across gender.

General procedure to test for measurement invariance across gender

For measurement invariance to be tenable, the factor model parameters discussed above should be the same in the male and female sample while allowing for group differences in the mean factor scores and the variance of the factor scores (Mellenbergh, 1989; Meredith, 1993). As discussed in the introduction section, most of the previous studies have compared the PCL-R across males and females in terms of psychometric properties like reliability, predictive validity, or classification rates. However, these psychometric properties do not account for possible differences in factor means and variances, therefore, differences in these properties are hard to interpret in terms of violations of measurement invariance.

To assess measurement invariance, we fitted explicit confirmatory factor models to the PCL-R data of the males and females in which factor model parameters are explicitly separated from the factor means and variances. Specifically, we considered the following increasingly restrictive models: For, *configural invariance* (Horn & McArdle, 1992), the factor loadings and item thresholds are free to vary across groups.¹ If this model shows acceptable fit to the data, it indicates that the same factor structure holds in both groups. For *metric invariance* (sometimes referred to as weak factorial invariance; Widaman & Reise, 1997), the factor loadings are constrained to be equal across groups whereas the factor variances are allowed to differ across groups (allowing free estimation of item thresholds across groups). Should the metric invariance model fit better than the configural model, this would indicate that the constraints imposed in the model (equal factor loadings across groups) are statistically tenable. In the third, *full measurement invariance* model, the item thresholds are constrained across groups and the factor means are allowed to freely differ across groups (Meredith, 1993). Note that in this final model, the factor variances are still allowed to differ across groups. When the full measurement invariance fits better than the metric model, it in turn reflects that the constraints imposed (equal item thresholds across groups) are statistically tenable.

The procedure outlined above can be repeated at the second-order level to establish invariance of the first-order factors (see Chen et al., 2005). In this procedure, the first-order factor means take over the role of the item thresholds above. That is, as a first step (second-order metric invariance), the second-order loadings are constrained and the second-order factor variances are freed across groups. In the next step (second-order full measurement invariance), the first-order factor means are constrained to be equal across groups while allowing for a mean difference on the second-order factor. Similarly, this procedure can be extended to test for measurement invariance at the level of the third-order factor. Although you can also test for higher-order variances to be equal across groups in principle (Chen et al., 2005) we followed Dolan et al. (2006) and left the higher-order residuals to be free across groups.

Model fit assessment and diagnosis of misfit

To assess model fit, various fit indices are available. Here, we used the Root Mean Squared Error of Approximation

¹Note that we treat the item scores explicitly as ordinal (which is equivalent to using the graded response model as Bolt et al., 2004, did.). As a result, similarly as in the graded response model, the factor models considered here do not include intercept parameters, but category thresholds (as there are three answer categories, each item contains two thresholds). In addition, the models considered do not contain residual variances. However, given appropriate identification restrictions (see Millsap & Yun-Tein, 2004) both the intercepts and the residual variances can be estimated and tested for invariance. Here however, we fixed the residual variances and intercepts to be equal across groups in all models, and we only freed specific residual variance parameters or intercepts parameters if this is indicated by the modeling results (i.e., poor model fit and/or large modification indices for the residual variances). Note that if the thresholds and factor loadings are shown to be invariant, it can be concluded that the intercepts and residual variances are also invariant.

Table 1. Fit indices for the four PCL-R models considered in this study.

Model	par	χ^2 (df)	RMSEA	CFI	TLI
1F	120	641.221 (340)	0.083 (0.073; 0.093)	0.842	0.823
2F-1S	122	522.771 (338)	0.065 (0.054; 0.076)	0.903	0.891
4F-1S	128	517.760 (332)	0.066 (0.055; 0.077)	0.902	0.888
4F-2S-1T ^a	126	484.981 (334)	0.059 (0.047; 0.071)	0.921	0.910

^aVariance of Facet 3 fixed to 0 in the female sample.

(RMSEA; Browne & Cudeck, 1989), the Comparative Fit Index (CFI; Bentler, 1990), and the Tucker-Lewis Index (Tucker & Lewis, 1973). For the RMSEA, it holds that values between 0.08 and 0.05 indicate acceptable model fit, and values smaller than 0.05 indicate good model fit. For CFI and TLI these cutoff values are 0.95-0.97 for acceptable model fit, and larger than 0.97 for good model fit (see Schermelleh-Engel et al., 2003).

In the process of fitting these series of models to data, it may be that a given restriction (e.g., the restriction of equal factor loadings) results in a deterioration in model fit indicating that at least one factor loading is not equal across groups. Such sources of misfit can be diagnosed using so-called “modification indices” which are available for each constrained parameter. In the case of a large modification index (a commonly used cutoff value is 10 as this is the default in Mplus; Muthén & Muthén, 1998–2017), the corresponding parameter (in this case the factor loading) can be freed across groups. If the resulting model fits better than the previous model (i.e., the model with all factor loadings to be free across groups), the measurement invariance tests may be continued. If for a minority of the items measurement invariance is violated, (i.e., a few items have factor loadings or item thresholds that are non-invariant across groups) this is referred to as partial measurement invariance (Byrne et al., 1989). The latent variable means and variances under partial measurement invariance can be meaningfully interpreted, but the non-invariant item thresholds and factor loadings do not contribute to the means and variances respectively.

All analyses were conducted in Mplus (Muthén & Muthén, 1998–2017). The observed item scores were explicitly treated as ordinal. Parameters were estimated using weighted least squares estimation. The resulting models are therefore not linear factor models but discrete factor models (Takane & De Leeuw, 1987; Wirth & Edwards, 2007).

Results

Factor models for the PCL-R

First, we fit the four competing PCL-R models to the data of the males and females. For the 4F-2S-1T, we fixed the variance of first-order Lifestyle Facet (Facet 3) in the female sample to 0 as this variance parameter approached 0 during parameter estimation. This implies, for this model, that the second-order Social Deviance Factor (F2) explains all the variance in the first-order Lifestyle Facet. Results concerning the fit of the four models is depicted in Table 1. As can be seen, all models but the one factor model showed acceptable fit according to the RMSEA statistic. However, according to the CFI and TLI values all models fit poorly. Of note, the

Table 2. Fit indices for the models to establish measurement invariance with respect to gender.

	Model	par	χ^2 (df)	RMSEA	CFI	TLI
1	Configural invariance ^a	126	484.980 (334)	0.059 (0.047; 0.071)	0.921	0.910
2	1 st -order loadings	111	484.593 (349)	0.055 (0.043; 0.066)	0.929	0.922
3	item thresholds	79	549.531 (381)	0.059 (0.047; 0.069)	0.911	0.912
3'	item thresholds ^b 1	83	520.061 (377)	0.054 (0.042; 0.065)	0.925	0.924
4	2 nd -order loadings	83	521.733 (377)	0.055 (0.043; 0.066)	0.924	0.923
5	1 st -order means	81	556.910 (379)	0.060 (0.049; 0.071)	0.907	0.906
5'	1 st -order means ^c	82	523.819 (378)	0.055 (0.043; 0.066)	0.923	0.923
6	3 rd -order loadings	80	537.171 (380)	0.057 (0.045; 0.068)	0.917	0.917
6'	3 rd -order loadings ^d	81	521.137 (379)	0.054 (0.042; 0.065)	0.925	0.925
7	2 nd -order means	77	559.161 (383)	0.060 (0.049; 0.070)	0.907	0.908
7'	2 nd -order means ^e	79	526.830 (381)	0.055 (0.043; 0.066)	0.923	0.924

^aVariance of Facet 3 fixed to 0 in the female sample. ^bItem thresholds of item 2, and item 10 free across groups. ^cMean of first-order Facets 3 and 4 free in the female sample, mean of second-order factor 2 fixed to 0 in the female sample. ^dFactor loading of item 11 free across groups. ^eItem thresholds of item 17 free across groups.

CFI and TLI are incremental fit indices, meaning that they measure the added value of the corresponding model to a baseline model in which all variables are uncorrelated. As generally correlations are medium to small in non-cognitive data (e.g., our average absolute inter-item correlation is 0.240), not much incremental fit can be expected. We therefore focused on the RMSEA, and inspected the CFI and TLI values only to compare subsequent models. As can be seen from the table, for the models considered, the Four-Facet model with two second-order factors and one third-order factor (4F-2S-1T) is associated with the best RMSEA (0.059). Additional inspection of the results for this model indicated that there are no obvious sources of misfit (the largest modification index was 11.582 for the residual covariance between items 1 and 2). Therefore, we accepted the 4F-2S-1T model and tested for measurement invariance with respect to gender in this model.

Measurement invariance analyses

The results concerning measurement invariance are in Table 2. As discussed above, we started with the configural invariance model (Model 1), which is the same as the 4F-2S-1T model in Table 1. Model fit was considered acceptable as judged by the RMSEA. Next, we fixed the first-order factor loadings to be equal across groups, while freeing the first-order facet variances in the female group. The resulting model (Model 2) yielded better fit than the configural invariance model in terms of RMSEA, CFI and TLI, indicating that the factor loadings are equal across gender. Next, we constrained the item thresholds to be equal across groups, while freeing the first-order facet means in the female sample. The resulting model (Model 3) fit worse as compared to the previous model as judged by the RMSEA, CFI, and TLI. The modification indices of the item threshold parameters indicated that the item threshold parameters of items 2 and 10 were a significant source of misfit. We freed these parameters resulting in Model 3'. The fit of this model is approximately the same as Model 2 indicating that except for the items 2 and 10 item thresholds were the same across genders. The standardized estimates of the thresholds

Table 3. Standardized parameter estimates for the non-invariant thresholds in Model 3'.

Item	Threshold	Estimate	SE	Z	P
Males					
2	1	-0.546	0.109	-5.002	0.000
	2	0.691	0.113	6.120	0.000
10	1	-1.181	0.136	-8.669	0.000
	2	-0.150	0.105	-1.421	0.155
Females					
2	1	0.321	0.143	2.242	0.025
	2	1.268	0.202	6.265	0.000
10	1	-1.812	0.192	-9.436	0.000
	2	-0.692	0.175	-3.962	0.000

of these two non-invariant items are depicted in Table 3. As can be seen, the thresholds for item 2 were smaller in the male group while for item 10, the thresholds were smaller in the female group.

As partial measurement invariance was now established with respect to the first-order facet structure, the first-order facet means and variances can now be meaningfully compared across groups. See Table 4 for the standardized parameters estimates of the first-order facet means and variance in the female sample. As the means in the males are all equal to 0 for identification purposes, the standardized estimates for the first-order facet means in the table can readily be interpreted as standardized mean differences between males and females. See Figure 3 for a graphical display of these differences. As can be seen, female patients scored significantly lower than their male counterparts on all facets, with the largest standardized effect on the Antisocial Facet (Facet 4) and the smallest effect on the Interpersonal Facet (Facet 1). In addition, the female sample showed less variance on all first-order facets.

Invariance of the second-order structure

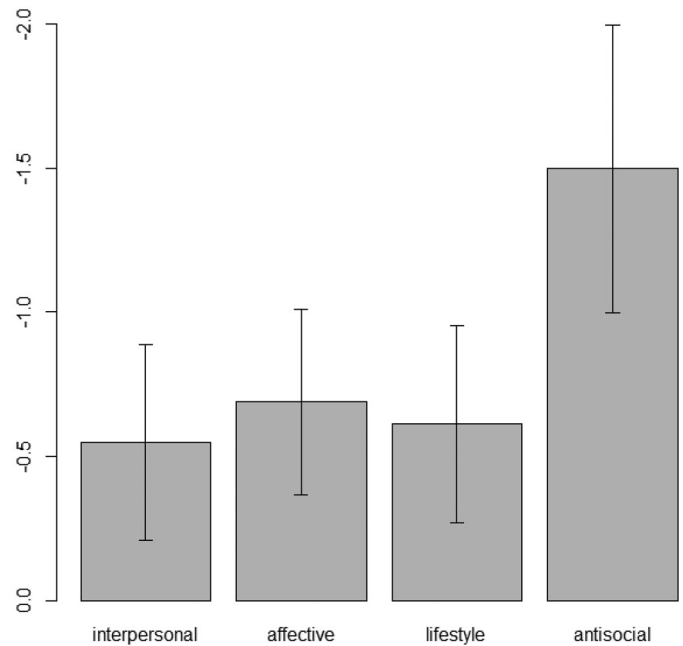
Next, in Model 3' we forced the second-order factor loadings to be equal across gender while freeing the variances of the two second-order factors (Model 4). As can be seen in Table 1 the fit of this model is identical to the previous model (Model 3') as the model contains the same number of parameters to model the correlations among the first-order factors.² We can thus accept Model 4 and fix the first-order means to be equal to 0 (i.e., equal across groups) while allowing for a mean difference on the second-order factors. The resulting model (Model 5) fits worse as compared to Model 4 as judged by the RMSEA, CFI, and TLI. The modification indices of the first-order factor mean parameters indicated that the means of the Lifestyle Facet (Facet 3) and the Antisocial Facet (Facet 4) cannot be explained by the mean difference in the second-order Social Deviance Factor only. We therefore freed the first-order facet means of the Lifestyle Facet and the Antisocial Facet and fixed the mean of second-order Social Deviance Factor to be equal to 0. The resulting model (Model 5') showed

²As the second-order factor loadings of first-order Facets 1 and 2 and the second-order factor loadings of Facets 3 and 4 are constrained to be the same, see above, we only had to fix 2 parameters, i.e., the two loadings, and we had to free 2 parameters, i.e., the second-order factor variances

Table 4. Standardized parameter estimates of the first-order facet means and variances in the female sample in Model 3'.

Factor	Estimate	SE	Z	P
Means				
1: Interpersonal Facet	-0.549	0.174	-3.163	0.002
2: Affective Facet	-0.689	0.164	-4.194	0.000
3: Lifestyle Facet	-0.612	0.175	-3.489	0.000
4: Antisocial Facet	-1.498	0.255	-5.885	0.000
Variances				
1: Interpersonal Facet	0.212	0.147	1.443	0.149
2: Affective Facet	0.457	0.111	4.097	0.000
3: Lifestyle Facet	0.000	-	-	-
4: Antisocial Facet	0.370	0.154	2.412	0.016

Note. Means in the male sample are fixed to 0 for all first-order facets. In addition, variances in the male sample are fixed to 1.00 for all first-order facets. The variance of first-order Facet 3 is fixed to 0 in the female sample as this variance approached 0 in model estimation.

**Figure 3.** Graphical display of the standardized mean differences in the first-order facets in Model 3'.

Note. Vertical lines represent 95% confidence intervals.

largely the same fit to the data as Model 4 with the RMSEA, and TLI being identical for Model 4 and 5', and with the CFI being 0.001 smaller for Model 5'. As this difference in CFI is negligible, and no misfit was evident from the modification indices (all modification indices are smaller than 10), we accepted Model 5'. In this model, the first-order facet means are thus modeled using one second-order mean parameter (standardized estimate: -0.816, SE: 0.193), and mean parameters for the first-order Lifestyle Facet (standardized estimate: -0.593, SE: 0.170) and the first-order Antisocial Facet (standardized estimate: -1.467, SE: 0.243). As indicated by the estimates of the first-order means, the females scores on these facets are smaller than that of the males.

Invariance of the third-order factor structure

In Model 5' we proceeded by fixing the third-order factor loading (note that we only have one third-order factor

loading as discussed above) and the factor-loadings of item 11 and 17 while allowing for a variance difference on the third-order general psychopathology factor. Note that item 11 and 17 were now also taken into account, as these items are direct indicators of the third-order factor. As can be seen in Table 2, the resulting model (Model 6) fit worse than the previous Model 5' in terms of all fit indices. The modification indices indicated that the misfit was mainly due to the factor loading of item 11. Freeing this parameter across groups resulted in a model (Model 6') that showed better than Model 5' for all fit indices. Result indicated that the factor loading of item 11 is smaller in the male group (standardized estimate: 0.405, se: 0.101) as compared to the female group (standardized estimate 0.735, se: 0.087). Finally, we equated the second-order means and the thresholds of item 11 and 17 to be equal across groups (Model 7). Model fit again deteriorated as compared to the previous Model 6'. The modification indices indicated that the misfit was mainly due to the thresholds of item 17. Freeing these parameters across groups resulted in Model 7' which has the same fit as Model 6'. However, as Model 7' contains fewer parameters, we accepted this model as the final model. In this final model, the thresholds for item 17 in the male group are: 0.859 (se: 0.127) and 1.189 (se: 0.145), while in the female group they are -0.105 (se: 0.157) and 0.202 (se: 0.168). That is, the thresholds were smaller in the female group, indicating that females scored higher on item 17 than would be expected on basis of their underlying third-order general psychopathology factor score. In the final Model 7', the standardized mean difference between males and females on the third-order General Psychopathy Factor equals -0.895 (SE: 0.198). Note that this mean difference is only based on the second-order Interpersonal/Affective Factor (F1) and on item 11, as the second-order Social Deviance Factor (F2) and item 17 were shown to have non-invariant means/thresholds.

Discussion

This study used confirmatory factor analysis within a structural equation modeling framework to test for measurement invariance of the PCL-R for gender, using 147 male and 110 female forensic psychiatric patients. The PCL-R was found to be only partially invariant for gender. More specifically, four items (i.e., items 2, 10, 11, and 17) violated measurement invariance.³ That is, for item 2 ("Grandiose sense of self-worth") the thresholds were larger in the female group than in the male group, indicating that females scored lower on these items than would be expected based on their position on the underlying first-order Interpersonal Facet. In addition, for items 10 ("Poor behavioral controls") and 17 ("Many short-term marital relationships"), the thresholds

were smaller in the female group indicating that females scored higher on these items than would be expected based on their position on the underlying facet or factor (which is the first-order Antisocial Facet for item 10, and the General Psychopathy Factor for item 17). With respect to item 11 ("Promiscuous sexual behavior"), we found that the factor loading of this item on the General Psychopathy Factor was larger in the female group as compared to the male group. This indicates that for females, item 11 was a stronger indicator of the General Psychopathy Factor. Moreover, the first-order Lifestyle and Antisocial Facets (Facets 3 and 4) were biased indicators of the second-order Social Deviance Factor (Factor 2).⁴ That is, the female scores on the Lifestyle and Antisocial Facets were smaller than what would be expected on the basis of their position on the underlying Social Deviance Factor. Thus, a female patient with a given position on Factor 2 will have a lower expected position on the Lifestyle and Antisocial Facets as compared to a male patient with exactly the same position on Factor 2. Hence, male and female forensic psychiatric patients cannot be meaningfully compared on the Social Deviance Factor, as they are not held to the same standard. For the first-order facet structure partial invariance was established, and accordingly male and female forensic psychiatric patients could be meaningfully compared on the first-order facets after removing items 2 and 10.

It is informative to compare our results to those of Bolt and colleagues (2004) in a group of incarcerated females. At item level our findings do not overlap, as Bolt et al. found differential item functioning in items 5, 12, 18, and 20, whereas we found items 2, 10, 11, and 17 to violate measurement invariance. Possible explanations for these differences are variations in sample (a criminal justice population versus forensic psychiatric patients) and the use of unmatched versus matched comparison groups. However, at the level of the Four-Facet model our findings are remarkably similar, with Facets 3 and 4 (Lifestyle and Antisocial) not functioning well. This may be a possible explanation for the finding that Factor 2, comprising of Facets 3 and 4 and predictive of violence in men (Yang et al., 2010), has not been consistently found to predict violence in women. Possibly, prevalence and predictive validity with regard to violent behavior will improve when the non-invariant items are removed. This would be an interesting follow-up study. Of Factor 1, in both studies only one item violated measurement invariance; item 5 ("Conning/manipulative" in Bolt et al., 2004) and item 2 (Grandiose sense of self-worth in our study). Both are part of Interpersonal Facet 1. Facet 2, relating to defective emotional functioning, often considered one of the core aspects of psychopathy, was found to be invariant. Previously, some authors have speculated whether the construct of psychopathy in females is inherently distinct

³As suggested by a reviewer, we verified these results in the Cooke and Michie (2001) second-order Three-Factor model. Similar to our third-order Four-Facet model, we found item 2 to violate threshold invariance in the Cooke and Michie model. The other items that we found to be non-invariant in the third-order Four-Facet model, items 10, 11, and 17, are not part of the Three-Factor model. The results are available upon request.

⁴In the Cooke and Michie (2001) second-order Three-Factor model (see previous footnote), Factor 3 was found to be unbiased with respect to the second-order factor. However, note that the second-order General Psychopathy Factor in the Cooke and Michie model is statistically and substantively different from our second-order Social Deviance Factor in the third-order Four-Facet model.

from psychopathy in males (Miller et al., 2011; Viljoen et al., 2015; Wynn et al., 2012). A tentative conclusion on the basis of these two studies would be that the personality features of psychopathy (the interpersonal and affective traits of Facets 1 and 2) are manifested in a sufficiently comparable way in males and females, while the behavioral features (impulsive, irresponsible lifestyle and antisocial behavior) are subject to gender differences. We feel that, although the PCL-R was (again) found not to be fully invariant with respect to gender, it does appear to capture the core features of psychopathy in women.

To speculate on the origin of the partial lack of MI, it may stand to reason that female forensic patients with narcissistic and callous features express their impulsive, irresponsible lifestyle and antisocial behavior systematically different than men with the same trait levels. For instance, several authors have speculated that, due to biological differences, the use of physical force or violence to achieve a desired outcome may be considerably less feasible and effective for women, whereas the manipulative use of flirtation, intimacy or sexual favors may work well to attain their personal agenda (Forouzan & Cooke, 2005; Nicholls & Petrila, 2005; Wynn et al., 2012). The literature which focuses specifically on gender differences in psychopathic behavior still appears to be rather sparse. An early review of behavioral gender differences (among other aspects of psychopathy) by Forouzan and Cooke (2005) reported that impulsivity and conduct disorder in females were characterized by self-harming behavior, manipulation, and complicity in theft and fraud, as opposed to violent behavior in males. This was recently replicated by De Vogel and Lancel (2016). More research is needed that explores typical behaviors associated with psychopathy in women.

If we accept the finding that the PCL-R does not optimally capture female behavioral manifestations, what are we to do with the assessment of females with psychopathic features? One possibility would be to uphold the items of Facets 1 and 2 (Factor 1), and to formulate a new female version of Facets 3 and 4 (Factor 2), based on female manifestations of impulsive, irresponsible, parasitic and antisocial behavior. For example, we found Item 11 (Promiscuous Sexual Behavior) to be more strongly related to the General Psychopathy Factor for females than for males. Item 11 describes sexual behavior that is impersonal and trivial, as reflected in frequent casual contacts, infidelities, and prostitution. In male populations Item 11 does not load on any facets or factors. However, for females this may well be a good candidate for the behavioral features typically associated with Factor 2. A less drastic alternative to rewriting all Factor 2 items would be to develop supplementary guidelines to the existing manual. In a series of studies, Morrissey and colleagues investigated the applicability of the PCL-R in individuals with intellectual disabilities in secure settings (Morrissey, 2003; Morrissey et al., 2005, 2007a, 2007b). The authors developed such supplementary guidelines for all PCL-R items, while maintaining the flavor or intent of the original items (Morrissey, 2007). Where appropriate, they expanded item descriptions to include examples of behavior

especially relevant for individuals with intellectual disabilities. For some items, the criteria for evidence of the behavior involved were slightly broadened, or different sources of information were added. In a validation study in three forensic settings ($N=203$), use of the supplementary guidelines in combination with the PCL-R showed adequate internal consistency ($\alpha = .82$) and good interrater reliability ($ICC = .89$) (Morrissey et al., 2005). What would this idea mean for the items we found to be non-invariant for females? First, female forensic psychiatric patients scored lower on item 2 (Grandiose Sense of Self-Worth) than would be expected based on their position on the underlying Interpersonal Facet. Possibly, the current phrasing of the item does not adequately capture female narcissistic manifestations. Perhaps, providing additional examples of grandiosity in females, possibly less overt and brazen, more subtle and covert, may improve the scoring of this item. Second, female psychiatric patients scored higher than would be expected on items 10 (Poor Behavioral Controls). Apparently, we tend to judge females more harshly when it comes to aggressive behavior, while we are more tolerant of such behavior in men. More elaborate specification of what type of behavior is needed for a score of 1 or 2 in females may improve veridical assessment. Finally, females also scored higher than would be expected on item 17 (Many Short-term Marital Relationships). The item description in this case precisely defines the number of live-in relationships needed for a score of 1 or 2, also based on age (under 30, or 30 and above). These norms could quite easily be adjusted for females.

There are a number of notable strengths and limitations to this study. First and foremost, as far as we are aware, our study drew on the largest published database of matched male and female forensic psychiatric patients. The principal limitations are inherent to the matched control design: although the case matching was conducted with great care, it is impossible to preclude hidden confounding variables. This limitation is especially relevant to the current study, as we have used a subset of the matched sample based on (homogeneity in) administration and scoring. Also, in the original matched sample, it was not always possible to match a female case with a male case from the same hospital. Thus, the majority of male cases came from the Van der Hoevenkliniek. It is not immediately clear how this may have affected the findings. Second, our study participants consisted of forensic psychiatric patients, all part of the Dutch TBS-measure. Therefore, the results are drawn from and hence applicable to a very specific forensic population, not necessarily generalizable to other populations. Also, at item level, findings were disparate from those reported by Bolt et al. (2004). Replication and testing the generalizability to other types of samples is warranted. Thirdly, based on the current study, the practical consequences of the partial MI remain unclear. Female forensic patients may, for example, have been unduly burdened with, or escaped consequences of, partly unreliable psychopathy assessments. Another consequence of partial MI may be the inconsistency in predictive validity of the PCL-R in males and females. In

several studies that have examined the predictive validity of the PCL-R for re-offending and/or institutional violence in women (De Vogel et al., 2019; De Vogel & Lancel, 2016; Geraghty & Woodhams, 2015; Loucks & Zamble, 2000; Richards et al., 2003; Salekin et al., 1998; Weizmann-Henelius et al., 2015), non-significant to modest relationships were found. Future studies will have to investigate these issues.

A final point of discussion is whether the sample size in our study (110 female and 147 male patients) is sufficient for measurement invariance analyses. There are no sample size recommendations for testing measurement invariance available that can be directly applied to our study. For instance, Meade and Bauer (2007) pointed out that a sample size of 100 participants per group is sufficient to detect violations of metric invariance, but their study focused on a Three-Factor model with standardized factor loadings around 0.8 and 20 continuous items. Thus, we did meet the requirement of 100 participants per group, but we used a more complex model and our standardized factor loadings were between 0.4 and 0.9. As we did find violations of measurement invariance in the present study, it can be concluded that sample size is at least not too small, but it cannot be ruled out that we missed some small effects on items with small factor loadings. However, these effects are of less practical significance, and the effects that we did find can be trusted.

In sum, to provide a straightforward answer to the question we posed in the title of our article (i.e., Do we hold males and females with psychopathy to the same standard?), we have to conclude that the PCL-R in its current form does not fully attain this essential outcome. On the other hand, it seems reasonable to expect that specific scoring adjustments (especially with respect to the Social Deviance Factor) might go a long way in bringing about more equivalent assessment of psychopathic features in men and women. Only then can we meaningfully compare the genders on the prevalence, structure, and external correlates of psychopathy.


Disclosure statement

We have no conflicts of interests to disclose.

ORCID

Evelyn Klein Haneveld  <http://orcid.org/0000-0003-1222-5920>

Dylan Molenaar  <http://orcid.org/0000-0002-7168-3238>

Vivienne de Vogel  <http://orcid.org/0000-0001-7671-1675>

Wineke Smid  <http://orcid.org/0000-0001-6440-5232>

Jan H. Kamphuis  <http://orcid.org/0000-0002-4050-0697>

Public significance and data availability statement

This study shows that the most widely used and validated instrument for the assessment of psychopathy, the PCL-R, needs modification to adequately assess psychopathy in females. Adequate assessment is a prerequisite for consistent and informative research on psychopathy in women, its societal consequences and role in judicial systems, as well as the development of effective interventions.

The data that support the findings of this study are available on request from the corresponding author. It is up to the author to

determine whether a request is reasonable. The data are not publicly available due to privacy restrictions.

References

- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- Beryl, R., Chou, S., & Völlm, B. (2014). A systematic review of psychopathy in women within secure settings. *Personality and Individual Differences*, 71, 185–195. <https://doi.org/10.1016/j.paid.2014.07.033>
- Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multi-group item response theory analysis of the Psychopathy checklist-revised. *Psychological Assessment*, 16(2), 155–168. <https://doi.org/10.1037/1040-3590.16.2.155>
- Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62–83. https://doi.org/10.1207/s15327906mbr2404_4
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456–466. <https://doi.org/10.1037/0033-2909.105.3.456>
- Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(3), 471–492. https://doi.org/10.1207/s15328007sem1203_7
- Cooke, D. J., & Michie, C. (2001). Refining the construct of psychopathy: Towards a hierarchical model. *Psychological Assessment*, 13(2), 171–188. <https://doi.org/10.1037/1040-3590.13.2.171>
- De Vogel, V., Bruggeman, M., & Lancel, M. (2019). Gender-sensitive violence risk assessment: Predictive validity of six tools in female forensic psychiatric patients. *Criminal Justice and Behavior*, 46(4), 528–549. <https://doi.org/10.1177/0093854818824135>
- De Vogel, V., & Lancel, M. (2016). Gender differences in the assessment and manifestation of psychopathy: Results from a multicenter study in forensic psychiatric patients. *International Journal of Forensic Mental Health*, 15(1), 97–110. <https://doi.org/10.1080/14999013.2016.1138173>
- De Vogel, V., Stam, J., Bouman, Y. H., Ter Horst, P., & Lancel, M. (2016). Violent women: A multicentre study into gender differences in forensic psychiatric patients. *The Journal of Forensic Psychiatry & Psychology*, 27(2), 145–168. <https://doi.org/10.1080/14789949.2015.1102312>
- Dillard, C. L., Salekin, R. T., Barker, E. D., & Grimes, R. D. (2013). Psychopathy in adolescent offenders: An item response theory study of the antisocial process screening device-self report and the Psychopathy Checklist: Youth Version. *Personality Disorders*, 4(2), 101–120. <https://doi.org/10.1037/a0028439>
- Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & Van De Sluis, S. (2006). Multi group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, 34(2), 193–210. <https://doi.org/10.1016/j.intell.2005.09.003>
- Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (2017). Personality in general and clinical samples: Measurement invariance of the Multidimensional Personality Questionnaire. *Psychological Assessment*, 29(9), 1111–1119. <https://doi.org/10.1037/pas0000408>
- Forouzan, E., & Cooke, D. J. (2005). Figuring out la femme fatale: Conceptual and assessment issues concerning psychopathy in females. *Behavioral Sciences & the Law*, 23(6), 765–778. <https://doi.org/10.1002/bsl.669>
- Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *Hare psychopathy checklist: Youth version*. Multi-Health Systems.
- Geraghty, K. A., & Woodhams, J. (2015). The predictive validity of risk assessment tools for female offenders: A systematic review. *Aggression and Violent Behavior*, 21(2), 25–38. <https://doi.org/10.1016/j.avb.2015.01.002>

- Hare, R. D. (1991). *Hare Psychopathy Checklist-Revised*. Multi-Health Systems.
- Hare, R. D. (2003). *Hare Psychopathy Checklist-Revised (PCL-R)* (2nd ed.). Multi-Health Systems.
- Hare, R. D., & Neumann, C. S. (2005). Structural models of psychopathy. *Current Psychiatry Reports*, 7(1), 57–64. <https://doi.org/10.1007/s11920-005-0026-3>
- Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annual Review of Clinical Psychology*, 4, 217–246. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091452>
- Hare, R. D., Neumann, C. S., & Mokros, A. (2018). The PCL-R assessment of psychopathy: Development, properties, debates, and new directions. In C. J. Patrick (Ed.), *Handbook of Psychopathy* (2nd ed, pp. 509–528). The Guilford Press.
- Hildebrand, M., de Ruiter, C., de Vogel, V., & van der Wolf, P. (2002). Reliability and factor structure of the Dutch language version of Hare's Psychopathy Checklist-Revised. *International Journal of Forensic Mental Health*, 1(2), 139–154. <https://doi.org/10.1080/14999013.2002.10471169>
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3-4), 117–144. <https://doi.org/10.1080/03610739208253916>
- Klein Haneveld, E., Smid, W., Timmer, K., & Kamphuis, J. H. (2021). Clinical appraisals of individual differences in treatment responsiveness among patients with psychopathy: A Consensual Qualitative Research study. *Criminal Justice and Behavior*, 48(8), 1031–1051. <https://doi.org/10.1177/0093854820970597>
- Leistico, A. M. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law and Human Behavior*, 32(1), 28–45. <https://doi.org/10.1007/s10979-007-9096-6>
- Loucks, A. D., Zamble, E. (2000). *Predictors of criminal behaviour and prison misconduct in serious female offenders*. Retrieved November 15, 2020, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.559.5450&rep=rep1&type=pdf>
- Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 611–635. <https://doi.org/10.1080/10705510701575461>
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2), 127–143. [https://doi.org/10.1016/0883-0355\(89\)90002-5](https://doi.org/10.1016/0883-0355(89)90002-5)
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Miller, J. D., Watts, A., & Jones, S. E. (2011). Does psychopathy manifest divergent relations with components of its nomological network depending on gender? *Personality and Individual Differences*, 50(5), 564–569. <https://doi.org/10.1016/j.paid.2010.11.028>
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4
- Morrissey, C. (2003). The use of the PCL-R in forensic populations with learning disability. *The British Journal of Forensic Practice*, 5(1), 20–24. <https://doi.org/10.1108/14636646200300003>
- Morrissey, C. (2007). *Guidelines for assessing psychopathy in offenders with intellectual disabilities using the PCL-R and PCL:SV*. Retrieved November 15, 2020, from https://www.researchgate.net/publication/257890698_Guidelines_for_use_of_the_PCL-R_and_SV_in_adults_with_ID
- Morrissey, C., Hogue, T., Mooney, P., Allen, C., Johnston, S., Hollin, C., Lindsay, W. R., & Taylor, J. L. (2007a). The predictive validity of the PCL-R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry & Psychology*, 18(1), 1–15. <https://doi.org/10.1080/08990220601116345>
- Morrissey, C., Hogue, T. E., Mooney, P., Lindsay, W. R., Steptoe, L., Taylor, J., & Johnston, S. (2005). Applicability, reliability, and validity of the Psychopathy Checklist-Revised in offenders with intellectual disabilities: Some initial findings. *International Journal of Forensic Mental Health*, 4(2), 207–220. <https://doi.org/10.1080/14999013.2005.10471225>
- Morrissey, C., Mooney, P., Hogue, T., Lindsay, W. R., & Taylor, J. (2007b). Predictive validity of the PCL-R for offenders with intellectual disability in a high security hospital: treatment progress. *Journal of Intellectual & Developmental Disability*, 32(2), 125–134. <https://doi.org/10.1080/13668250701383116>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Nicholls, T. L., & Petrila, J. (2005). Gender and psychopathy: An overview of important issues and introduction to the special issue. *Behavioral Sciences & the Law*, 23(6), 729–741. <https://doi.org/10.1002/bsl.677>
- Neumann, C. S., Hare, R. D., & Newman, J. P. (2007). The superordinate nature of the Psychopathy Checklist-Revised. *Journal of Personality Disorders*, 21(2), 102–117. <https://doi.org/10.1521/pe.2007.21.2.102>
- Ogloff, J. R., Wong, S., & Greenwood, A. (1990). Treating criminal psychopaths in a therapeutic community program. *Behavioral Sciences & the Law*, 8(2), 181–190. <https://doi.org/10.1002/bsl.2370080210>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Richards, H. J., Casey, J. O., & Lucente, S. W. (2003). Psychopathy and treatment response in incarcerated female substance abusers. *Criminal Justice and Behavior*, 30(2), 251–276. <https://doi.org/10.1177/2F0093854802251010>
- Salekin, R. T., Rogers, R., Ustad, K. L., & Sewell, K. W. (1998). Psychopathy and recidivism among female inmates. *Law and Human Behavior*, 22(1), 109–128. <https://doi.org/10.1023/A:1025780806538>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408. <https://doi.org/10.1007/BF02294363>
- Tsang, S., Schmidt, K. M., Vincent, G. M., Salekin, R. T., Moretti, M. M., & Odgers, C. L. (2015). Assessing psychopathy among justice involved adolescents with the PCL:YV: An item response theory examination across gender. *Personality Disorders*, 6(1), 22–31. <https://doi.org/10.1037/per0000094>
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1–10. <https://doi.org/10.1007/BF02291170>
- Verschueren, B., van Ghesel Grothe, S., Waldorp, L., Watts, A. L., Lilienfeld, S. O., Edens, J. F., Skeem, J. L., & Noordhof, A. (2018). What features of psychopathy might be central? A network analysis of the Psychopathy Checklist-Revised (PCL-R) in three large samples. *Journal of Abnormal Psychology*, 127(1), 51–65. <https://doi.org/10.1037/abn0000315>
- Verona, E., & Vitale, J. (2018). Psychopathy in women: Assessment, manifestations, and etiology. In C. J. Patrick (Ed.), *Handbook of psychopathy* (2nd ed., pp. 509–528). The Guilford Press.
- Viljoen, S., Cook, A. N., Lim, Y. L., Layden, B. K., Bousfield, N. K., & Hart, S. D. (2015). Are psychopathic and borderline personality disorder distinct, or differently gendered expressions of the same disorder? An exploration using concept maps. *International Journal of Forensic Mental Health*, 14(4), 267–279. <https://doi.org/10.1080/14999013.2015.1114535>
- Vitale, J. E., & Newman, J. P. (2001). Using the Psychopathy Checklist-Revised with female samples: Reliability, validity, and implications for clinical utility. *Clinical Psychology: Science and Practice*, 8(1), 117–132. <https://doi.org/10.1093/clipsy.8.1.117>
- Weizmann-Henelius, G., Virkkunen, M., Gammelgård, M., Eronen, M., & Putkonen, H. (2015). The PCL-R and violent recidivism in a prospective follow-up of a nationwide sample of female offenders. *The Journal of Forensic Psychiatry & Psychology*, 26(5), 667–685. <https://doi.org/10.1080/14789949.2015.1049192>

- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. <https://doi.org/10.1037/10222-009>
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79. <https://doi.org/10.1037/1082-989X.12.1.58>
- Wong, S. C. P., & Hare, R. D. (2005). *Guidelines for a psychopathy treatment program*. Multi-Health Systems Inc.
- Wynn, R., Høiseth, M. H., & Pettersen, G. (2012). Psychopathy in women: Theoretical and clinical perspectives. *International Journal of Women's Health*, 4, 257–263. <http://dx.doi.org/10.2147/IJWH.S25518>
- Yang, M., Wong, S. C. P., & Coid, J. W. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, 136(5), 740–767. <https://doi.org/10.1037/a0020473>