CTT and IRT analysis of the Triarchic Psychopathy Measure

Yeong Hwangbo (6147172)

Eberhard Karls Universität Tübingen

Item response theory (QDS-PS3)

Dr. Stefano Noventa

April 20, 2024

CTT and IRT analysis of the Triarchic Psychopathy Measure

In measurement theory, there are two most used approaches to test development and its analysis: *Classical Test Theory (CTT)* and *Item Response Theory (IRT)*. CTT, which is not a theory but a collection of different theories, is a traditional method of analyzing tests. CTT has been used as a fundamental theory on which many test developments and measurements are based. CTT has been widely used and applied up until now in many fields such as Educational Measurement and Evaluation or psychological test development by virtue of the fact that it requires weak assumptions, which increases applications of CTT. However, as IRT addresses and solves the shortcomings of CTT (e.g., local independence and flexibility of IRT), IRT has been dominating as a modern approach to a measurement theory with its superiority (Bichi & Talib, 2018).

CTT and IRT are different from each other in many aspects, having their own advantages and disadvantages over one another. The biggest difference is the units of analysis. In CTT, analyses are done by a total score of a test on a test-level, accordingly unit of analysis is a whole test. Therefore, items are exchangeable, and the entire test is considered instead of each item (Hoffman, 2014). The total score in CTT consists of the sum of two components: True score (T) + Measurement error (E). True score is what has to be estimated in CTT, which is an unobserved and unknown mean of score distribution obtained from replicated administration of a test under the same condition (Hoffman, 2014). The other component, measurement error, is also unknown and constant across the test. Measurement error is assumed to be normally distributed with the mean of zero, uncorrelated with true score (Hoffman, 2018). Also, there is no correlation between errors from different tests, meaning that measurement error from one test does not affect the measurement error from another test. Due to the weak assumption and simple mathematical analyses, CTT has been widely used and applied (Idaka & Idaka, 2014).

IRT, which is also referred to as latent trait theory or strong true score theory, analyzes data on item-level (Idaka & Idaka, 2014). Therefore, the unit of analysis in IRT is individual items, and the item-ability relationship is expressed by the *Item characteristic curve (ICC)*. In IRT models, true score is replaced with ability scores. One of the advantages of IRT over CTT is that item difficulty and a person's ability can be compared because they are on the same latent metric (Bichi & Talib, 2018), while in CTT, comparison of scores from different measures are possible only when tests are parallel. This property in IRT is called conjoint scaling (Hoffman, 2018), which enables researchers to perform measurement

equivalence tests across groups or to do comparisons of persons measured in different groups with different items on the same scale (Bichi & Talib, 2018; Embretson & Reise, 2000). This can also be seen as a result of equating, which occurs automatically when person and item parameters are linked to the same metric (Embretson & Reise, 2000). Equating, again, allows comparison of parameters without making assumptions of the score distribution, whereas CTT requires an assumption about population score distributions (Embretson & Reise, 2000). Additionally, IRT property of comparing item responses of different scales measuring the same latent trait is useful in creating item banks (Idaka & Idaka, 2014). Especially, this facilitates to development large-scale testing programs such as computerized adaptive testing (CAT), where a person's ability is used to tailor the exam, and test difficulty gets modified each time depending on whether a person answered correctly (Bichi & Talib, 2018; Idaka & Idaka, 2014).

Another big advantage is that item characteristics in IRT are sample-independent, also called as invariance item parameters. This means that person and item parameters of IRT models do not change when different sample or test forms are used (Bichi & Talib, 2018; Idaka & Idaka, 2014). This property gives IRT flexibility, which enables IRT to investigate the contribution of each item of a test when different samples are used (Bichi & Talib, 2018). Thus, developing tests based on IRT is more beneficial since it examines and analyzes each item separately (Udoudoh & Umoobong, 2016). Item parameters in CTT, however, are sample dependent, meaning true score and item statistics change depending on test takers and test context. This is a shortcoming of CTT because item statistics are not generalizable to a setting where similar tests are administrated or different samples take similar tests (Krishnan, 2013). That is, item statistics work only under the same conditions where they were first estimated. IRT produces significantly less measurement error than CTT as well as allows items to have different response categories (Idaka & Idaka, 2014). These are beneficial for parameter estimation. However, CTT requires all items on the scale to have the same response categories, and measurement error is group-dependent and constant across groups (Embretson & Reise, 2000; Jabrayilov et al., 2016). Unlike CTT, measurement error in IRT is independent of groups, and thus can be estimated for a single individual based on item parameters (Embretson & Reise, 2000).

Although IRT demonstrated its superiority and usefulness in many different situations, CTT has its own strengths that have made CTT mainstream for a long time. CTT requires simpler mathematical analyses and weaker assumptions compared to those in IRT, increasing its use and application (Idaka &

Idaka, 2014). Accordingly, this serves as a shortcoming in IRT. IRT models require more complex procedures and statistical techniques for analyses (Idaka & Idaka, 2014). In contrast to IRT, parameter estimation in CTT is straightforward and comprehensible, due to its simplicity. Another advantage of CTT is that CTT requires a relatively small sample size for item parameter estimation. 200 to 500 samples are needed in CTT, while IRT requires a larger sample size, over 500 in general (Hambleton & Jones, 1993).

Despite some drawbacks to IRT, IRT has been dominating CTT as a modern approach to test theory. IRT assumes a non-linear relationship between latent trait and item response, whereas, in CTT, linear relationships are assumed between a total score and a true score that represents ability. That is, if a total score on a test increases, one's ability also increases. However, this is not necessarily true, as there is an exception that one with higher ability gets a low test score. Thus, considering the nature of the data being analyzed, assuming non-linear models appears to be more suitable for analyses of test and item responses. Also, it is advantageous that shorter test is considered more reliable in IRT, and allows mixed item formats that produce optimal test scores (Embretson & Reise, 2000). This is not the case in CTT because longer tests are preferred, and mixed item formats have an unbalanced impact on test scores (Embretson & Reise, 2000).

In CTT and IRT, item parameters are defined in different way. CTT specifies difficulty as a proportion of correct responses of each item, also called p-value in item analysis in CTT. IRT explains the difficulty of an item (d) using ICC. It is often referred to as a location parameter because in ICC, difficulty indicates a location on the ability scale that corresponds to a 0.5 probability of answering the item correctly. Item difficulties are also different in terms of the range of values. In most cases, item difficulty in IRT ranges from -3 to 3, with values around 0 (-0.5 - 0.5) being moderately difficult items, and higher values indicating difficult items (Bichi & Talib, 2018). In CTT, on the other hand, the interpretation of difficulty value is the opposite of that in IRT. Items with higher values indicate an easy item in the range of 0 to 1 (Bichi & Talib, 2018).

As an index of item discrimination power in CTT, a biserial correlation is computed and can be obtained in two ways: item-total correlation or item-rest correlation. It ranges from -1 to 1 as it is a correlation. A large negative biserial correlation implies that students with high test scores answer the item incorrectly and students who answer the item correctly score high on the test (Varma, 2006). Items with low biserial correlations are considered to be problematic, and a value above .25 is recommended to be

good test items. Item discrimination in IRT, often called the *a* parameter, is expressed by the slope of ICC, which corresponds to factor loadings in latent trait models. Items have higher discrimination values when the slope of ICC is steeper. Items with good discrimination power have values 0.5 to 2, which means items can discriminate well among participants with different abilities (Bichi & Talib, 2018). In general, a discrimination value above 1 is desirable for good test items, but a value above 0.75 is also acceptable (Bichi & Talib, 2018).

CTT and IRT also differ from each other in terms of reliability. In CTT, several reliability indices such as alpha, KR20, or omega are reported as reliability estimates. For the general formula for reliability, a ratio of the variance of a true score to the variance of a total score : Var(T) / Var(Y) is used. On the other hand, the concept of reliability in IRT is extended to the degree to which measurement error is free (Bichi & Talib, 2018). Measurement precision in IRT is explained as information (Hoffman, 2018). There are two types of information in IRT: *Test Information Function (TIF)* and *Item Information Function (IIF)*. Based on TIF and IIF, a reliability estimate is obtained by calculating conditional standard error of measurement (Bichi & Talib, 2018). This allows selecting items that best measure abilities at each level (Bichi & Talib, 2018).

The purpose of this essay is to investigate how the results of CTT analysis and IRT analysis are different from each other in terms of values of item parameters and interpretation of results by analyzing the dataset. A dataset containing participants' responses to the items measuring three psychopathy constructs was used for the analysis. The analyses started with data preprocessing, analyses of dimensionality and reliability were followed by estimation of item parameters with different packages in R. Finally, measurement invariance and differential item functioning (DIF) were investigated to assess the equivalence of measurement across groups that are chosen to be analyzed.

Methods

Data

Participants' item responses to Triarchic Psychopathy Measure (TriPM) were used for the analyses. Triarchic Psychopathy Measure is an inventory assessing the three constructs of the triarchic model of psychopathy, which encompasses Boldness, Disinhibition, and Callousness (Meanness) (Patrick et al., 2009). Each scale consists of 6 to 9 subscales, also called facets, with items assigned to one of these three scales. The responses to 58 four-point Likert items were collected by 1,678 participants.

Process

In this essay, the comparisons of results of IRT and CTT analyses were discussed. With the dataset preprocessed, an analysis of dimensionality was conducted to examine the factor structure that best fits the given dataset by doing a reliability analysis. The reliability estimates for each factor structure were considered together as an indicator of how well a factor structure represented the data, and provided insights into determining factor structures. Then, considering the previous research suggesting alternative factor structures, exploratory factor analysis (EFA) was performed using omega estimates. Thirdly, item parameters were estimated under both CTT framework and IRT framework. Item parameters estimated from IRT models with different packages in R were compared to see how values change with the packages. Based on the results, the fit and misfits of the items were discussed. Finally, to examine if there are significant differences between groups with regard to specific variables, measurement invariance was investigated, along with DIF analysis.

Results

Data preprocessing and data description

Missing data

The dataset was loaded with missing data and reverse-coded items included. It consists of 62 columns, where the first four columns indicate demographic questions such as `SEX` and `AGE`, and the rest indicate 58 items measuring the three traits. Firstly, to decide how to treat the missing data, '999's standing for missing values were converted to "NA". It was shown that each of the first four columns and each of the 58 item columns had the same number of missing values. 22 missing values were for the first four columns, 242 for items 1 to 30, and 255 for items 31-58, exhibiting the same pattern of missing values for individuals. The number of missing values was mostly 58, and 59 and 61 in some cases, and missing values of 28 were found in only a few cases. This means that in most of the missing cases, individuals did not answer to all 58 items. For cases with 59 and 61 missing responses, items were not answered including one of the four demographic questions. 28 missing cases can be interpreted as ones where the entire last part of the questionnaire (items 31 - 58) was left blank, which matches `summary(tripm.NA)` that there were more missing cases ($N = 252$) in items 31 - 58 than in items 1 - 30 ($N = 242$). As the specific pattern in missing values was observed where a row (i.e., an individual) containing missing values in any item has missing

values in the other items as well, it is reasonable to exclude the rows with missing values from the analysis. Thus, rows containing missing values were dropped from the data.

Reverse-coding

`ID`, `SEX`, and `PSYCH PROB` were converted to factors as they are categorical variables. `AGE` were converted to integers to be treated as numerical values. Then, reverse-coded items were recoded by flipping the order of the response categories. In this analysis, "+" keying items such as item 1 and item 3 were reverse-coded to "–" keying items so that choosing higher categories indicates a higher possibility of one having a trait being measured by the item (Patrick, 2009).

Descriptive statistics

As a consequence of dropping rows with missing values, the dataset contained 1,423 individuals (810 males and 623 females). `PSYCH PROB` indicates that approximately one-third of the participants ($N = 400$) have asked for help with psychological issues. Ages range widely from 16 to 89 with a mean of 31, and more than half the participants are under the age of 30.

Furthermore, the frequency of each response category of items was examined on the original categorical data. There was a specific response pattern where 1 was the most frequently answered category for almost all items, and for each response category, as moving to next categories, decreasing frequencies were observed. This indicates most of the participants answered the response categories of 1, 2, or 3 and resulted in a relatively low overall score, implying they are less associated with boldness, disinhibition, and callousness.

Dichotomization

For analyses, the items were dichotomized to 0 and 1. In order for the meaning of dichotomized item scores to be consistent with that of the original 4-point scale, items were dichotomized in a way that 0 indicated lower response categories (i.e., 1 and 2), and 1 indicated higher response categories (i.e., 3 and 4), so that higher response categories are still indicative of having the traits related to psychopathy.

Reliability and dimensionality

To determine the number of dimensions of TriPM, models with different factor structures were tested. First of all, a single-factor model was suggested for testing for unidimensionality. The model was assumed to be a 1PL IRT model, holding all loadings the same across all items (i.e., tau-equivalence). To

assess model fit, a cut-off ciriteria suggested by Hu and Bentler (1999) was used : CFI $\geq$ .95, TLI $\geq$ .95, and RMSEA $\geq$ .06 for excellent fit.

CFA results for the single-factor model with fixed loadings showed very poor model fit (CFI = .372, TLI = .371, RMSEA = .075), indicating that factor loadings should be allowed to vary. It is more reasonable to assume the 2PL model because given many items and a large sample size, it does not make sense for all items to have the same difficulty and discrimination for all of the participants.

Accordingly, the unidimensional model with varying factor loadings was estimated. The loadings varied substantially with mostly negative loadings, ranging from |-.004| to |-.903| and providing evidence that the tau-equivalence assumption does not hold for TriPM. The model fit indices were better than those of the 1PL single factor model (CFI = .587, TLI = .572, RMSEA = .062), but generally indicated poor fit. As expected from the fit of the two models, the results of `anova` and `lavTestLRT` presented that model with varying loadings fitted the data better at a significance level of 0.01 ($p < .001$). This was also supported by Δ CFI of .215.

Reliability was estimated despite the poor model fit. Because the single factor 1PL model is the unidimensional model with the same loadings, ordinal alpha that accounts for the categorical nature of the dataset was used. For the 2PL model, `omega 1` was used. `omega 2` can be also used as it gives the same value in a simple factor structure (Flora, 2020). Reliablity estimates were high for both two models (α = .94, ω = .87 for 1PL model, α = .94, ω = .82 for 2PL model). However, the reliability estimates were not trustworthy because of the poor fit. It is important to note that although alpha is most widely and commonly used as an internal consistency measure in many fields, it is problematic, especially in that (a) alpha increases with the number of items, (b) requires assumptions that are hard to be met such as essential tau-equivalence, and (c) alpha does not contain information on an internal consistency of a test (Sijtsma, 2009).

Based on the results above, the 2PL was assumed for the estimation of measurement models. With the strong evidence that unidimensionality does not hold for TriPM scale, a three-factor structure was tested, on which traits related to psychopathy were originally defined. The items were loaded on their respective factors mostly with loadings > .40, and all loadings were significant and positive. The model fit indices did not indicate acceptable fit (CFI = .693, TLI = .681, RMSEA = .053), however, the model fit improved compared to those in the single-factor 2PL model (Δ CFI = .106). This was also shown in a

model comparison of `anova` and `lavTestLRT`, both indicating the same result that the three-factor structure represented the data better ($p < .001$). Interestingly, disinhibition and meanness were strongly correlated to each other. This is due to the fact that the items comprising two traits are from the same scale, Externalizing Spectrum Inventory (ESI), which is developed to assess externalizing behaviors (Stanton et al., 2021). The reliability estimate was examined and indicated that the model measured what it was supposed to measure with all values above .70 ($\alpha = .85$, $\omega = .77$), however, it was not reliable due to the unstable model fit.

Next, bifactor model was estimated, where the general factor was added to the three-factor structure. Most of the items loaded negatively onto a general factor, varying substantially (|.006| - |-.912|), some of which were not significant. The model fit indices still indicated poor fit (CFI = .848 TLI = .836, RMSEA = .038), but yielded the the best model fit among the models discussed previously. There was a significant improvement in model fit in comparison with those of the three-factor model ($p < .001$, Δ CFI = .154), due to the addition of the general factor. Table 1 shows reliability estimates for the bifactor model. `omega 2` and `omega 3` indicated the total score variance is more explained by boldness, rather than by general factor. This provides evidence for the multidimensionality of the data, indicating unidimensionality should not be assumed for this scale, as shown in the model fit of single-factor models. This also implies that although the bifactor model indicated the best model fit, another factor structure should be considered that represents and fits the scale better.

Table 1

Reliablity estimates for bifactor model

```
reliability(tri.bif.2PL)
             GEN  Boldness Disinhibition Callousness
alpha     0.8395397 0.7590852    0.80264026   0.8288411
alpha.ord 0.9368629 0.8504525    0.92732057   0.9399853
omega     0.6948356 0.7270055    0.23533502   0.3641752
omega2    0.5815948 0.7737772    0.07697527   0.1761607
omega3    0.5955170 0.7962279    0.07591311   0.1726348
avevar           NA        NA            NA          NA
```

Testing alternative factor structures

Accordingly, two alternative factor structures were proposed. The first model was derived from the three-factor model where disinhibition and meanness had a strong correlation of .809. To account for the highly correlated factors, two factors and their items were combined with one single factor, and thus, the first alternative had a two-factor structure. The model fit indices (CFI = .671, TLI = .659, RMSEA = .055) indicated the factor structure was not suitable for the scale. The same results were shown in the comparison to the three-factor model and bifactor model, indicating that the two models were significantly better than the two-factor model ($p < .001$ for both models).

The second model was suggested based on the first alternative model by adding a general factor. The number of specific factors was the only difference with the bifactor model with three factors. The fit of the second model indicated that the model did not differ significantly from the bifactor model with three factors (Δ CFI = .004). The reliability estimates yielded similar values to those of the bifactor model (α = .94, ω = .58). Taken all the fit indices and model comparisons together, it can be concluded that the bifactor structure defined by three factors represented the data best among all models suggested and two alternatives. That is, unidimensionality does not hold and the bifactor structure is the most appropriate for the scale.

Calculation of reliability

With the bifactor model that showed the best model fit of all models, further investigation on reliability estimates was conducted by manually calculating them and comparing results from `reliability` function. ω_t for scale and subscales for bifactor model were estimated and produced relatively higher values than values from `reliability` function (0.96 for all general factor and three subscales, 0.86, 0.96, 0.95 for each subscale, respectively). The results were very similar to ordinal alpha (0.93, 0.85, 0.93, 0.94 for general and three factors respectively).

ω_h was then calculated by hand as well as using `compRelSEM` function. Omega estimates were 0.78 for all general factor and three subscales, 0.85, 0.05, 0.24 for each subscale, respectively. Table 2 gives the reliability estimates calculated from `compRelSEM`. The result of Manual calculation corresponded to the estimate from `compRelSEM` function with Green and Yang's correction removed. `compRelSEM` with correction yielded similar estimates to `omega 2` for bifactor model from `reliability` function (0.58, 0.78 0.08, 0.18 for general and the three factors respectively). Given that data is categorical, Green and

Yang (2009)'s approach should be applied, and therefore estimates from `compRelSEM` with `ord.scale=T` are preferable.

Table 2

Omega hierarchical for bifactor model using `compRelSEM`

```
compRelSEM(tri.bif.2PL, obs.var=T, ord.scale=F)
GEN      Boldness Disinhibition   Callousness
0.795         0.902         0.044         0.241
compRelSEM(tri.bif.2PL, obs.var=F, ord.scale=F)
GEN      Boldness Disinhibition   Callousness
0.775         0.870         0.045         0.241
compRelSEM(tri.bif.2PL, obs.var=T, ord.scale=T)
GEN      Boldness Disinhibition   Callousness
0.596         0.796         0.076         0.173
compRelSEM(tri.bif.2PL, obs.var=F, ord.scale=T)
GEN      Boldness Disinhibition   Callousness
0.582         0.774         0.077         0.176
```

The ECV of 0.62 also backed up the conclusion of reliability estimates, meaning not much of variance of total score are explained by the general factor, and indicating again that unidimensional does not hold.

Sum scores

Summing over all the sum scores of subscales corresponded to the sum scores of the main scale, which is defined as a total score. Table 3 shows the sum scores of the scale and three subscales.

According to Liu and Pek (2024), factor scores were found to perform best on a correctly specified model with a large sample size. Therefore, given a large number of participants (N =1,423), a factor score as a latent variable would be more appropriate than a sum score if the model represents the data well.

Table 3

Sum scores of the scale and subscales

```
# Sum scores of main scale
sum(tripm.dic) # 23090
# Sum scores of sub scales
sum(tripm.dic[items.b]) # 12717
sum(tripm.dic[items.d]) # 5957
sum(tripm.dic[items.m]) # 4416


sum(tripm.dic[items.b])+sum(tripm.dic[items.d])+sum(tripm.dic[items.m])
# 23090
```

Alternative factor structure using EFA

As prior studies on the factor structure of TriPM have not converged to one conclusion, an optimal factor structure for the scale was explored. Unlike the alternative factor structures discussed earlier, factor structures were proposed by conducting EFA and reporting omega estimates. Factor structures were tested starting from the three-factor structure, which is the default, and investigated until there was no improvement in reliability estimates.

As a result, five structures were estimated and the four-factor solution was found to be suitable to data the most, presenting higher estimates than others (ω_t = .90, ω_h = .65, ECV = .45). Items on TriPM were reassigned to each of four factors as suggested in the four-factor solution. Because many of the factor loadings were less than .2, items were retained (a) that had strong loadings $\geq |.40|$ on one factor and (b) that weakly loaded on other factors $\leq |.30|$ (Clark & Watson, 2019). There was one item cross-loading on two factors (item 36) with loadings of .48 and .51. Since the item cross-loaded highly on both factors, it was not eliminated but was assigned to the factor that the item loaded on higher. Thus, 23 items were excluded from the analysis, and the rest 35 items were assigned to each of four factor : F1, F2, F3, and F4. The resulting factor structure solution can be found in Appendix A.

F1 contains 16 items consisting of 9 items from disinhibition and 7 items from meanness, F2

consists of five items assessing the same facet in scale, empathy in meanness. For F3, four of which are from boldness and the other is from meanness. Finally, F4 consists of 9 items, where 7 items are from boldness and two are from disinhibition. The retained items and their respective factors are listed in Appendix B. F2 can be labeled 'Empathy' since it is defined only by the items belonging to the facet of empathy. Labels were not specified in the other three factors, as the factors did not contain similar items assessing one or its associated traits, but rather defined by diverse content of items.

To investigate how well the suggested four-factor structure represents the data, CFA was performed. The same result was provided by `omegaFromSem`. The four-factor solution indicated a better model fit (CFI = .865, TLI = .855, RMSEA = .048) than the bifactor model with three factors that showed the best fit (CFI = .848, TLI = .836, RMSEA = .038). The improvement in model fit is shown in reliability estimates, indicating increased omega estimates (.87, .76, .70, .49 for F1, F2, F3, and F4, respectively). As expected, all items were cleanly loaded on each factor that they were assigned to, with loadings $\geq |.50|$ and significant p-values ($p < .001$ for all loadings). Also, a significant covariance between two factors was found : F1 was strongly correlated to F2 ($r = .744$). This could be due to that (a) meanness and disinhibition are from the same scale (i.e., ESI) (Stanton et al., 2021) or (b) a finding by Patrick (2010), the author of the TriPM, is reflected : scores on meanness scale has a stronger correlation with scores on disinhibition scale ($r = .4$) than with scores on boldness scale ($r = .2$). Finally, (c) a quarter of the items (4 items out of 16) consisting of F1 are the items assessing facet 'Empathy', by which F2 was defined. In other words, 25% of F1 items are items from F2.

Discussion on alternative factor structures of TriPM

Given that conclusions on factor structures and assignment of items to factors on TriPM are divergent in prior studies, some suggestions can be made. As proposed factors in the four-factor solution consisted of items assessing diverse facets in different factors, it is hard to define and label the content of items as one, specifically F1. Therefore, as Stanton et al. (2021) and his colleagues did in their study, a suggestion could be that conducting EFA separately on each of four factor, extracting subfactors defining facets within each factor if there exists, and then selecting items on each subfactor representing facets. By doing so, items can be more clearly specified in both factors and subfactors in TriPM (Stanton et al., 2021). Also, item selection can lead to the development of a brief form of the test. However, further research on factor structure on TriPM is still needed, so that consistent results are drawn and generalized in different

settings and on different groups of people.

Parameter estimation in CTT and IRT

Item parameter estimates in CTT

Item difficulty in CTT, also referred to item easiness, is defined as a proportion of correct responses of an item, ranging from 0 to 1 where higher values indicate easier items in contrast to difficulty in IRT. Since items were reverse-coded and dichotomized into 0 and 1, the number of correct responses was calculated by summing all responses of an item across all participants. That is, the number of responses '1' was summed and then divided by the number of participants to be in a range of 0 to 1.

Given that reverse-coding was conducted in a way that choosing a higher response category (i.e., '1') indicates how likely one has the traits related to psychopathy, item easiness close to 1 indicated that many of the participants were highly likely to have a trait assessed by an item. Six of the items were considered as easy items ($\geq$ .6), and two of them were greater than .7 : item 50 (.72) and item 22 (.74). This means many of the people see themselves who function very well in a new situation, even when unprepared (item 22), and stack up well against most others (item 50). More than 30 items were found to be difficult (< .4), meaning that only a few people answered 1 to the item.

To estimate item discrimination in CTT, a biserial correlation between individuals' total scores and responses on each item was used instead of a point-biserial correlation, because the dichotomous item response variables in this data were created from polytomous variables, which were not dichotomous originally. Most of the items were found to well discriminate among individuals, with all values positive except item 16 (-.03). This means that individuals with high scores are more likely to get an item correctly, and less likely for individuals with low scores. However, Six of the items showed weak biserial correlation (< .25), which are items 1, 10, 16, 21, 22, and 50. Of these, items with higher item easiness(> .7) such as items 22 and 50 indicated a low discrimination of less than 0.25. This can be interpreted as the number of people who answered '1' to this item is not much related to the possibility of one possessing three traits associated with psychopathy (i.e., individuals' total scores). In other words, answering '1' to item 50 or item 22, which indicates having self-confidence, is less relevant to one's total score.

Also, it is noteworthy that low easiness (< .25) items with very high discrimination (> .90) such as items 29, 34, 8, 55, and 58 can be strongly indicative of one having the traits associated with psychopathy. Because item discrimination of .90 indicates strong correlations between item responses of '1' and

individuals' high total scores, those items were most likely answered '1's by participants with high total scores. This can also interpreted that the items differentiate people with a normal range of traits from people with psychopathic traits very well. Interestingly, this was consistent with the four-factor solution discussed above in that all five items except item 8 were retained due to their strong relationships with respective factors and were assigned to the same factor, F1.

Item parameter estimates in IRT

Next, item parameters were estimated under IRT framework, starting with 1PL models using five different packages: `ltm`, `eRm`, `mirt`, `TAM`, and `lavaan`. The difficulty parameters estimated from the five packages were transformed to IRT parameters and then ordered from easiest to hardest items by rescaling the estmates, so that they can be compared on the same scale. Table 4 shows the three easiest items. The five packages estimated item difficulties very similar to one another, yielding slightly lower values in `lavaan`. However, as shown in the fit of the unidimensional model, when estimating the parameter in `lavaan` holding factor loadings the same across all items, CFA result indicated very poor fit (CFI = .372, TLI = .371). This finding implies a 1PL model does not fit the data well, providing strong evidence that assuming items with varying discriminations (i.e., 2PL model) is appropriate for the data. Additionally, the order of items sorted by item difficulties was mostly the same as in CTT, which indicates that CTT and IRT analyses provide similar results in terms of estimation of item difficulty.

Table 4

Item difficulties estimated from IRT models

```
head(diff1PL)
```

	CML-eRm	MML-ltm	EM-mirt	lavaan	MML-TAM
Item.22	-2.565408	-2.619275	-2.578502	-2.716291	-2.578005
Item.50	-2.492318	-2.544249	-2.504502	-2.638872	-2.504012
Item.1	-2.298156	-2.344585	-2.307899	-2.431828	-2.307431

Based on the 1PL models, person parameters θ were estimated. Figure 1 shows a density of estimates calculated from 5 different packages, where the x-axis is a latent dimension for a person parameter θ. This refers to how strongly a person is associated with the three traits, with high values being

high association with the traits and low values being low association. As shown, all parameter estimates except one estimated from `eRm` were normally distributed with most of the values around zero. Estimates from `eRm` were right-skewed, resulting in a lower mean of around -2. Figure 1 also shows estimates 'ksi' in `lavaan` slightly deviated from the normal distributions and were left-skewed.

Figure 1

Density of Person Parameter Estimates for 1PL

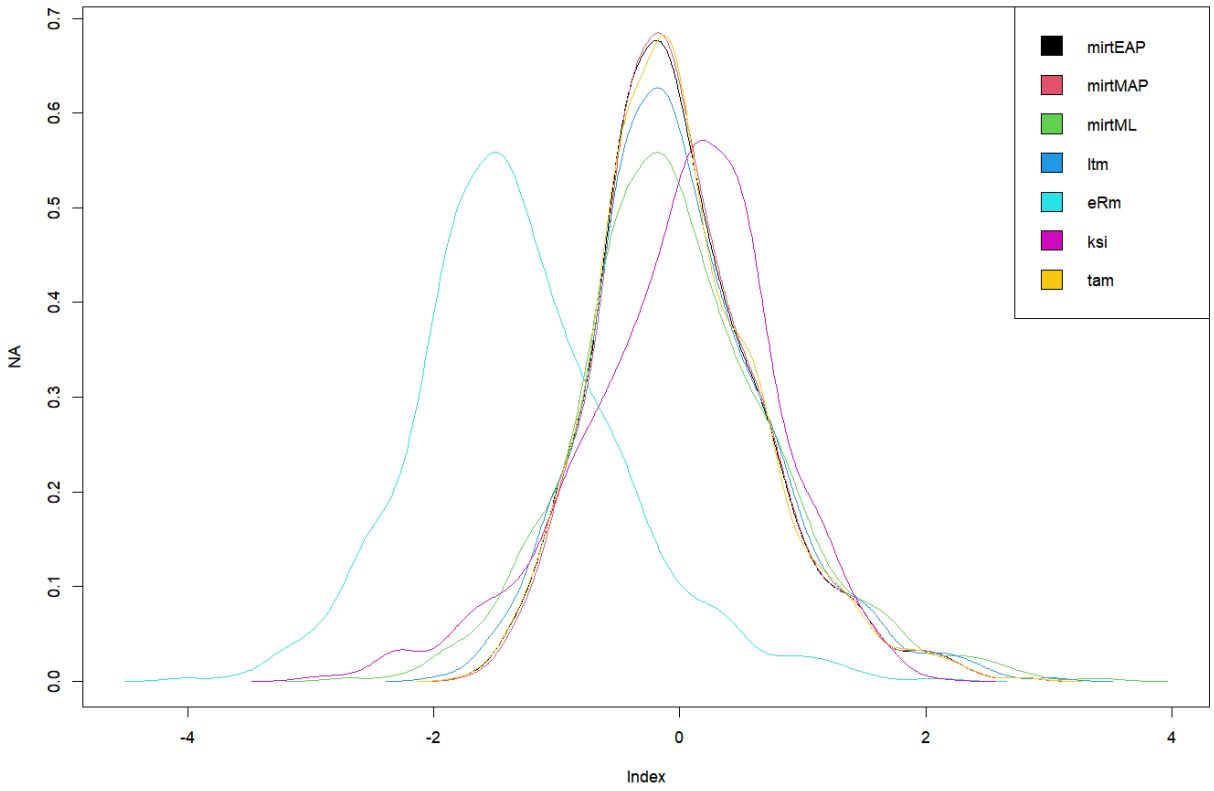
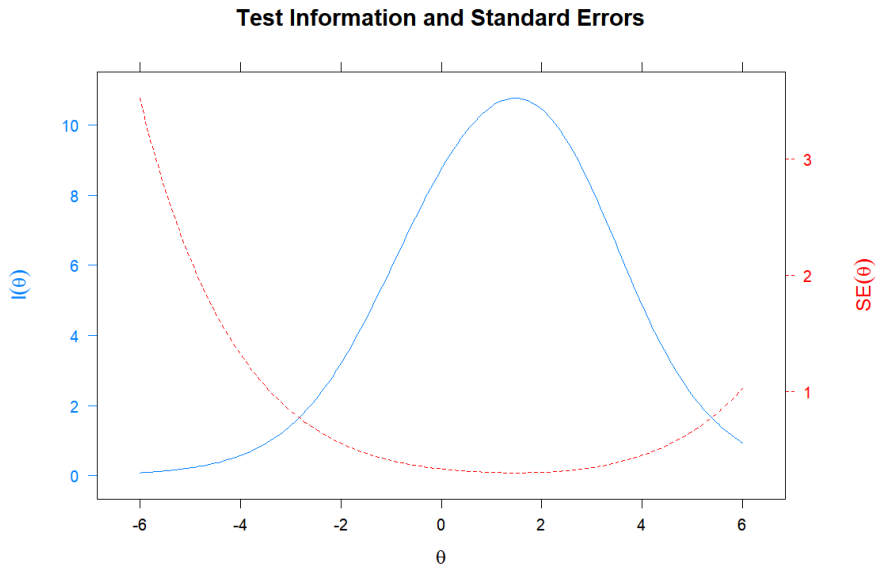


Figure 2 displays how much information the test provides on each ability estimate. Unlike the person parameters where most estimates were around 0, TriPM measured most of the information in a range of -2 to 4, and the most precise information was provided on ability around 2 with the smallest standard error. This means TriPM can be the most useful tool, especially for people with higher total scores on the scale.

One big limitation of 1PL models is that discrimination differences for items of the same difficulty

Figure 2

Test Information Function for 1PL model



are not represented in ICCs. Item 22 and 50, for example, had almost the same difficulties and were the two easiest items in both CTT and IRT (0.73 and 0.72 respectively in CTT, -2.61 and -2.54 in `ltm` in IRT) but with different biserial correlation (0.22 and 0.10, respectively). As Figure 3 presents, as long as they are in the same location (i.e., same difficulty), the 1PL model will consider them as the same item with the same item characteristics. Moreover, ICCs expressed that those items had steeper slopes than they actually did, which had to be flatter. Assuming the same discrimination across all the items can lead to biased parameter estimates and give inaccurate information on item parameters. Therefore, those findings support that a 2PL model is required for the scale.

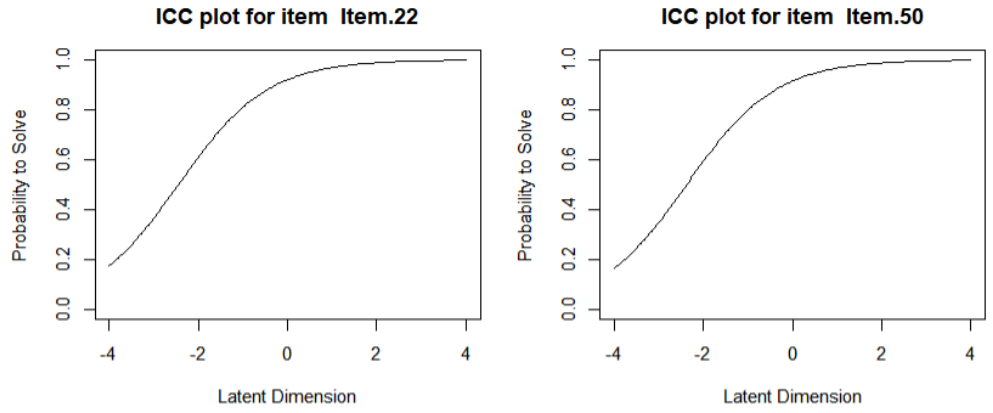
Item fit analysis

Item fit was estimated using packages `TAM`, `eRm`, and `mirt` and gave similar values to one another. Outfit values widely ranged from 0.49 (item 58) to 1.3 (item 25). An item was considered to have a good item fit when its value was around 1. Items were regarded as underfitting items when the fit was much larger than 1 and as overfitting items when the fit was much smaller than 1. In this data, many items were classified into misfit items with t statistics outside the range of (-2, 2) and significant p values, showing a similar number of underfitting and overfitting items.

Unlike outfit, infit showed less extreme fit and infit t statistics, leading to a smaller number of

Figure 3

Item Characteristic Curve for 1PL model

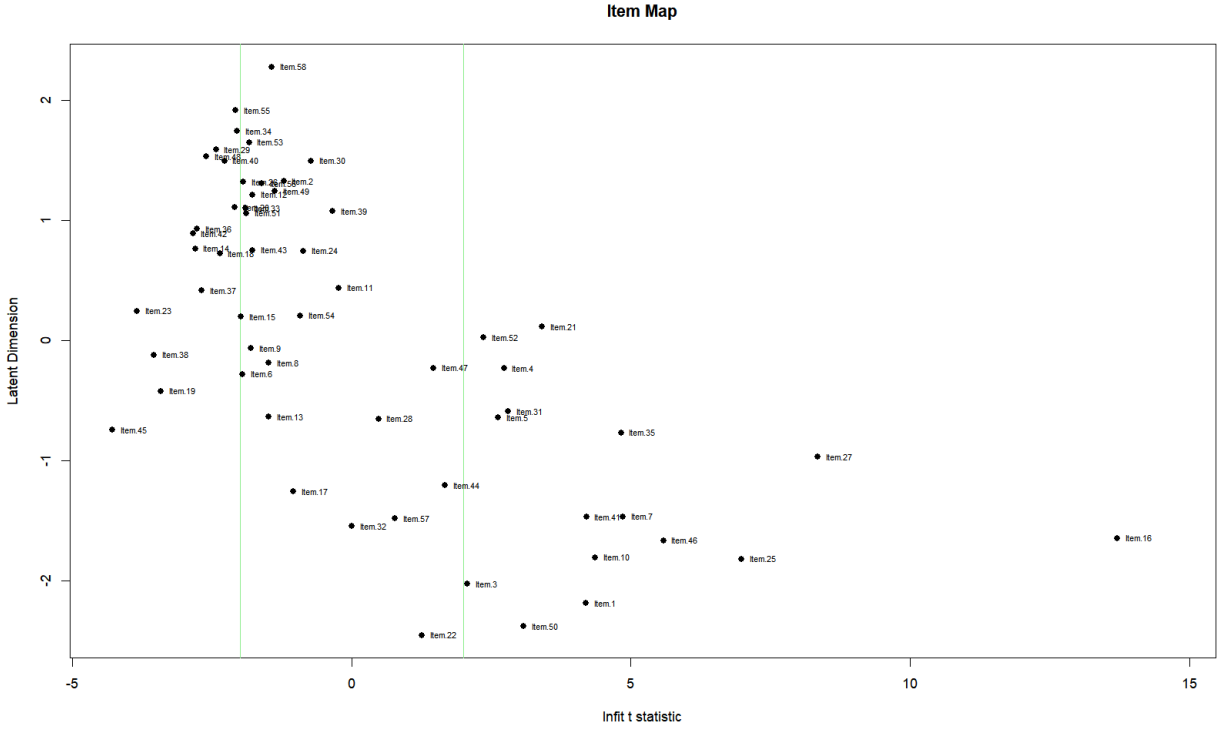


misfitting items, which is less than half. As shown in Figure 4, many of the items are outside the *t* value of (-2, 2). The left side of the green line indicates overfitting items and the right side indicates underfitting items. The key finding was that overfitting items and items on the left upper side of Figure 4 mostly consisted of high difficulty items, and in the middle and bottom right, most of them were underfitting items and items with moderate and low difficulties. This means the items with higher difficulties predicted better than easy difficulty items, although they were overfitting items. This also matches the result of item estimation in CTT, where easier items were associated with low discrimination, and difficult items mostly had higher discrimination.

Figure 5 demonstrates how item fits are related to item discriminations. The top two graphs are the ICCs for item 16 and item 50, which had relatively easier difficulties with low discrimination (< .25) in the CTT analyses. As presented in figure 4, item 16 (infit = 1.232, infit t = 12.443, $p < .001$ in `TAM`) and item 50 (infit = 1.098, infit t = 3.412, $p = .001$ in `TAM`) also indicated underfit deviating from t value range (-2, 2) with infit greater than 1. Both ICCs showed that observed ICCs of the items displayed flatter slopes than theoretical ICCs, and did not follow the theoretical ICCs, meaning that underfitting items (i.e., infit > 1) did not disrminate well (Wu, 2022). The bottom two ICCs are for item 29 (infit = 0.862, infit t = -1.618, $p = .106$ in `TAM`) and item 48 (infit = 0.856, infit t = -1.747, $p = .081$ in `TAM`), which were considered to be difficult (item easiness < .10) and high-discriminating (item discrimination > .90) items in CTT analyses. It should be noted items 29 and 48 were classified into overfitting items in Figure 4, while their infit t-statistics were within the range of (-2, 2). However, regardless of the misfit of the items, the ICCs of items

Figure 4

Item Infit t-Statistics



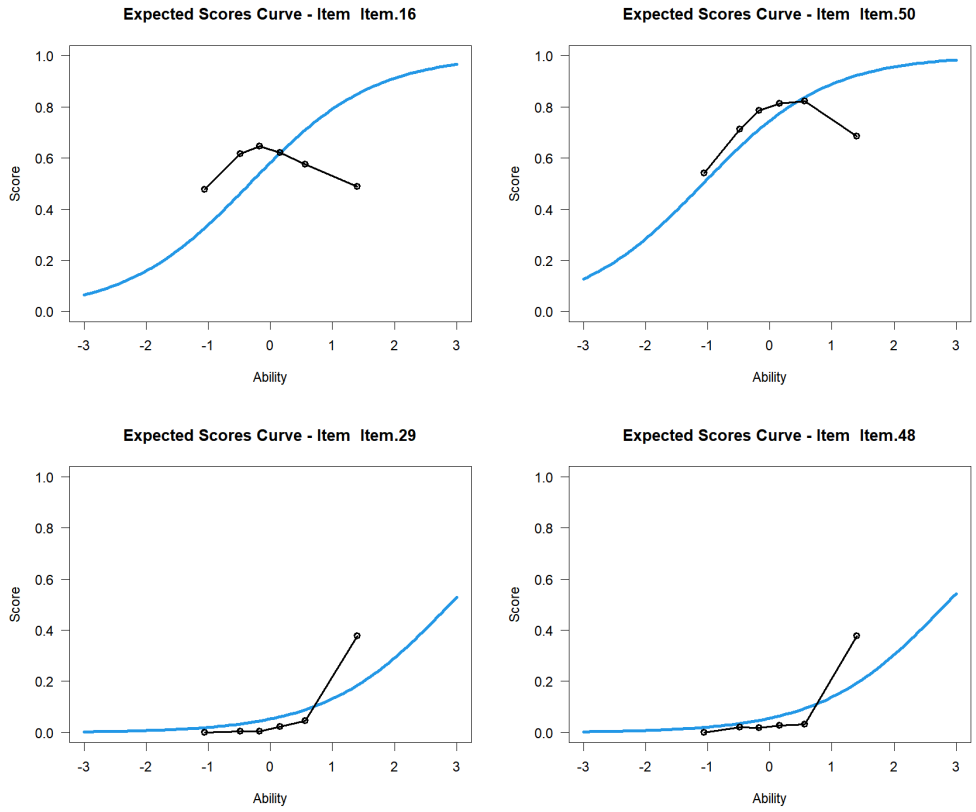
29 and 48 showed that items with smaller infit than 1 had steeper slopes than theoretical ICCs, which means those items discriminated better than the average items did (Wu, 2022). All these ICCs and misfits of the 1PL model considered, those findings again imply that the 1PL model should be extended to 2PL to account for item discriminations and to improve the item fits.

Comparison of 1PL and 2PL model

2PL models were estimated using the packages `TAM` and `mirt`. Then, the models were compared with the 1PL models in terms of item and person parameters, and item fit indices. Item difficulties estimated from the two packages yielded very similar results to each other as in 1PL results. However, item difficulty estimates for 2PL models were completely different from estimates for 1PL models. The order of items by difficulty also showed different orders from one another, unlike it was the same in CTT and 1PL results. It was because item difficulties were estimated with item discrimination simultaneously. Although difficulty estimates and the order of items sorted by their difficulties were different from the results of CTT

Figure 5

Theoretical and Observed ICCs for 1PL



and 1PL, there still was a tendency that easier items in CTT and 1PL analyses showed relatively low item difficulties, and the same was applied to difficult items.

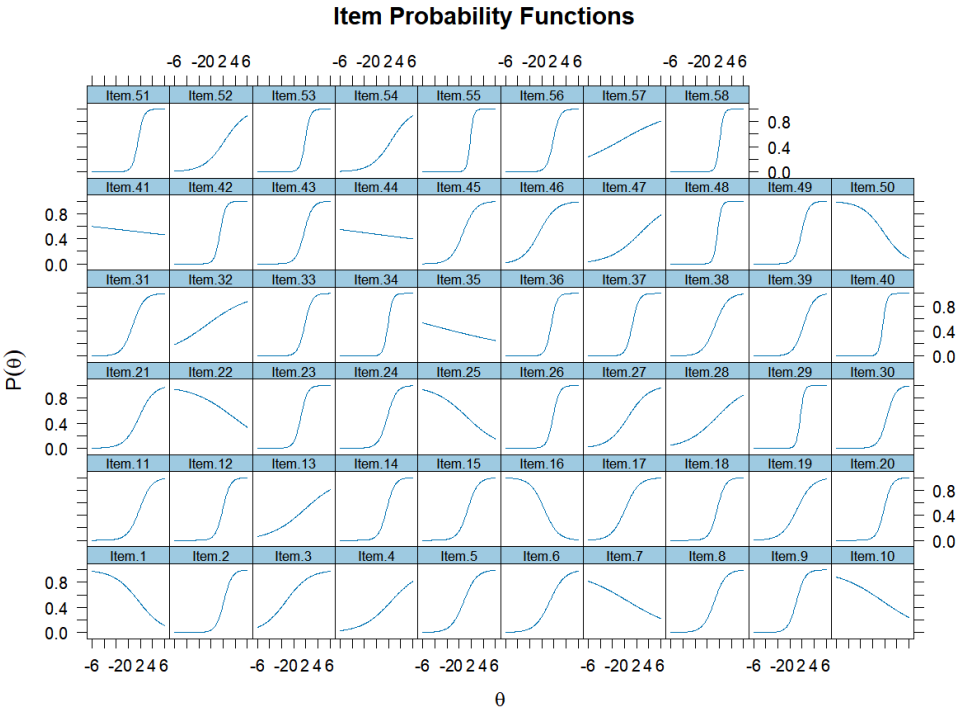
Item discriminations were estimated along with item difficulties for 2PL. Parallel to the estimation of item difficulties, item discriminations exhibited a nearly identical pattern where the estimates from `TAM` and `mirt` were almost the same each other with only small differences in values. The order of items sorted by discrimination was largely preserved, compared to the biserial correlation in CTT analysis.

Figure 6 presents the ICCs of all items. The varying slopes of the ICCs again indicate the necessity of discrimination for this dataset for better interpretation. The key finding was that item 22, the easiest item in the 1PL models, became the most difficult item in the 2PL models. This big difference is due to the introduction of discrimination, which is expressed by the negative slope of item 22 in Figure 6. despite the high probability of answering the item correctly (-2.58 in `mirt` in 1PL, 0.74 of easiness in CTT), The negative discrimination lead the item to the most difficult one, resulting in unreliable and

misleading interpretation of item parameters. This also implies that the inclusion of items with negative or low discrimination should be considered with caution.

Figure 6

Item Characteristic Curves for 2PL model

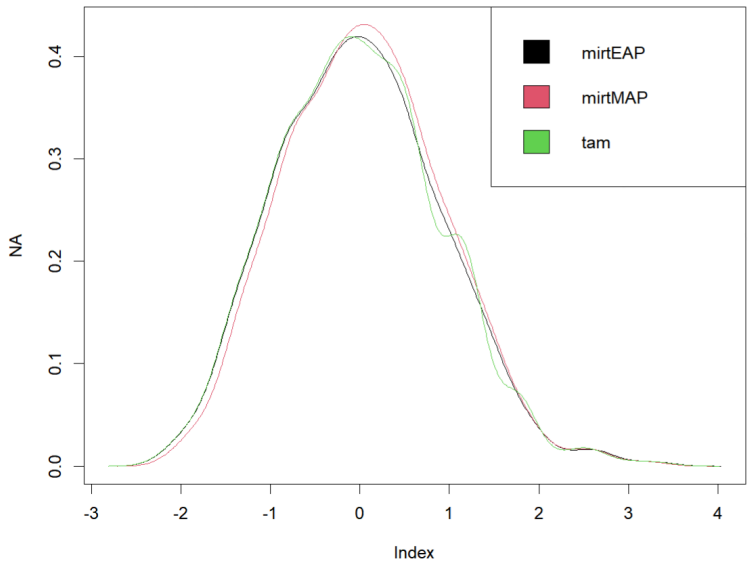


Person parameters were then estimated. Figure 7 presents the density of person parameter estimates, where the x-axis is a latent dimension for a person parameter θ. Consistent with the density for 1PL models (see Figure 1), both `mirt` and `tam` produced almost the same densities, and the estimates were normally distributed around zero, which means most of the participants had normal (i.e., not too high) trait scores associated with three psychopathy traits. Also, person parameters mostly ranged from -2 to 2 as in 1PL, but with lower probabilities around .4 and a larger width, meaning the estimates were more spread over the latent dimension with larger variance than estimates of 1PL. This may be due to the introduction of discrimination, and as a consequence, it might have produced the various person parameter estimates.

Figure 8 shows the TIF of TriPM and its standard error. Most of the information was provided on the range of ability estimates from -2 to 3. TriPM measured the most precise information around 2. Compared to TIF for 1PL (see Figure 2), the maximum standard error for 2PL decreased from 3 to 2.5 with its minimum around 2. Also, the information provided by the test increased three times (10 to 30) as

Figure 7

Density of Person Parameter Estimates for 2PL



expressed in sharper and narrower distribution. This indicates the test measures and distinguishes individuals with higher trait scores most accurately (i.e., $\theta = 2$).

Finally, 1PL and 2PL models were compared in terms of item fit indices. In contrast to the result for item fit for 1PL models where infit and outfit substantially varied with extreme t statistics and significant p values, the result of 2PL models showed that t values of most items were within a range of (-2, 2) with insignificant p values. Especially, p values for infit for 2PL model estimated from `TAM` were all significant.

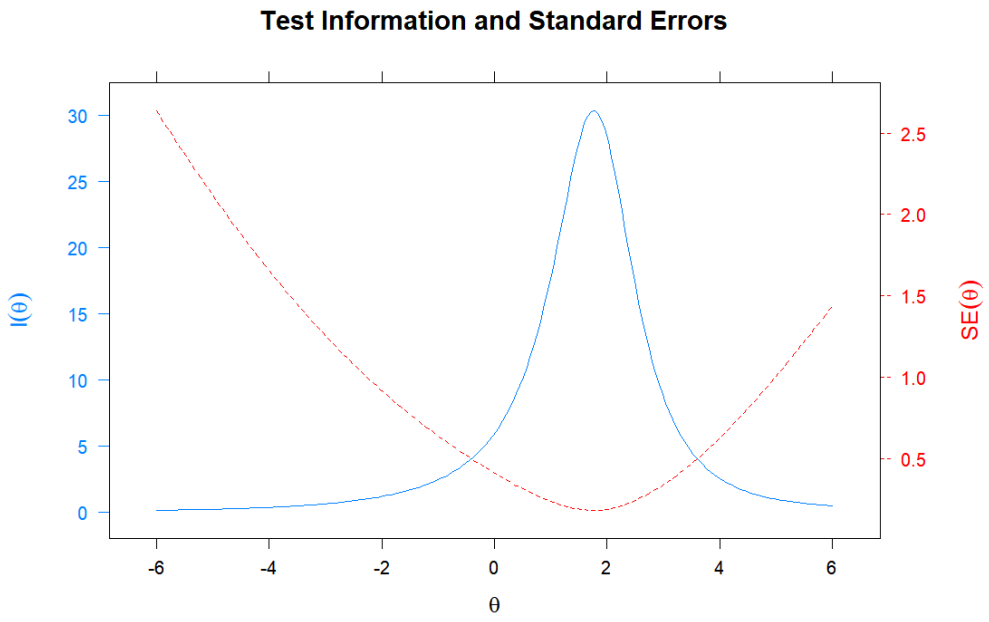
Altogether, all the results of the analyses discussed above consistently indicated that item discrimination should be included, and a 2PL model fitted the data better than 1PL. This was supported by the ANOVA results between 1PL and 2PL models from each package, providing much smaller AIC and BIC values in all 2PL models. The result is shown in Table 5.

Differential item functioning of the dichotomized data

On the basis of the 1PL and 2PL models estimated, DIF and measurement invariance were investigated to check if the traits were measured equally across groups. Uniform and non-uniform DIF items with respect to `SEX` of the participants were first examined, followed by detection of DIF with respect to `PSYCH PROB`, and four groups based on the combination of each condition of the two variables.

Figure 8

Test Information Function for 2PL model



DIF with respect to `SEX`

The participants were grouped by sex with males being a focal group. Mantel-Haenszel method reported that nearly half of the items (26 items) were found to perform differently in each group, 10 of which showed a large effect size labeled as C, indicating many of uniform DIF items have large differences in males and females. Most of the 10 items exhibited positive effect size under the column `DeltaMH`, which indicates items were easier for males and answered correctly more by males than by females. Since the data was reverse-coded in a way that higher response categories such as 3 or 4 are more likely to have traits associated with psychopathy, and then dichotomized into 0 and 1 in the same way. Thus, getting items correctly (i.e., answering 1 in dichotomized data, and 3 or 4 in original data) means that males are more likely to have traits associated with psychopathy than females.

Breslow-Day method was used to detect non-uniform DIF. Six items were found to be non-uniform items, meaning the discriminations of these items differ across male and female groups. Considering the result of Mantel-Haenszel method together, items 10, 11, and 44 were flagged as both uniform and non-uniform DIF items.

DIF results were compared using `dichoDif` with five methods: Mantel-Haenszel, Standardization,

Table 5

Comparison of 1PL and 2PL models

```
anova(tripm.tam, tripm.tam.2PL)
          Model   loglike Deviance Npars      AIC      BIC    Chisq df  p
1     tripm.tam -38235.91 76471.82    59 76589.82 76900.19 3188.328 57  0
2 tripm.tam.2PL -36641.74 73283.49   116 73515.49 74125.71       NA NA NA


anova(tripm.mirt, tripm.mirt.2PL)
                    AIC    SABIC       HQ      BIC    logLik       X2 df p
tripm.mirt     76589.39 76712.34 76705.32 76899.76 -38235.69
tripm.mirt.2PL 73509.14 73750.87 73737.07 74119.36 -36638.57 3194.249 57 0
```

Logistic regression, Raju, and Lord. For 1PL model, six items were flagged as DIF items by all five methods, nine items were flagged by four methods, and 17 items were never detected as DIF items. To account for non-uniform DIF, Breslow-Day method was added to the five methods for the comparison of DIF results for the 2PL model. Similar to the result of 1PL, six methods did not give matching results, where Raju and Lord methods detected most of the items as DIF items, whereas only a few items were identified as DIF by Standardization and Breslow-Day methods. It is interesting to note that all of the items were detected as DIF items by at least one method, seven items by five methods, and only item 7 was found to be DIF items by all six methods. Additionally, the results showed Raju and Lord methods, which are both methods of IRT framework, detected most items as DIF, whereas Breslow-Day and Standardization methods detected only a few as DIF. Given that many items were found to have a uniform and non-uniform DIF across male and female groups, `SEX` can be regarded as the meaningful and relevant variable for measuring the traits related to psychopathy. This implies that for interpretation of items measuring psychopathy or similar traits, sex should be taken into account and be examined if there are any differences depending on group membership.

DIF with respect to PSYCH PROB

The participants were then grouped by PSYCH PROB variable, a question asking experience requesting assistance for psychological issues. People who answered that they had the experience were chosen as a focal group. One-third of all items were detected as uniform DIF, four of which exhibited a large DIF effect. Both two large effect size items (items 9 and 51) assessed impulsivity with positive DIF effects. This indicates participants who had asked for help with psychological distress were aware that people they know were concerned about their impulsivities (e.g., "My impulsive decisions have caused problems with loved ones", "Others have told me they are concerned about my lack of self-control"), which might have led them to request assistance for their psychological issues. Items 1 and 50 assessing boldness had a large negative DIF effect, meaning participants with experience in psychological help-seeking think of themselves as less optimistic than the people without the experience and rate themselves they are not as good as others.

Non-uniform DIF items were detected using Breslow-Day method. Five items functioned differently across two groups, including items 10 and 44, which were also found to have both uniform and non-uniform DIF effects across male and female groups. Therefore, by creating four groups based on the conditions of SEX and PSYCH PROB, it can be investigated how items 10 and 44 function differently depending on the four conditions.

Unlike the comparison of DIF methods for 1PL model for SEX, the result for 1PL PSYCH PROB showed the five methods identified similar items as DIF items, yielding relatively consistent results across all methods used. For 2PL, most of the items were flagged as DIF by at least one method. Same as above, Lord identified almost all items as DIF items, whereas Breslow-Day and Standardization methods detected only a few items as DIF. Two items (items 35 and 41) were never detected as DIF items.

DIF with respect to SEX and PSYCH PROB

As some of the items (e.g., items 10 and 44) were found to be both uniform and non-uniform DIF items, a multi-group DIF analysis in terms of both SEX and PSYCH PROB was performed. The participants were divided into groups according to their responses on two variables, resulting in four groups : females with experience in requesting assistance for psychological distress as a reference group, labeled as Female.ReqX, female without the experience (Female.Req), male without the experience (Male.ReqX), and male with experience (Male.Req).

Similar to single-group DIF analyses conducted, generalized Mantel-Haenszel, generalized Lord, and generalized logistic regression were used as methods for the detection of DIF items. For the 1PL model, almost half of the items were found to be uniform DIF items, most of which were agreed by all three methods. This indicates the difficulties of half of the items varied depending on sex and experience requesting assistance for psychological issues. In DIF analysis for the 2PL model, about half of the items were detected as non-uniform DIF items by all three methods as well as all items were detected by at least one method, except item 55 (i.e., "It does not bother me when people around me are hurting") assessing empathy in meanness scale. It implies item difficulties and discriminations of item 55 does not depend on sex and experience requesting assistance for psychological issues. More items were flagged as DIF items for the 2PL model, which means many of the items exhibited uniform DIF, non-uniform DIF, or both of the effects. Consistent with the result of multi-group analysis for 1PL, this indicates there were many items whose difficulties and discriminations were measured differently depending on sex and experience requesting assistance for psychological issues.

Given that many of the items were flagged as DIF items, one can conclude that measuring boldness, meanness, disinhibition, or similar traits associated with psychopathy are highly affected by sex, experience in psychological help-seeking, and interactions of the two variables. This suggests future item development assessing traits related to psychopathy should thoroughly examine DIF related to the two variables for a more accurate measurement of traits.

Measurement invariance of original categorical data

Measurement invariance of `SEX`

As the equivalent of DIF, measurement invariance was tested using original 4-point Likert 58 items with respect to `SEX`, `PSYCH PROB`, and `AGE`. Before investigation on measurement invariance, a baseline model (i.e., configural invariance) was assessed. In order to proceed to testing of next level of invariance, a fit of a baseline model should be sufficient. Thus, a bifactor structure was chosen, as it showed the best fit of all models in the dimensionality analysis discussed previously. For baseline model fit, Van De Schoot et al. (2012)'s criteria for adequate fit were considered : CFI > .90, TLI > .90, and RMSEA < .08.

As shown in Table 6, a baseline model for `SEX` yielded poor model fit (CFI = .853, TLI = .841, RMSEA = .051), not reaching the cut-off for acceptable fit suggested by Van De Schoot et al. (2012). Nevertheless, for analytical purposes, the weak fit of configural invariance was accepted. Adding constraint

of equal loadings indicated a non-significant difference (p = .981) in chi-square values between configural and metric invariance models. However, as chi-square tests are sensitive to sample size, alternative fit indices such as changes in CFI were considered, and Chen (2007)'s suggestion was chosen to assess changes in model fit : Δ CFI $\geq -.01$ and Δ RMSEA $\leq .015$ for equal sample sizes, and Δ CFI $\geq -.005$ and Δ RMSEA $\leq .010$ for unequal sample sizes.

Given unequal sample sizes of males and females (N = 810 for males, N = 613 for females), fit criteria for unequal sample sizes were used. With the insignificant difference between configural and metric invariance models (p = .98), changes in CFI (Δ CFI = .032) were considered together, indicating metric invariance across males and females. Testing for scalar invariance of `SEX` indicated .02 of decrease in CFI. This indicates a violation of scalar invariance, which was expected since DIF was already found with respect to `SEX`. The results suggested that items load on the traits with similar magnitude across males and females, and therefore covariances of factors and factor loadings can be meaningfully compared between groups (Hirschfeld & Brachel, 2014). For the establishment of a higher level of invariance, the investigation on partial scalar invariance can be suggested as one of the solutions. It can be done using a modification index by releasing an item parameter constraint that has the largest impact on a chi-square value one at a time, and repeating it until differences in chi-square values between partial scalar and full scalar invariance models are non-significant.

Table 6

Model fit for measurement invariance of SEX

	χ^2	df	p	RMSEA	CFI	TLI
Configural	8727.724	3074.000	0.000	0.051	0.853	0.841
Metric	7616.247	3186.000	0.000	0.044	0.884	0.880
Scalar	8491.495	3298.000	0.000	0.047	0.865	0.864

Measurement invariance of `PSYCH PROB`

Table 7 gives model fits at each level of invariance. As sample sizes were unequal across groups (N = 400 for request experience, N = 1,023 for no request experience), the cut-off for alternative fit indices for unequal sample sizes was considered. With the insufficient fit of the baseline model for `PSYCH PROB` (CFI = .867, TLI = .857, RMSEA = .052), equal loadings were added to all groups to test for metric

invariance. Changes in CFI from configural to metric was .036, which is larger than -.005, while the result of the chi-square test indicated the insignificant difference (p = 1.0). Given that chi-square tests are sensitive to sample size, Δ CFI was also used as an alternative fit index. This resulted in a lack of metric invariance, and thus threshold invariance was considered instead, by constraining thresholds to be equal across the groups. As Table 7 shows, changes in CFI at each invariance level were larger than -.005 (-.002 and .025 respectively). Although the chi-square test indicated configural and threshold models differed significantly at a significance level of 0.05 (p = .02), threshold invariance was established given sensitivity to a sample size of a chi-square test and small difference in fit indices (Δ CFI = .002, Δ RMSEA = .001). The subsequent testing for scalar invariance yielded Δ CFI of .025 with the insignificant difference (p = 1.0). Scalar invariance, therefore, can be established, meaning that scores in each trait can be meaningfully compared across all two `PSYCH PROB` groups, in addition to factor variance and covariance. Thus, a higher level of invariance was established for `PSYCH PROB` groups than the level of invariance of `SEX`, which might be attributed to the fact that fewer items were found to have DIF effects with respect to `PSYCH PROB`. However, as discussed in the measurement invariance of `SEX`, since some items were found to have DIF related to `PSYCH PROB`, interpretation of scalar invariance of the variable should be done with caution.

Table 7

Model fit for measurement invariance of PSYCH PROB

	χ^2	df	p	RMSEA	CFI	TLI
Configural	8884.010	3074.000	0.000	0.052	0.867	0.857
Thresholds	9084.461	3186.000	0.000	0.051	0.865	0.860
Scalar	8104.270	3298.000	0.000	0.045	0.890	0.890

Measurement invariance of `AGE`

To test for measurement invariance of `AGE`, age was divided into four groups. Participants ranged from 16 to 89, with most of them in their early 20s. Using the median age of 24, the age group was first divided into two groups : 'Under 25' (16 - 25, $N = 834$) and 'Over 25' (26 - 89, $N = 589$). Then, to deal with unbalanced the sample size and the wide range of age, the two groups were divided again into two groups, resulting in four groups : 16 - 20 ($N = 226$), 21 - 25 ($N = 608$), 26 - 50 ($N = 306$), 50 + ($N = 283$). Due to `lavaan` warnings, a large number of groups (e.g., 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65,

65 +) were not used. Considering the unequal sample sizes of four groups, the cut-off for alternative fit indices for unequal sample sizes was used. Unlike the invariance testing of variables with two groups done previously, running a CFA with different age groups sometimes gave the lavaan warning, which seems to be caused by dividing continuous variables manually into groups. Nevertheless, again, for the analytical purposes, a configural model with an unstable fit was used for invariance testings.

With a configural invariance model (CFI = .865, TLI = .847, RMSEA = .049), metric invariance was tested. CFI was increased (Δ CFI = .037) with the non-significant difference between configural and metric invariance models (p = 1.0), indicating the establishment of metric invariance. The result of scalar invariance indicated that Δ CFI smaller than -.005 (Δ CFI = -.016), and the chi-square test result with a warning message. Thus, instead of metric invariance, threshold invariance was considered. Table 8 gives model fit at each level of invariance. As shown, threshold invariance was established (Δ CFI = -.005, p = .98). The comparison between threshold and scalar invariance models suggested that Δ CFI larger than -.005 with insignificant chi-square test result (Δ CFI = .026, p = 1.0). This indicated the establishment of scalar invariance, and therefore comparisons of scores on traits across the age groups can be done.

Table 8

Model fit for measurement invariance of AGE

	χ^2	df	p	RMSEA	CFI	TLI
Configural	11450.868	6148.000	0.000	0.049	0.865	0.854
Thresholds	11965.352	6484.000	0.000	0.049	0.860	0.857
Scalar	11274.965	6820.000	0.000	0.043	0.886	0.890

References

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142–151. https://doi.org/10.11591/ijere.v7i2.12900

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2000-03918-000

Flora, D. B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Green, S. B., & Yang, Y. (2009). Reliability of Summed Item Scores Using Structural Equation Modeling: An Alternative to Coefficient Alpha. *Psychometrika*, *74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hirschfeld, G., & Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment Resaerch and Evaluation*, *19*, 1–12. https://doi.org/10.7275/qazy-2946

Hoffman, L. (2014). *Classical Test Theory for Assessing Scale Reliability and Validity*. [PowerPoint slides]. lesahoffman.com. https://www.lesahoffman.com/PSYC948/948_Lecture3_CTT.pdf

Hoffman, L. (2018). *Latent Trait Measurement Models for Binary Responses: IRT and IFA*. [PowerPoint slides]. lesahoffman.com. https://www.lesahoffman.com/CLDP948/CLDP948_Lecture5_Binary_Responses.pdf

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Idaka, I., & Idaka, E. (2014). From Classical Test Theory (CTT) to Item Response Theory (IRT) in Research Instrument. *Lwati: A Journal of Contemporary Research*, *11*(3), 36–44. https://www.ajol.info/index.php/lwati/article/view/119740

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Krishnan, V. (2013). *The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta's data*. Early Child Development Mapping (ECMap) Project. https://www.ualberta.ca/community-university-partnership/media-library/community-university-partnership/research/ecmap-reports/ediitemanalysisctt.pdf

Liu, Y., & Pek, J. (2024). Summed versus estimated factor scores: Considering uncertainties when using observed scores. *Psychological Methods*. https://doi.org/10.1037/met0000644

Patrick, C. J. (2009). *TriPm scoring key*. [Microsoft Excel spreadsheet]. Patrick CNS Lab. https://patrickcnslab.psy.fsu.edu/wiki/images/5/5a/TriarchicPsychopathyMeasure_key.xls

Patrick, C. J. (2010). *Tripm manual: Operationalizing the Triarchic Conceptualization of Psychopathy: Preliminary Description of Brief Scales for Assessment of Boldness, Meanness, and Disinhibition*. Patrick CNS Lab. https://patrickcnslab.psy.fsu.edu/wiki/images/b/b2/TPMmanual.pdf

Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology*, *21*(3), 913–938. https://doi.org/10.1017/S0954579409000492

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Stanton, K., Brown, M. F. D., & Watson, D. (2021). Examining the Item-Level Structure of the Triarchic Psychopathy Measure: Sharpening Assessment of Psychopathy Constructs. *Assessment*, *28*(2), 429–445. https://doi.org/10.1177/1073191120927786

Udoudoh, J. F., & Umoobong, M. (2016). Item Response Theory: A Tool for Education Measurement and Evaluation. *African Journal Of Theory And Practice Of Educational Assessment (AJTPEA)*, *3*(1). https://earnia.org/?p=journal-article&id=61

Van De Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Varma, S. (2006). *Preliminary Item Statistics Using Point-Biserial Correlation and P-Values*. Educational Data Systems Inc.: Morgan Hill CA. Retrieved. https://eddata.com/wp-content/uploads/2015/11/EDS_Point_Biserial.pdf

Wu, M. (2022). *Residual-based item fit statistics*. [Online tutorials]. A Course on Test and Item Analyses. https://www.edmeasurementsurveys.com/residual-based-item-fit-statistics.html

Appendix A

Alternative factor structure suggested by exploratory omega

```
omegaFromSem(tri.omg.4fac)
 Omega Hierarchical from a confirmatory model using sem =  0.11
 Omega Total  from a confirmatory model using sem =  0.94
With loadings of
            g  F1*  F2*  F3*   h2   u2   p2
Item.1   0.65                0.42 0.58 1.01
Item.7   0.56                0.32 0.68 0.98
Item.10  0.55                0.31 0.69 0.98
Item.16  0.74                0.54 0.46 1.01
Item.21- 0.65                0.42 0.58 1.01
Item.22  0.60                0.36 0.64 1.00
Item.31- 0.59                0.35 0.65 0.99
Item.44  0.47                0.22 0.78 1.00
Item.50  0.74                0.55 0.45 1.00
Item.12       0.66           0.44 0.56 0.00
Item.18       0.62           0.39 0.61 0.00
Item.20       0.76           0.58 0.42 0.00
Item.26       0.73           0.53 0.47 0.00
Item.29       0.89           0.80 0.20 0.00
Item.34       0.81           0.66 0.34 0.00
Item.37       0.68           0.46 0.54 0.00
Item.40       0.84           0.70 0.30 0.00
Item.42       0.77           0.59 0.41 0.00
Item.43       0.63           0.40 0.60 0.00
Item.48       0.91           0.82 0.18 0.00
Item.49       0.70           0.49 0.51 0.00
```

```
Item.51       0.76           0.58 0.42 0.00
Item.55       0.93           0.87 0.13 0.00
Item.56       0.69           0.48 0.52 0.00
Item.58       0.87           0.76 0.24 0.00
Item.2             0.86      0.75 0.25 0.00
Item.11            0.73      0.54 0.46 0.00
Item.33            0.91      0.83 0.17 0.00
Item.36            1.00      0.99 0.01 0.00
Item.52            0.52      0.27 0.73 0.00
Item.13                 0.61 0.38 0.62 0.00
Item.19                 0.87 0.76 0.24 0.00
Item.38                 0.92 0.84 0.16 0.00
Item.45                 0.56 0.31 0.69 0.00
Item.57                 0.51 0.26 0.74 0.00
```

Appendix B

Items of alternative four$_c$-factor structure

F1 (16 items)			
43	Disinhibition	Theft	I have taken items from a store without paying for them.
58	Disinhibition	Theft	I have stolen something out of a vehicle.
51	Disinhibition	Problematic Impulsivity	Others have told me they are concerned about my lack of self-control.
37	Disinhibition	Problematic Impulsivity	things are more fun if a little danger is involved.
34	Disinhibition	Fraud	I have conned people to get money from them.
49	Disinhibition	Irresponsibility	I have lost a friend because of irresponsible things I've done.
56	Disinhibition	Irresponsibility	I have had problems at work because I was irresponsible.
12	Disinhibition	Irresponsibility	I have missed work without bothering to call in.
18	Disinhibition	Irresponsibility	I've gotten in trouble because I missed too much school.
40	Meanness	Destructive Aggression	I've injured people to see them in pain.
26	Meanness	Relational Aggression	I taunt people just to stir things up.
42	Meanness	Relational Aggression	I sometimes insult people on purpose to get a reaction from them.
20	Meanness	Empathy	It doesn't bother me to see someone else in pain.
48	Meanness	Empathy	I don't care much if what I do hurts others.
55	Meanness	Empathy	It doesn't bother me when people around me are hurting.
29	Meanness	Empathy	I don't see any point in worrying if what I do hurts someone else.
F2 (5 items) ; F2∼F4			
2	Meanness	Empathy	How other people feel is important to me.
11	Meanness	Empathy	I sympathize with others' problems.
33	Meanness	Empathy	I am sensitive to the feelings of others.
36	Meanness	Empathy	I don't have much sympathy for people.
52	Meanness	Empathy	It's easy for me to relate to other people's emotions.
F3 (5 items) ; F3∼F4			
13	Boldness	Dominance	I'm a born leader.
19	Boldness	Persuasiveness	I have a knack for influencing people.

38	Boldness	Persuasiveness	I can convince people to do what I want.
57	Boldness	Persuasiveness	I'm not very good at influencing people.
45	Meanness	Excitement Seeking	Things are more fun if a little danger is involved.
F4 (9 items)			
1	Boldness	Optimism	I'm optimistic more often than not.
16	Boldness	Optimism	I have a hard time making things turn out the way I want.
7	Boldness	Resilience	I am well-equipped to deal with stress.
10	Boldness	Courage	I get scared easily.
22	Boldness	Tolerance for Uncertainty	I function well in new situations, even when unprepared.
50	Boldness	Self Confidence	I don't stack up well against most others.
44	Boldness	Social Assurance	It's easy to embarrass me.
31	Disinhibition	Boredom Proneness	I often get bored quickly and lose interest.
21	Disinhibition	Planful Control	I have good control over myself.
Dropped (23 items)			
28	Boldness	Courage	I'm afraid of far fewer things than most people.
41	Boldness	Dominance	I don't like to take the lead in groups.
4	Boldness	Intrepidness	I have no strong desire to parachute out of an airplane.
47	Boldness	Intrepidness	I stay away from physical danger as much as I can.
32	Boldness	Resilience	I can get over things that would traumatize others.
25	Boldness	Self Confidence	I don't think of myself as talented.
54	Boldness	Social Assurance	I never worry about making a fool of myself with others.
35	Boldness	Tolerance for Uncertainty	It worries me to go into an unfamiliar situation without knowing all the details.
27	Disinhibition	Alienation	People often abuse my trust.
5	Disinhibition	Dependability	I've often missed things I promised to attend.
30	Disinhibition	Dependability	I keep appointments I make.
3	Disinhibition	Impatient Urgency	I often act on immediate needs.
46	Disinhibition	Impatient Urgency	I have a hard time waiting patiently for things I want.
9	Disinhibition	Problematic Impulsivity	My impulsive decisions have caused problems with loved ones.
15	Disinhibition	Problematic Impulsivity	I jump into things without thinking.
24	Disinhibition	Theft	I have taken money from someone's purse or wallet without asking.
53	Disinhibition	Theft	I have robbed someone.

8	Meanness	Empathy	I don’t mind if someone I dislike gets hurt.
6	Meanness	Excitement Seeking	I would enjoy being in a high-speed chase.
39	Meanness	Honesty	For me, honesty really is the best policy.
14	Meanness	Physical Aggression	I enjoy a good physical fight.
17	Meanness	Relational Aggression	I return insults.
23	Meanness	Relational Aggression	I enjoy pushing people around sometimes.