The assessment of paranoia in young people: Item and test properties of the Bird Checklist of Adolescent Paranoia

Jessica C. Bird,[1,2] Bao S. Loe,[3] Miriam Kirkham,[1,2] Emma C. Fergusson,[2] Christina Shearn,[2] Hannah Stratford,[2] Ashley Teale,[1,2] Felicity Waite,[1,2] & Daniel Freeman[1,2]

Accepted manuscript in Schizophrenia Research

Date: 29th January 2020

Financial support

The study was funded by an NIHR Research Professorship to DF (RP-2014-05-003).

* Address for correspondence: Dr Jessica Bird, Oxford Cognitive Approaches to Psychosis, Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, United Kingdom, OX3 7JX.
Email: Jessica.bird@psych.ox.ac.uk.

[1] Department of Psychiatry, University of Oxford

[2] Oxford Health NHS Foundation Trust

[3] The Psychometrics Centre, University of Cambridge

ABSTRACT

Background: Precise assessment tools for psychotic experiences in young people may help identify symptoms early and facilitate advances in treatment. In this study we provide an exemplar - with a paranoia scale for youth – for improving measurement precision for psychotic experiences using item response theory (IRT). We evaluate the psychometric properties of the new measure, test for measurement invariance, and assess its potential for computerised adaptive testing (CAT).

Method: The 18-item Bird Checklist of Adolescent Paranoia (B-CAP) was completed by 1102 adolescents including 301 patients with mental health problems and 801 from the general population. After excluding outliers (n=10), IRT was used to examine item properties, test reliability, and measurement invariance. The properties of an adaptive B-CAP were assessed using a simulation of 10,000 responses.

Results: All B-CAP items were highly discriminative (a=1.15-2.76), whereby small shifts in paranoia led to a higher probability of item endorsement. Test reliability was high (a>0.90) across a wide range of paranoia severity (θ= -0.46-3.36), with the greatest precision at elevated levels. All items were invariant for gender, age, and population groups. The simulated adaptive B-CAP performed with high accuracy and required only 5-6 items at higher levels of paranoia severity.

Conclusions: The B-CAP is a reliable assessment tool with excellent psychometric properties to assess both non-clinical and clinical levels of paranoia in young people, with potential as an efficient adaptive test. In future, these approaches could be used to develop a multidimensional CAT to assess the full range of psychotic experiences in youth.

Key words: questionnaire development; computerised adaptive testing; psychosis; early intervention

1. Introduction

For many young people psychotic experiences such as hallucinations and delusional ideas are a transitory part of normal development (Kelleher et al. 2012a; Wigman et al. 2011). Yet for others, persistent psychotic experiences throughout adolescence may indicate a pluripotent risk for a range of mental health problems including schizophrenia, depression, and anxiety disorders (Kelleher et al. 2012b; Linscott & Van Os, 2013; McGorry & Mei, 2018). Reliable assessment of psychotic experiences in young people accessing services is therefore important. Although clinician-rated tools indicating the presence of psychotic symptoms and/or disorder have been well established (Yung et al. 2009), the reliability and clinical utility of self-report measures have not (Kelleher et al. 2011; Lee et al., 2016). In this study we evaluate the measurement of one of the most common psychotic experiences in young people: paranoia (the unfounded idea that other people are intending you harm) (Freeman & Garety, 2000). Despite its association with a range of psychopathology in youth (Ronald et al. 2014; Wigman et al. 2011), a lack of well-defined and age appropriate measures has perhaps been a barrier to the recognition of paranoia in young people. Most existing measures are primarily designed for adult populations with language that may be less appropriate for adolescents. We recently presented an initial validation of a new dimensional measure of paranoid thoughts specifically for youth: The Bird Checklist of Adolescent Paranoia (B-CAP; Bird et al. 2019). We now extend this initial validation to evaluate the item and test properties of the B-CAP using item response theory (IRT), whilst also arguing for the wider application of IRT to the measurement of psychotic experiences in youth.

There are several key issues potentially affecting the validity and reliability of measures of psychotic experiences in young people. First, using measures created for adults is a common practice with young people. However, it is possible that item content based on adult descriptions do not adequately capture adolescent psychotic experiences. Further, age may influence the way items are interpreted. The rates of false positives in self-report measures of psychotic experiences are thought to be high in young people, with certain items producing more false positives than others (Kelleher et al. 2011). Second, assessment tools vary in the overall severity of psychotic experiences they measure, potentially impacting precision across different populations. For example, a test consisting of items representing mildly elevated presentations may perform well in non-clinical populations but have low reliability in those with a diagnosis of schizophrenia (where a ceiling effect may instead occur). Third, measures of psychotic experiences frequently include a broad range of symptoms in a single scale, summed together to provide a total score. Yet individual psychotic experiences such as paranoia, hearing voices, and cognitive disorganisation are qualitatively distinct phenomenon, with numerous studies demonstrating they form separate

factors and differ in underlying aetiology (Zavos et al. 2014). These differences are lost in tools that sum together multiple psychotic experiences, often including an unequal balance of items for individual domains. Further, items within a scale often vary in the level of severity they represent, leading to items with potentially little clinical relevance being assigned the same value as ones that strongly discriminate clinical symptomology. This can lead to imprecise estimates of psychotic experiences whilst reducing the ability to meaningfully interpret and compare summed scores (Gibbons et al. 2016).

Modern psychometric methods such as item response theory (IRT) provide innovative opportunities to improve measurement precision. Rather than relying on a count of item endorsement, IRT estimates severity on a continuum derived from the relationship between participant responses and differences in the ability of each item to measure the problem. Thus, IRT produces more precise estimates. The approach also evaluates the level of severity each item typically measures, allowing inferences to be made about which population the test may be most reliable for. IRT can also be used to examine differential item functioning between demographic groups to prevent bias within a scale. Another advantage of IRT is the ability to create computerised adaptive tests (CAT) that dynamically select items matched to a person's severity. Fewer items are typically required to reach a similar level of precision as the full questionnaire, providing the opportunity to reduce patient burden in lengthy clinical and research assessments (Gibbons et al. 2016). CAT presents a novel solution to ensure the reliable measurement of individual psychotic experiences whilst also minimising assessment burden.

In this study we offer an exemplar for using IRT to develop precise assessments for psychotic experiences in young people. With a combined sample of adolescent patients attending mental health services and adolescents from the general population, we use IRT to evaluate the item properties and test reliability of the B-CAP across the continuum of paranoia severity. Next, we test for differential item functioning (DIF) between genders, older and younger adolescents, and between the clinical and non-clinical groups. Finally, using a CAT simulation we evaluate the potential to administer the B-CAP adaptively to limit the number of items required to estimate paranoia reliably.

2. Method

2.1. Participants

Participants were 1102 adolescents aged 11-17 years including 301 help-seeking patients attending child and adolescent mental health services (mean age=15.1, SD=1.75, female *n*=184, male *n*=117,

White British *n*=240) and 801 adolescents from the general population (mean age= 13.3, SD=1.16, female *n*=410, male *n*=382, other gender *n*=9, White British *n*=629). Participants in the clinical sample were seeking help for a range of presenting problems, most commonly affective disorders and neurodevelopmental disorders (see supplementary materials). Seven participants had a diagnosis of psychosis (2.0%).

Participants from the clinical sample were recruited during routine appointments at a community outpatient child and adolescent mental health service (*n*=271) and an adolescent inpatient unit (*n*=30) in Oxfordshire (Bird et al. in prep). Patients aged 11-17 years accessing these services were invited to take part, regardless of diagnosis. Exclusion criteria were a moderate/severe learning disability or inability to complete questionnaires in English. Informed parental consent and child assent was obtained for young people prior to completing the questionnaire in the clinic. Participants from the general population were recruited from a secondary school in Leicestershire, United Kingdom, as reported in Bird et al. (2019). All pupils aged 11-15 years were invited to take part using opt-out parental consent and pupils who provided written assent completed the questionnaire as part of the larger study pack within a 60-minute lesson.

2.2. Assessments

The B-CAP is an 18-item scale, developed by our research team (Bird et al. 2019), assessing the frequency of paranoid thoughts in the last two weeks on a 6-point scale (0=*Never*, 5=*All the time*). Higher scores indicate higher paranoia. The full scale is provided in the appendix at the end of this paper. Items were generated from the clinical expertise of the research team, consideration of existing paranoia measures, and comments made by young people. Eleven adolescents aged 12-16 from a secondary school in Oxfordshire and three adolescents aged 15-16 years receiving inpatient mental health care met with JB and provided their perspective on important components paranoia for young people and gave suggestions for the content, wording, response format, and layout of questionnaire. The final 18 items were selected from a broader item pool following exploratory and confirmatory factor analysis. The final hierarchical model of a second-order paranoia factor and three sub-factors of social harm, conspiracy ideas, and physical threat showed excellent fit statistics (Bird et al. 2019).

The B-CAP has concurrent validity with other measures of paranoia, with data from Bird et al. (2019) showing a correlation of r=0.84 (p<0.001) with the paranoia subscale of the Specific Psychotic Experiences Questionnaire (SPEQ; Ronald et al. 2014) and r=0.68 (p<0.001) with the Social Mistrust Scale (SMS; Wong et al. 2014). Although all paranoia measures are limited in their ability to determine whether thoughts are truly unfounded, we have shown that B-CAP scores are

moderate correlated (r=0.41, p<0.001) with participant ratings on a visual analogue scale (VAS) that they are 'more fearful of others than they should be' (Bird et al. 2019). This was significantly larger than the small correlation of r=0.25 between the same VAS rating and bullying scores ($\tilde{\chi}$=5.35, p<0.001). This provides additional confidence that the B-CAP is a valid tool to identify likely excessive concerns about others that are distinct from genuine victimisation.

2.3. Statistical analysis

All analyses were conducted in R, version 3.6.1 (R Core Team, 2013). Individual packages used included "psych" (Revelle, 2017), "mirt" (Chalmers, 2012), "mokken" (Ark, 2015), "lordif" (Choi et al. 2011), and "catR" (Magis & Raîche, 2011). To assess the full spectrum of paranoia severity, the analysis was conducted on the combined clinical and non-clinical data. Rates of missing data were very low (<2%) and missing values were imputed using the mice package (van Buuren & Groothuis-Oudshoorn, 2011).

Although the B-CAP has been shown to consist of three subtypes of paranoia (Bird et al. 2019), the single overarching paranoia factor that explained these sub-factors in a hierarchical model indicates the scale can be considered unidimensional for the purpose of IRT. A confirmatory factor analysis in the combined clinical and non-clinical sample confirmed that this hierarchical model had a good fit to the data (supplementary materials). We used Mokken analysis to further evaluate item homogeneity, with Loevinger's H coefficients ≥0.3 indicating unidimensionality (Stochl et al. 2012). This showed all 18 items conformed to a single dimension with item coefficients above 0.3 and an overall homogeneity coefficient of 0.474 (SE=0.018).

As the response options were polytomous, a two-parameter graded response model (GRM; Samejima, 1969) was fitted to the B-CAP items. To identify outliers with atypical response patterns, participants with extreme person fit statistic scores (±3) were excluded (Felt et al. 2017). The IRT parameters are expressed as a function of theta, representing the continuum of the latent trait (i.e. paranoia), with values denoting standard deviations from average trait paranoia (i.e. theta of 0). Higher theta values therefore indicate greater severity of paranoia. Unlike ability constructs often used in the IRT literature (e.g. intelligence), the population distribution of paranoia is expected to be positively skewed with most people reporting minimal levels (Bebbington et al. 2013). As a result, average trait paranoia would represent the lower end of the severity spectrum. The discrimination parameters (a) describe how well each item discriminates different levels of theta, with higher values signifying small shifts in paranoia severity produce rapid increases in the probability of item endorsement. Discrimination values ≥1 are highly discriminative whilst those ≤0.5 are unacceptable (Baker & Kim, 2017). The difficulty parameter indicates the severity level

each item response typically represents, with higher values suggesting the item assesses more severe presentations. Five difficulty parameters are given for each item, representing the theta level where there is a 50% probability of responding between the threshold of each of the six response options (*b*1=0-1, *b*2=1-2, *b*3=2-3, *b*4=3-4, *b*5=4-5).

To assess measurement invariance between groups, differential item functioning (DIF) analysis was conducted for gender, age, and sample population. Item variance, or DIF, indicates a bias in measurement whereby participants from different demographic groups with the same level of trait paranoia respond differently to the items (Holland & Wainer, 2012). A beta change above 10% and a pseudo R-square above 0.13 were used as the criteria to identify items with DIF (Crane et al. 2007; Choi et al. 2011).

The overall reliability of the B-CAP was primarily assessed using the test information (TI) function. This denotes the scale precision as a function of theta, showing at which levels of severity the scale has high and low reliability. For interpretability, the formula $1/\sqrt{TI(\theta)}$ was used to convert TI values at specific theta levels to an equivalent alpha on a scale of 0-1 (O'Connor, 2018). The expected score function based on the GRM was used to assess the likely score for individuals at different points of the severity spectrum and establish interpretative score ranges.

Using the IRT parameters derived from the current sample, we conducted a CAT simulation with 10,000 simulated responses to evaluate the mean number of items required at different levels of paranoia severity. Item selection is determined using the maximum Fisher Information criterion. A Bayesian modal estimation that temporally assumes a normal distribution is used to estimate theta at the start of the simulation. A non-Bayesian maximum likelihood estimation is employed to determine a participant's final theta (i.e. severity). The simulation stopping rule was a standard error (SE) of ≥0.32, equivalent to a reliability of ≥0.90. The correlation between theta scores derived from the CAT and those obtained from all 18 items were then computed.

3. Results

Following removal of the 10 participants with extreme person fit statistics (non-clinical *n*=5, clinical *n*=5), IRT analysis was conducted on the final sample of 1092 participants (non-clinical *n*=796, clinical *n*=296). The two parameter GRM provided a good fit to the data (CFI=0.97, TLI=0.96, RMSEA=0.053, SRMSR=0.076). The item parameters are displayed in Table 1, and the item category response curves (CRCs) for all items are shown in the supplementary materials.

Item properties

Discrimination parameters were high for all 18 items (*a*=1.14-2.77), suggesting small shifts in paranoia severity lead to an increased probability that items will be endorsed. The item "*People are making sly comments to upset me*" was the most discriminating item (*a*=2.77). The difficulty parameters show all items measure a broad range of paranoia severity from average to severe across the response options (0-5). Full endorsement of all items (*b*5, item response 4-5) represented a high severity of paranoia at 2.10-3.90 standard deviations above the average level. High difficulty parameters for *b*1, representing a response of 0-1, suggested that any endorsement of the following items were particularly indicative of heightened paranoia severity (>0.85 S.D. above average): *"people are collecting my information or photos to use against me"* (*b*1=1.23), "*people will try to kidnap me*" (*b*1=0.96), *"I feel like I am being followed or stalked"* (*b*1=0.96), *"Nasty tricks are being played on me"* (*b*1=0.88), and *"Groups of people are planning against me"* (*b*1=0.85).

Table 1. IRT parameters for B-CAP items with combined non-clinical and clinical sample (*n*=1091). Standard errors are shown in parentheses.

Item	*a*	b_1	b_2	b_3	b_4	b_5
1	2.41 (0.13)	0.10 (0.05)	0.58 (0.05)	1.23 (0.06)	1.77 (0.08)	2.36 (0.12)
2	2.09 (0.12)	0.21 (0.05)	0.71 (0.06)	1.51 (0.08)	2.10 (0.10)	2.46 (0.13)
3	2.40 (0.13)	0.09 (0.05)	0.70 (0.05)	1.40 (0.07)	1.94 (0.09)	2.57 (0.13)
4	1.75 (0.11)	0.70 (0.06)	1.43 (0.08)	1.95 (0.11)	2.51 (0.15)	3.02 (0.19)
5	1.99 (0.12)	0.24 (0.05)	0.96 (0.06)	1.58 (0.08)	2.16 (0.11)	2.63 (0.14)
6	2.77 (0.16)	0.21 (0.04)	0.83 (0.05)	1.41 (0.07)	2.03 (0.09)	2.41 (0.12)
7	2.44 (0.13)	-0.24 (0.05)	0.41 (0.05)	1.09 (0.06)	1.66 (0.08)	2.10 (0.10)
8	2.24 (0.12)	0.08 (0.05)	0.69 (0.05)	1.30 (0.07)	1.92 (0.09)	2.34 (0.12)
9	2.27 (0.15)	0.88 (0.06)	1.51 (0.08)	2.29 (0.12)	2.75 (0.15)	2.94 (0.17)
10	1.88 (0.11)	0.41 (0.05)	1.08 (0.07)	1.79 (0.09)	2.29 (0.12)	2.73 (0.15)
11	2.57 (0.17)	0.85 (0.05)	1.36 (0.07)	1.82 (0.09)	2.18 (0.11)	2.48 (0.12)
12	1.85 (0.14)	1.23 (0.08)	1.83 (0.11)	2.45 (0.15)	2.78 (0.17)	3.20 (0.21)
13	1.83 (0.12)	0.80 (0.06)	1.51 (0.09)	2.06 (0.11)	2.50 (0.14)	2.79 (0.16)
14	1.51 (0.11)	0.96 (0.07)	1.63 (0.10)	2.34 (0.14)	2.94 (0.19)	3.36 (0.23)
15	1.22 (0.08)	-0.29 (0.07)	0.61 (0.07)	1.33 (0.10)	2.00 (0.13)	2.53 (0.16)
16	1.14 (0.09)	0.96 (0.09)	1.71 (0.13)	2.42 (0.18)	3.24 (0.25)	3.90 (0.32)
17	1.49 (0.09)	0.11 (0.06)	0.89 (0.07)	1.49 (0.09)	1.97 (0.12)	2.35 (0.14)
18	1.90 (0.12)	0.44 (0.05)	0.97 (0.06)	1.47 (0.08)	1.84 (0.10)	2.39 (0.13)

3.1. Differential Item Functioning (DIF)

There was no evidence of significant DIF between younger (aged 11-13 years, *n*=509) and older (aged 14-17, *n*=583) adolescents, girls (*n*=590) and boys (*n*=494), or participants in the non-clinical

(*n*=796) and clinical (*n*=296) samples. In each of these three DIF analyses, none of the 18 items showed a pseudo R^2 change of more than 0.13 or a beta change of more than 10%.

We then adopted a stricter criterion of a beta change of more than 5% to assess milder levels of DIF. With this stricter criterion, one item was flagged for DIF in the participant group analysis (item 5: *"People are trying to embarrass me in class on purpose")* and 4 items were flagged for DIF in the gender analysis (item 9: *"Nasty tricks are being played on me"*; item 13: *"I'm sure people are seeking revenge on me"*; item 15: "*I am scared of what strangers will do to me"*; and item 16: *"People will try to kidnap me"*). No items were flagged using the stricter criterion for the age analysis. The item characteristic plots and the test characteristic curves for the items with mild DIF are shown in the supplementary materials. The items flagged for DIF under the stricter criterion did not have an impact on the total scores, with high correlations between theta scores from all 18 items and the DIF adjusted scores for both participant group (*r*=0.997, item 5 omitted) and gender (*r*=0.992, items 9, 13, 15, & 16 omitted). The mild DIF in these items can therefore be disregarded and differences between demographic groups can be meaningfully interpreted.

Paranoia scores were significantly higher in the clinical sample (mean=20.1, SD=18.2) compared to the general population sample (mean=12.5, SD=14.0; *t*=6.56, *df*=440.7, *p*<0.001), and in girls (mean=18.2, SD=16.8) compared to boys (mean=9.84, SD=12.2; *t*=9.44, *df*=1070.2, *p*<0.001). Older adolescents also had significantly higher paranoia scores (mean=16.5, SD=16.5) than younger adolescents (mean=12.3, SD=14.2; *t*=4.44, *df*=1100.9, *p*<0.001), although the magnitude of this difference was smaller.

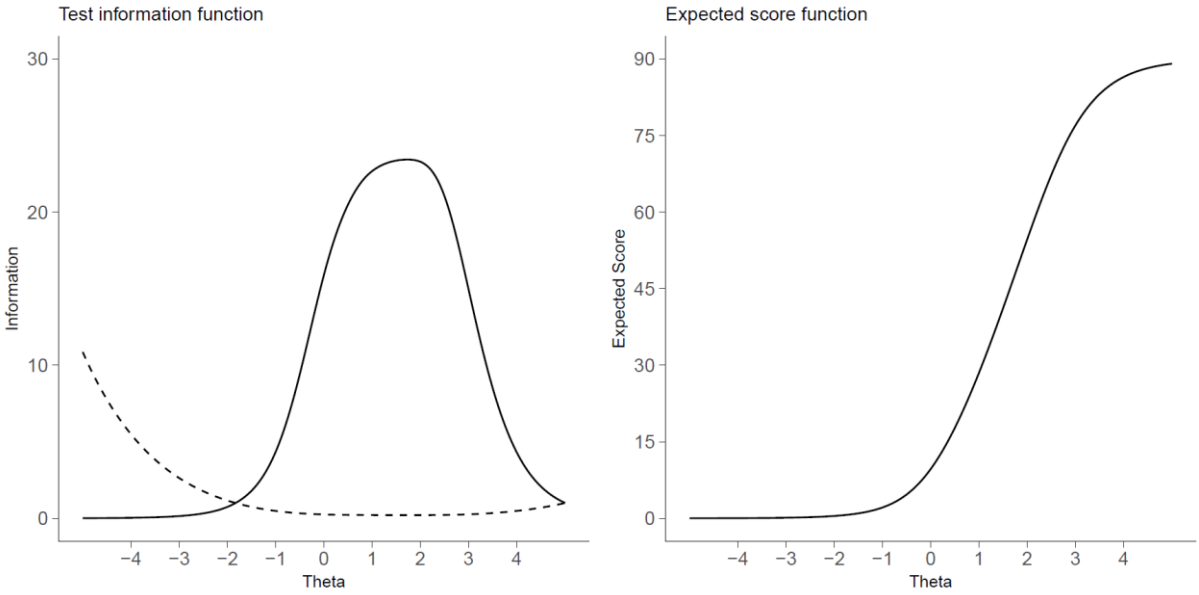

Figure 1. B-CAP test information (TI) with standard errors (----) and expected score across the theta distribution.

3.2. Expected scores

The B-CAP total score had high precision, with a correlation of 0.92 between scores derived from summing the 18 items and theta scores from the GRM. The expected score function in Figure 1 highlights the anticipated positive skew of paranoia where the average adolescent would be expected to endorse the paranoid thought items to a small degree, with expected scores of 9.68 out of 90 at the average level of trait paranoia (theta=0) and 17.8 at 0.5 SDs above average (i.e. theta = 0.5). Higher scores reflect higher levels of paranoia severity, with expected scores of 28.5 at 1 SD above average, 41.2 at 1.5 SD, 54.7 at 2 SD, 67.3 at 2.5 SD, and 77.0 at 3 SD above average trait paranoia. Using the expected score function, we provide descriptive score categories to facilitate interpretation of the total B-CAP score (Table 2).

Table 2. Interpretive ranges for B-CAP total score

Category	Score range	Theta range	Comment
Average	0-22	≤ 0.70	Scores in this range represent normal levels of suspicious thinking at less than 0.70 SDs above average for adolescents
Mildly elevated	23-39	0.75 to 1.40	Scores in this range represent mildly elevated suspiciousness at 0.75-1.40 SDs above average for adolescents.
Moderate	40-53	1.45 to 1.95	Scores in this range represent moderate paranoia at 1.45-1.95 SDs above average for adolescents.
High	54-70	2.00 to 2.60	Scores in this range represent high levels of paranoia at 2.00-2.60 SDs above average for adolescents.
Severe	71-90	≥ 2.65	Scores in this range represent severe levels of paranoia at greater than 2.65 SDs above average for adolescents

3.3. Test reliability

The test information (TI) function (Figure 1) represents the reliability of the B-CAP as a function of paranoia severity (i.e. theta) (see supplementary materials for individual information functions). As shown in Figure 1, the B-CAP demonstrated excellent reliability and precision across a wide range of the paranoia spectrum. Equivalent alpha values were greater than 0.90 (TI=10) between 0.46 SD below and 3.36 SD above average trait paranoia (SE=0.21-0.31), representing expected total scores between 5 and 82 (maximum score = 90). The highest reliability was between 0.44-2.60 SD above average, representing expected total scores between 17 and 69, with equivalent alpha values ≥0.95 (TI=20) and standard errors below 0.22 in this range. The maximum TI of 23.4, equivalent to a=0.96, was at a theta of 1.73 (SE=0.21). Reliability only fell into the acceptable range ($\alpha < 0.80$) beyond 3.90 SDs above (expected scores

of 87-90) and 0.94 below (expected scores of 0-1) average. These findings suggest the B-CAP has high reliability for assessing both non-clinical and clinical levels of paranoia.

3.4. Computer adaptive testing simulation

The results of the CAT simulation are shown in Table 3. The sample of 10,000 simulated participants are split into 10 equal decile (D) ranks representing the theta spectrum. The average test length was 10.8 items (SD= 5.15) with a mean Root Mean Square Error (RMSE) of 0.361 and a mean bias of 0.0036. The number of items administered decreased incrementally across each decile rank of the theta spectrum, indicating a smaller number of items are sufficient to get an accurate estimate of paranoia as severity increases. Only 5-6 items were administered on average from D8-D10 (mean theta=0.68-1.75). This pattern is consistent with the TI showing the B-CAP has the greatest reliability across the higher end of the severity spectrum (i.e. D8-D10). The CAT demonstrated high accuracy with a correlation of 0.93 between theta scores derived from all 18 items and the CAT estimated scores.

Table 3. CAT simulation of 10,000 respondents showing average test length across 10 decile (D) ranks of the theta spectrum

	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
Mean Theta	-1.78	-1.07	-0.69	-0.39	-0.11	0.14	0.40	0.68	1.04	1.75
Mean test length	18.0	17.8	16.4	13.7	10.2	8.10	6.70	6.15	5.67	5.68
RMSE	0.62	0.35	0.38	0.36	0.30	0.28	0.29	0.29	0.31	0.31
Mean SE	0.52	0.45	0.38	0.33	0.31	0.31	0.30	0.30	0.30	0.30
Mean bias	0.45	0.03	-0.03	-0.02	0.00	-0.04	-0.03	-0.06	-0.10	-0.16
Stop rule satisfied	0.00	0.05	0.25	0.59	0.88	0.97	1.00	1.00	1.00	1.00
Number of simulees	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Note: RMSE = Root mean square error, SE = standard error

4. Discussion

Paranoia and excessive mistrust may be an overlooked issue in youth with potentially detrimental effects on social functioning (Bird et al. 2019). Determining its true impact will depend on precise measurement. Here we use IRT to extensively evaluate the psychometric properties and precision of a new measure of paranoia specifically designed for young people. The IRT analyses show the B-CAP items are highly discriminative of shifts in adolescent paranoia across the spectrum of severity, with higher scores representing more severe presentations. Reliability was excellent across a wide range of paranoia severity, from the average levels expected in most adolescents to the more

severe presentations likely in clinical populations. Importantly, reliability was highest for elevated levels of paranoia and remained high even at the extreme end of the severity spectrum. Furthermore, all items functioned similarly between boys and girls, between older and younger adolescents, and between young people from the general population and those seeking help from mental health services. This measurement invariance indicates that differences in total score between these demographic groups are unlikely due to bias within the questionnaire. Overall, these findings suggest the B-CAP is a reliable and sensitive tool to assess both non-clinical and clinical levels of paranoia in young people.

The B-CAP should have utility as a clinical measure in adolescent mental health services. The broad similarity of items, both in their ability to discriminate shifts in paranoia and in the level of severity each item response represents, suggest that total summed scores can be meaningfully interpreted. Indeed, the high correlation between the summed scores and the IRT derived theta scores suggest the total score has good precision for routine clinical use. However, computerised administration to calculate theta scores would lead to even higher precision in estimating paranoia severity, whilst also allowing the possibility of adaptive testing to reduce the number of items administered. Our CAT simulation showed that only 5-6 items on average were required at the higher end of the severity spectrum to reliably estimate paranoia. Although an 18-item scale is already relatively short, when administered alongside many other measures within clinical and research assessments this item reduction may help decrease patient burden. This is especially important considering help-seeking young people with higher levels of paranoia severity will more likely be distressed, experience emotional dysregulation, and have difficulties concentrating. Shorter assessments including only relevant items may therefore improve the patient experience and facilitate engagement.

Improvements in technology within health care settings mean computerised assessments are becoming feasible in routine clinical practice. Yet only a few studies have so far used IRT and CAT to evaluate assessments for psychosis (Batterham et al. 2016; Kim et al. 2013; Laurens et al. 2012; van Bebber et al. 2017), most of which have included measures that combine a range of separable psychotic experience within the same scale. This study is the first to evaluate the item properties and application of CAT to an individual psychotic experience, and we are working to make the adaptive B-CAP freely available online. However, the extension of our findings to calibrate multiple item banks to each assess different psychotic experiences will now be an important task. Using IRT to assess individual item properties within each bank will ensure items can reliably discriminate psychotic experiences at clinically relevant levels of severity, whilst adaptive testing will allow multiple domains to be administered in a shorter time to reduce patient burden. This

would offer more precise estimates of an individual's current level of different psychotic experiences, providing reliable information about which domains may be the most pertinent for treatment.

4.1. Limitations

There are notable limitations of this study. Although the combination of participants from the general population and patients attending mental health services allowed a range of the paranoia spectrum to be assessed, only a small minority of our sample had psychosis and we did not include data from young people with diagnosed persecutory delusions. As a result, it is possible our analysis underrepresented the extreme end of the spectrum. Obtaining normative scores for the B-CAP from young people with persecutory delusions will now be helpful. However, we were able to use the IRT model to derive expected scores at different levels of paranoia with score ranges to aid clinical interpretation. Notably, the item categorical response curves (supplementary materials) suggest fewer response options would likely be adequate to sample paranoia severity from the B-CAP items. However, our view was that collapsing response categories would reduce the scale's sensitivity to detect individual fluctuations in paranoia in clinical practice. This is in line with service user input in the design phase where young people said a proposed four point scale was not specific enough and it was expanded to six at their suggestion.

Importantly, there is always an unavoidable level of measurement error in self-report paranoia questionnaires from genuine experiences of hostility. Although we recently showed B-CAP scores are moderately associated with participant ratings that their fears of others are excessive and distinct from bullying (Bird et al. 2019), validation of the B-CAP with clinician-rated assessment tools or experimental procedures such as virtual reality (Freeman et al. 2010) would be beneficial. It will also be necessary to determine the test-retest reliability of the B-CAP in future studies. Another limitation is the use of a small number of items within the CAT analysis, since a broader range of items to choose from when matching items to participants can improve reliability. Yet even with only 18 items our CAT simulation was still able to substantially reduce the number of items without a loss of precision. However, as our whole sample was used to calibrate the item bank, we were only able to use simulated responses in the CAT analysis. Assessing the functioning of the adaptive version of the B-CAP in a separate validation sample will now be required. Adopting a multidimensional IRT approach would also be beneficial to assess the capacity of the CAT to sample from each of the three sub-domains of paranoia in the B-CAP. Overall, the study potentially provides an exemplar of how to address common issues in the measurement of psychotic experiences in young people. Precise assessment tools are essential to facilitate advances in research and treatment.

Acknowledgement

We would like to thank the young people who participated in this study and the child and adolescent mental health services in Oxford Health NHS Foundation Trust who supported the research.

References

Baker, F., Kim, S., 2017. The basics of item response theory Using R. Springer International Publishing, Cham, Switzerland.

Batterham, P.J., Sunderland, M., Carragher, N., Calear, A.L., 2016. Development and community-based validation of eight item banks to assess mental health. Psychiatry Research 243, 453–462.

Bebbington, P.E, McBride O, Steel C, Kuipers E, Radovanovic, M, Brugha T, Jenkins R, Meltzer HI, Freeman D. 2013. The structure of paranoia in the general population. Br. J. Psychiatry, 202, 419-427.

Bird, J.C., Evans, R., Waite, F., Loe, B.S., Freeman, D., 2019. Adolescent paranoia: Prevalence, structure, and causal mechanisms. Schizophr. Bull. 45(5), 1134-1142.

Chalmers, R.P., 2012. mirt: A multidimensional item response theory package for the R environment. J. Stat. Soft. 48, 1–29.

Choi, S.W., Gibbons, L.E., Crane, P.K., 2011. lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. J. Stat Soft. 39, 1–30.

Crane, P.K., Gibbons, L.E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R.D., Teresi, J.A., 2007. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual. Life. Res. 16, 69–84.

Felt, J.M., Castaneda, R., Tiemensma, J., Depaoli, S., 2017. Using person fit statistics to detect outliers in survey research. Front. Psychol. 8, 1–9.

Freeman, D., Garety, P., 2000. Comments on the contents of persecutory delusions: does the definition need clarification? Br. J. Clin. Psychol. 39, 407–414.

Freeman, D., Pugh, K., Vorontsova, N., Antley, A., Slater, M., 2010. Testing the continuum of delusional beliefs: An experimental study using virtual reality. J. Abnorm. Psychol. 119, 83–92.

Gibbons, R.D., Weiss, D., Frank, E., Kupfer, D., 2016. Computerized adaptive diagnosis and testing of mental health disorders. Annu. Rev. Clin. Psychol. 12, 83–104.

Holland, P., Wainer, H., 2012. Differential item functioning. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Kelleher. I., Connor, D., Clarke, M.C., Devlin, N., Harley, M., Cannon, M., 2012a. Prevalence of psychotic symptoms in childhood and adolescence: A systematic review and meta-analysis of population-based studies. Psychol. Med. 42, 1857–1863.

Kelleher, I., Harley, M., Murtagh, A., Cannon, M., 2011. Are screening instruments valid for psychotic-like experiences? A validation study of screening questions for psychotic-like experiences using in-depth clinical interview. Schizophr. Bull. 37, 362–369.

Kelleher, I., Keeley, H., Corcoran, P., Lynch, F., Fitzpatrick, C., Devlin, N., Molloy, C., Roddy, S., Clarke, M.C., Harley, M., Arseneault, L., Wasserman, C., Carli, V., Sarchiapone, M., Hoven, C., Wasserman, D., Cannon, M., 2012b. Clinicopathological significance of psychotic experiences in non-psychotic young people: Evidence from four population-based studies. Br. J. Psychiatry. 201, 26–32.

Kim. Y., Chang, J.,S., Hwang, S., Yi, J.S., Cho, I.H., Jung, H.Y., 2013. Psychometric properties of Peters et al. Delusions Inventory-21 in adolescence. Psychiatry Res. 207, 189–194.

Laurens, K.R., Hobbs, M.J., Sunderland, M., Green, M.J., Mould, G.L., 2012. Psychotic-like experiences in a community sample of 8000 children aged 9 to 11 years: An item response theory analysis. Psychol. Med. 42, 1495–1506.

Lee, K.W., Chan, K.W., Chang, W.C., Lee, E.H.M., Hui, C.L.M., Chen, E.Y.H., 2016. A systematic review on definitions and assessments of psychotic-like experiences. Early Interv. Psychiatry. 10, 3–16.

Linscott, R.J., Van Os, J., 2013. An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: On the pathway from proneness to persistence to dimensional expression across mental disorders. Psychol. Med. 43, 1133–1149.

Magis, D., Raîche, G., 2011. catR: An R package for computerized adaptive testing. Appl. Psychol. Meas. 35, 576–577.

McGorry, P.D., Mei, C., 2018. Early intervention in youth mental health: Progress and future directions. Evid. Based Ment. Heal. 21, 182–184.

O'Connor, B.P., 2018. An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability. Psychol. Assess. 30, 991–1003.

Revelle, W., 2017. psych: Procedures for psychological, psychometric, and personality research. Software.

Ronald, A., Sieradzka, D., Cardno, A.G., Haworth, C.M.A., McGuire, P., Freeman, D., 2014. Characterization of psychotic experiences in adolescence using the specific psychotic experiences questionnaire: Findings from a study of 5000 16-year-old twins. Schizophr. Bull. 40, 868–877.

Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. Psychometrika 34, 1–97.

Stochl, J., Jones, P.B., Croudace, T.J., 2012. Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. BMC Med. Res. Methodol. 12.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2013. http://www.R-project.org/2013.

van Bebber, J., Wigman, J.T.W., Meijer, R.R., Ising, H.K., van den Berg, D., Rietdijk, J., Dragt, S., Klaassen, R., Nieman, D., de Jonge, P., Sytema, S., Wichers, M., Linszen, D., van der Gaag, M., Wunderink, L., 2017. The Prodromal Questionnaire: A case for IRT-based adaptive testing of psychotic experiences? Int. J.Methods Psychiatr. Res. 26, 1–11.

Van Buuren, S., Groothuis-oudshoorn, K., 2011. MICE: Multivariate imputation by chained equations in R. J. Stat. Soft. 45.

Van der Ark, L.A., 2015. New developments in mokken scale analysis in R. J. Stat. Soft. 48(5).

Wigman, J.T.W., Vollebergh, W.A.M., Raaijmakers, Q.A.W., Iedema, J., Van Dorsselaer, S., Ormel, J., Verhulst, F.C., Van Os, J., 2011. The structure of the extended psychosis phenotype in early adolescence: A cross-sample replication. Schizophr. Bull. 37, 850–860.

Wong, K.K., Freeman, D., Hughes, C., 2014. Suspicious young minds: Paranoia and mistrust in 8- To 14-year-olds in the UK and Hong Kong. Br. J. Psychiatry 205, 221–229.

Yung, A.R., Pan Yuen, H., Mcgorry, P.D., Phillips, L.J., Kelly, D., Dell'olio, M., Francey, S.M., Cosgrave, E.M., Killackey, E., Stanford, C., Godfrey, K., Buckby, J., 2009. Mapping the onset of psychosis: The comprehensive assessment of at-risk mental states. Aust. New Zeal. J. Psychiatry 39, 964–971.

Zavos, H.M.S., Freeman, D., Haworth, C.M.A., Mcguire, P., Plomin, R., Cardno, A.G., Ronald A., 2014. Consistent etiology of severe, frequent psychotic experiences and milder, less frequent manifestations: A twin study of specific psychotic experiences in adolescence. JAMA Psychiatry 71, 1049–1057.

Appendix 1 - The Bird Checklist of Adolescent Paranoia

This form is about worries you may have about other people. Please circle how often you have had each thought over the last *2 weeks*.

	Never	Once	Couple of times	Few times a week	Every day	All the time
1. People at school are trying to make me feel unwanted	0	1	2	3	4	5
2. I'm sure people are gossiping about me on social media	0	1	2	3	4	5
3. I am being pushed out of conversations on purpose	0	1	2	3	4	5
4. My friends or partner are ignoring my messages to upset me	0	1	2	3	4	5
5. People are trying to embarrass me in class on purpose	0	1	2	3	4	5
6. People are making sly comments to upset me	0	1	2	3	4	5
7. I think people are lying to me on purpose	0	1	2	3	4	5
8. People say things under their breath to wind me up	0	1	2	3	4	5
9. Nasty tricks are being played on me	0	1	2	3	4	5
10. People are trying to confuse me on purpose	0	1	2	3	4	5
11. Groups of people are planning against me	0	1	2	3	4	5
12. People are collecting my information or photos to use against me	0	1	2	3	4	5
13. I'm sure people are seeking revenge on me	0	1	2	3	4	5
14. I feel like I am being followed or stalked	0	1	2	3	4	5
15. I am scared of what strangers will do to me	0	1	2	3	4	5
16. People will try to kidnap me	0	1	2	3	4	5
17. I could be attacked at any time	0	1	2	3	4	5
18. I feel unsafe around people everywhere I go	0	1	2	3	4	5

Scoring instructions:

Add together responses for all 18 items to obtain the total paranoia score. To obtain subscale scores, add together responses for items 1-8 (social harm), 9-13 (conspiracy), and 14-18 (physical threat).

Appendix 2 - Supplementary materials

1. Diagnostic characteristics of clinical sample

Table S1. Primary diagnoses / presenting problems for participants in clinical sample included in the final IRT analysis (*n*=296).

Presenting problem	*n*	%
Anxiety / depression	191	63.5
Autism spectrum disorder	79	26.2
Emotion dysregulation, self-harm, & suicidality	78	25.9
Attention deficit hyperactivity disorder	41	13.6
Anger / conduct problems	30	9.6
Disordered eating	24	8.0
Trauma	23	7.6
Sleep problems	19	6.3
Gender identity issues	8	2.7
Family relationship issues	8	2.7
Psychosis	7	2.3
Substance misuse	6	2.0
Tic disorders	5	1.7

Note: Diagnostic characteristics obtained from patients' clinical records.

2. Confirmatory factor analysis (CFA)

In Bird et al. (2019) we show that the B-CAP consists of three first-order factors (social harm, conspiracy ideas, & physical threat) that are explained by an overarching second-order paranoia factor. Here we conducted a CFA using the combined non-clinical and clinical data used in this study to evaluate the fit of this hierarchical three-factor model to the data. CFA was conducted in lavaan (Rosseel, 2012) using the MLR robust maximum likelihood estimator to account for non-normality in the data. Based on recommended criteria (Hu & Bentler, 1999), the CFA demonstrated that the model had a good fit to the data (CFI=0.951, TLI=0.943, RMSEA=0.060, SRMR=0.046). The first and second order factor loadings are shown in Table S2.

Table S2. First and second-order factor loadings

First order loadings	**Social harm**	**Conspiracy**	**Physical threat**
1. Feel unwanted	0.779		
2. Gossip on social media	0.715		
3. Pushed out of conversations	0.783		
4. Ignoring messages	0.615		
5. Embarrass in class	0.711		
6. Sly comments	0.814		
7. Lying on purpose	0.764		
8. Say things under their breath	0.745		
9. Nasty tricks		0.743	
10. Confuse me on purpose		0.694	
11. Groups planning		0.815	
12. Collecting photos		0.655	
13. Seeking revenge		0.719	
14. Followed or stalked			0.607
15. Scared strangers			0.729
16. Kidnap me			0.709
17. Attacked anytime			0.802
18. Unsafe everywhere			0.755
Second order loadings	**Paranoia**		
Social harm	0.920		
Conspiracy	0.916		
Physical threat	0.729		

3. Category response curves

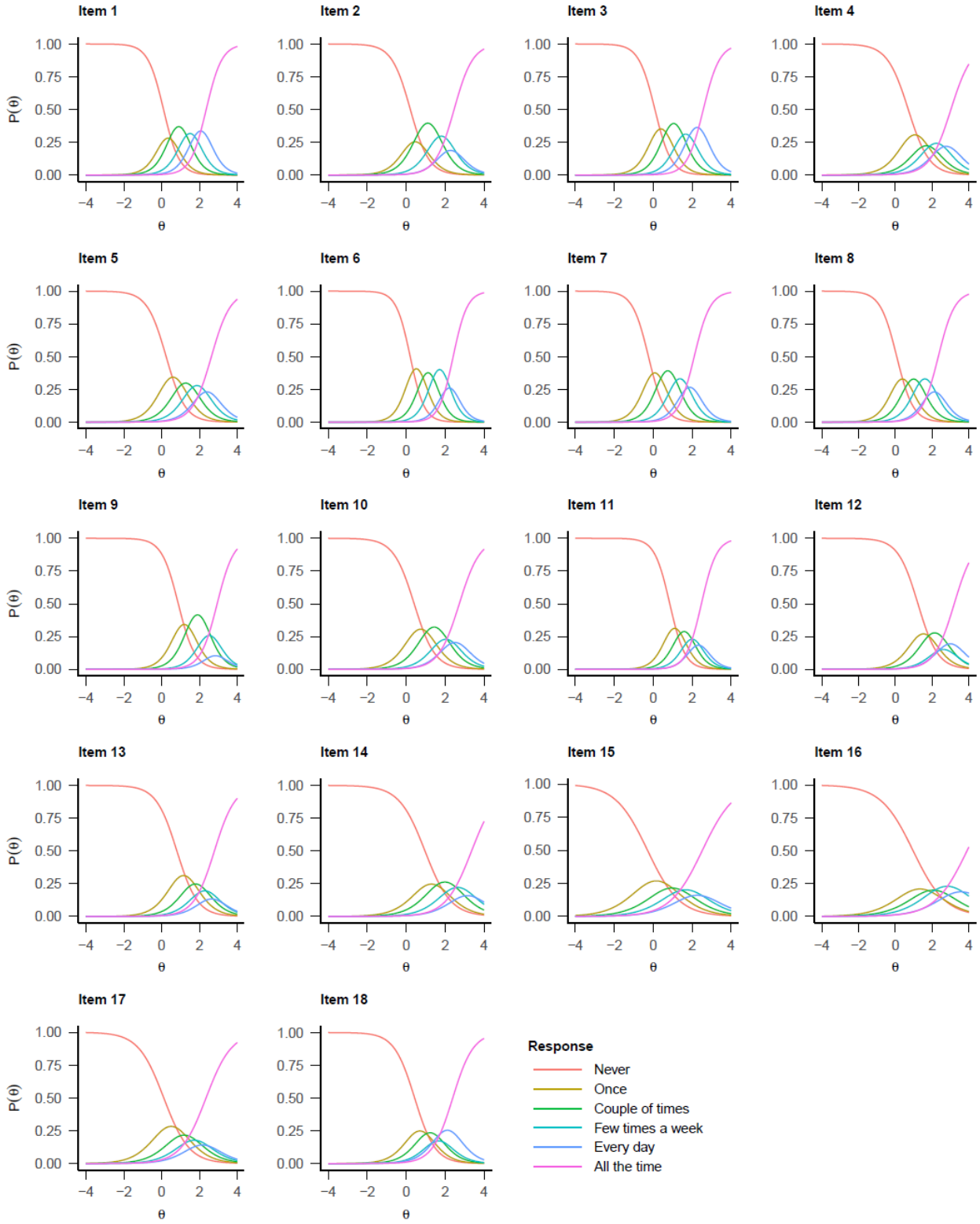

Figure S1. Category response curves (CRCs) for all 18 items. The lines represent the probability (y axis) of responding to each Likert scale option (0-5) across the distribution of theta (x axis) for each item.

4. Item information functions

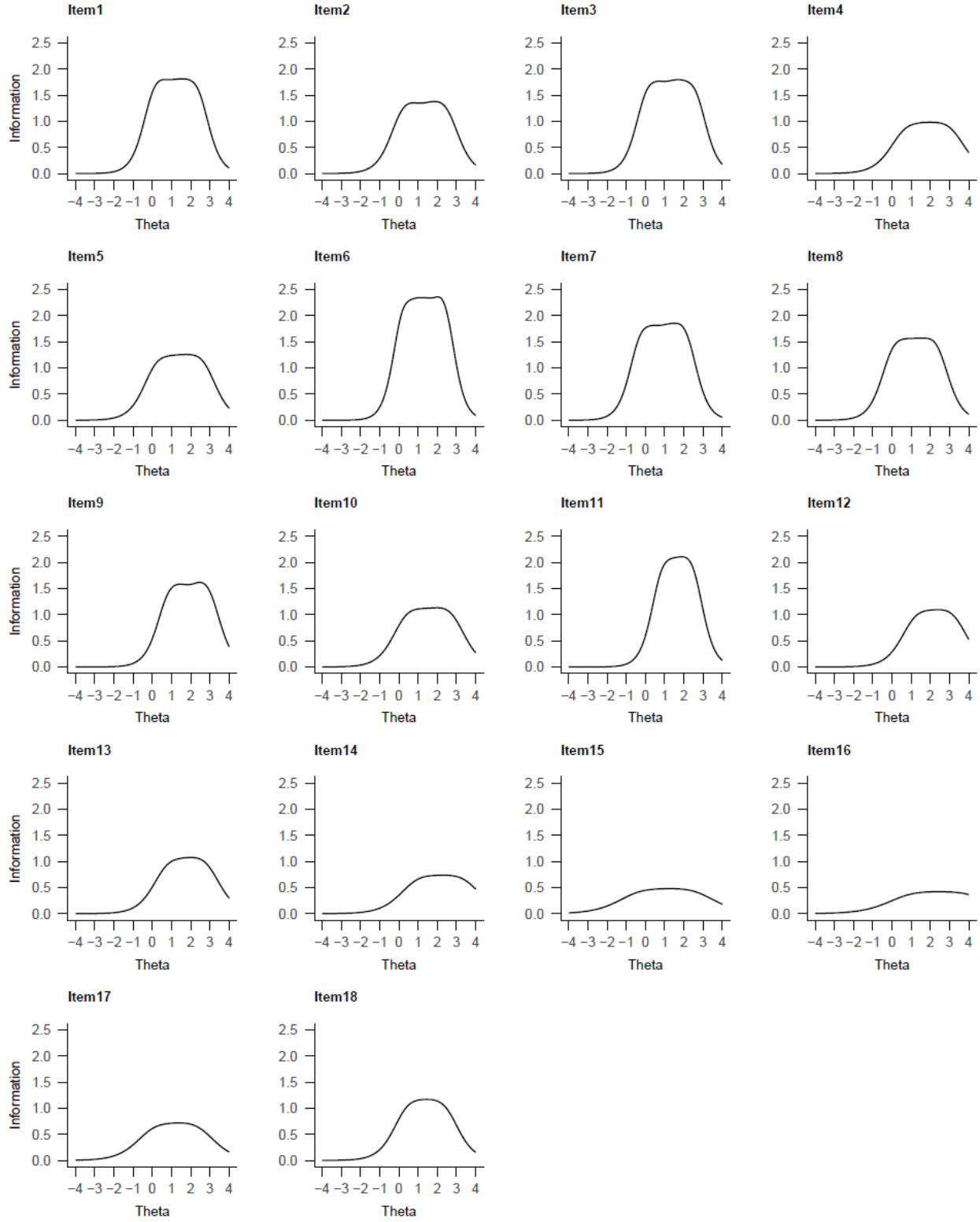

Figure S2. Item information curves for all 18 items.

5. Differential Item Functioning (DIF) analysis

5.1. Density plots

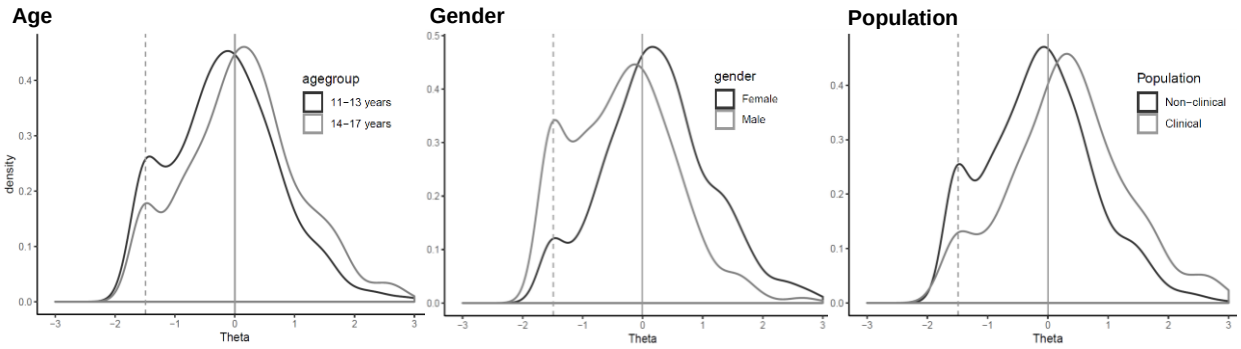

Figure S3. Density plots comparing distribution of theta (θ) between groups for Age (11-13 years, *n*=508; 14-17 years, *n*=583), Gender (female *n*=591; male *n*=493); and Population group (non-clinical *n*=795; clinical *n*=296). Solid line at theta = 0 represents average trait paranoia. Dashed line at theta = -1.49 represents an expected B-CAP score of 0.

5.2. Gender DIF analysis: Test characteristic curve

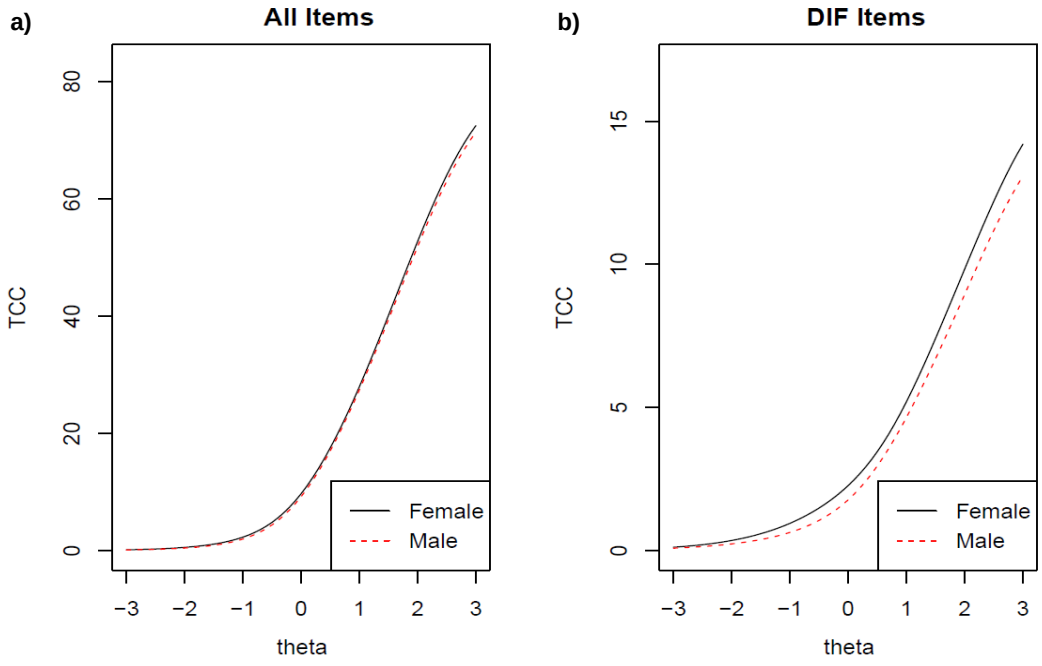

Figure S4. Impact of DIF items on the test characteristic curves (TCC). **a)** TCCs of all items (both items with and without DIF) for males and females. **b)** TCCs for the subset of items found to have DIF between males and females. At the overall test level, these curves suggest a minimal difference in the total expected score across the paranoia (theta) spectrum between males and females.

5.3. Gender DIF analysis: Item 9 curves

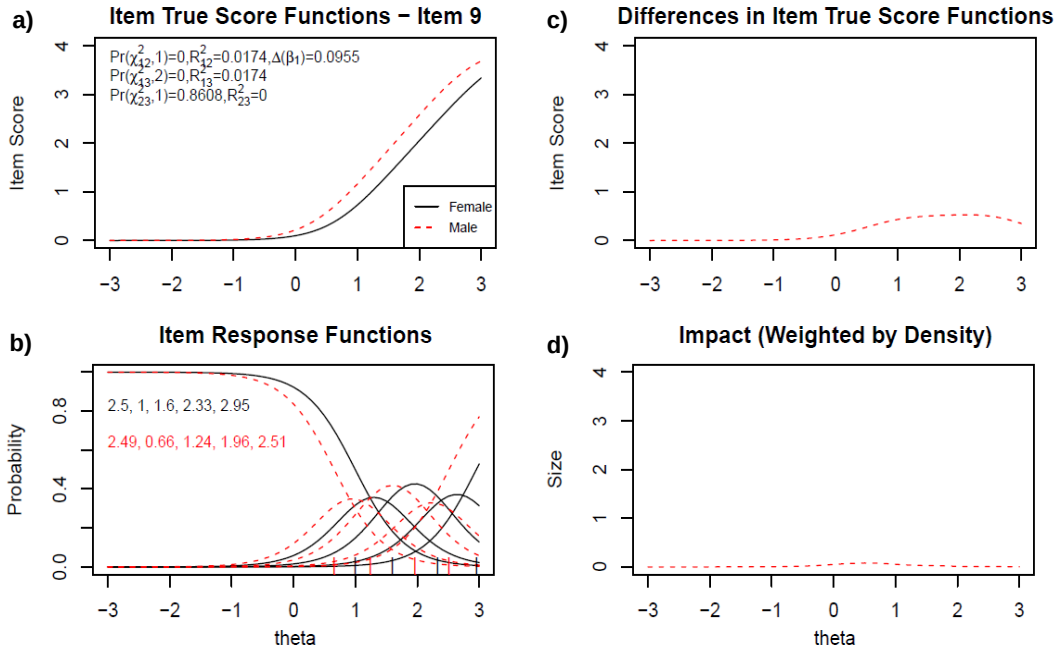

Figure S5. Graphical display of the item *'Nasty tricks are being played on me'* showing a uniform DIF relating to gender. **a)** Item characteristic curves (ICCs) for both genders **b)** Item response functions based on gender-specific item parameter estimates. **c)** Absolute difference between the ICCs of both groups, indicating the difference is at higher levels of paranoia (theta). **d)** Absolute difference between the ICCs weighted by the score distribution of the focal group (i.e. males), indicating minimal impact.

5.4. Gender DIF analysis: Item 13 curves

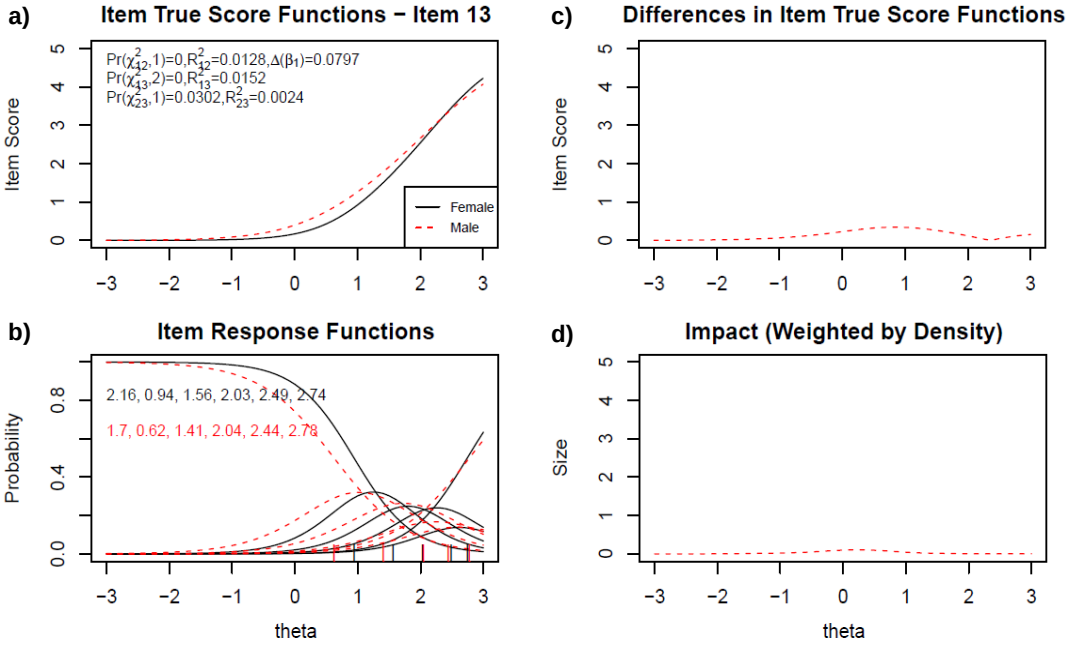

Figure S6. Graphical display of the item *'I'm sure people are seeking revenge on me'* which shows a uniform DIF with respect to gender. **a)** ICCs for both genders; **b)** Item response functions based on gender-specific item parameters; **c)** Absolute difference between the ICCs for both groups, indicating the difference is at moderate to higher levels of paranoia (theta); **d)** Absolute difference between the ICCs weighted by the score distribution for the focal group (i.e. males), indicating minimal impact.

5.5. Gender DIF analysis: Item 15 plots

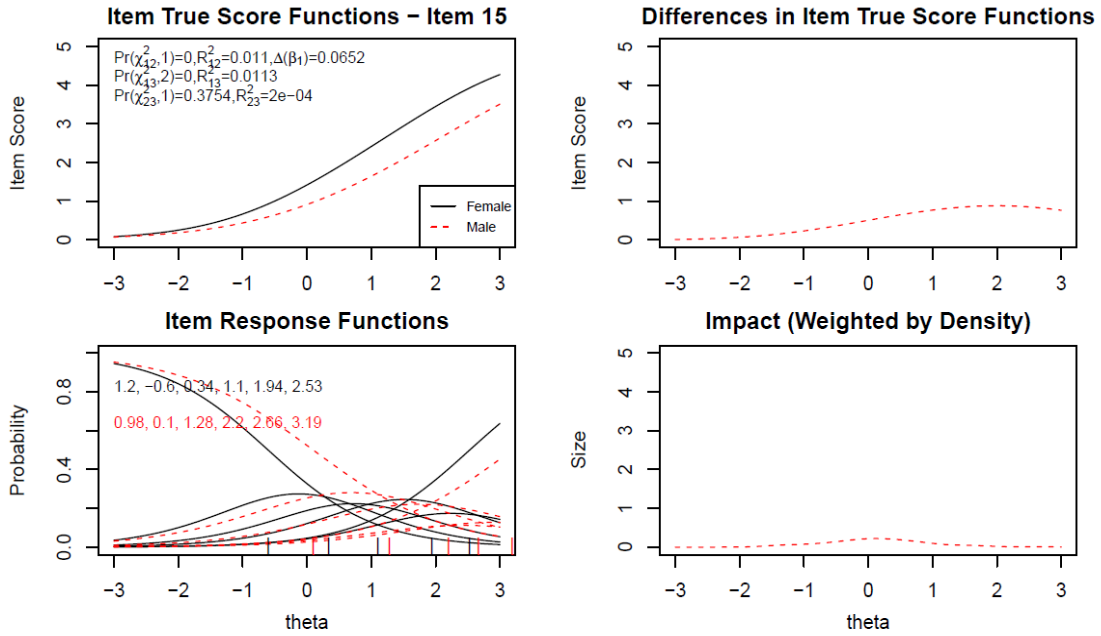

Figure S7. Graphical display of the item *'I am scared of what strangers will do to me'* which shows a uniform DIF with respect to gender. **a)** ICCs for both genders; **b)** Item response functions based on gender-specific item parameters; **c)** Absolute difference between the ICCs for both groups, indicating the difference is at moderate to higher levels of paranoia (theta); **d)** Absolute difference between the ICCs weighted by the score distribution for the focal group (i.e. males), indicating minimal impact.

5.6. Gender DIF analysis: Item 16 plots

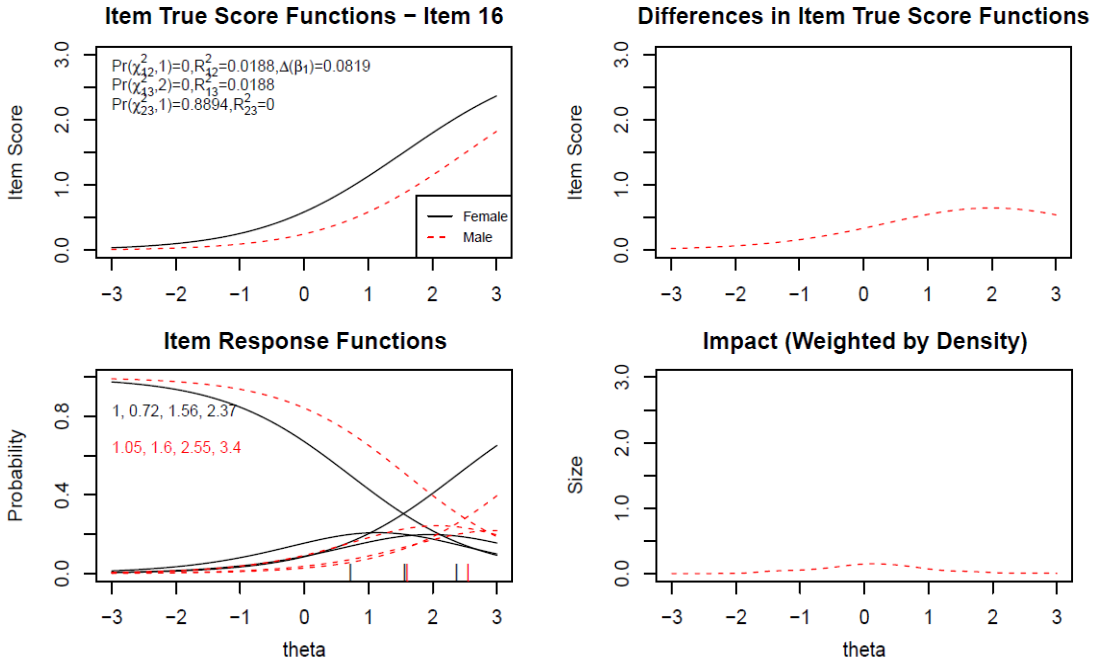

Figure S8. Graphical display of the item *'People will try to kidnap me'* which shows a uniform DIF with respect to gender. **a)** ICCs for both genders; **b)** Item response functions based on gender-specific item parameters; **c)** Absolute difference between the ICCs for both groups, indicating the difference is at moderate to higher levels of paranoia (theta); **d)** Absolute difference between the ICCs weighted by the score distribution for the focal group (i.e. males), indicating minimal impact.

5.7. Population DIF analysis: Test characteristic curves

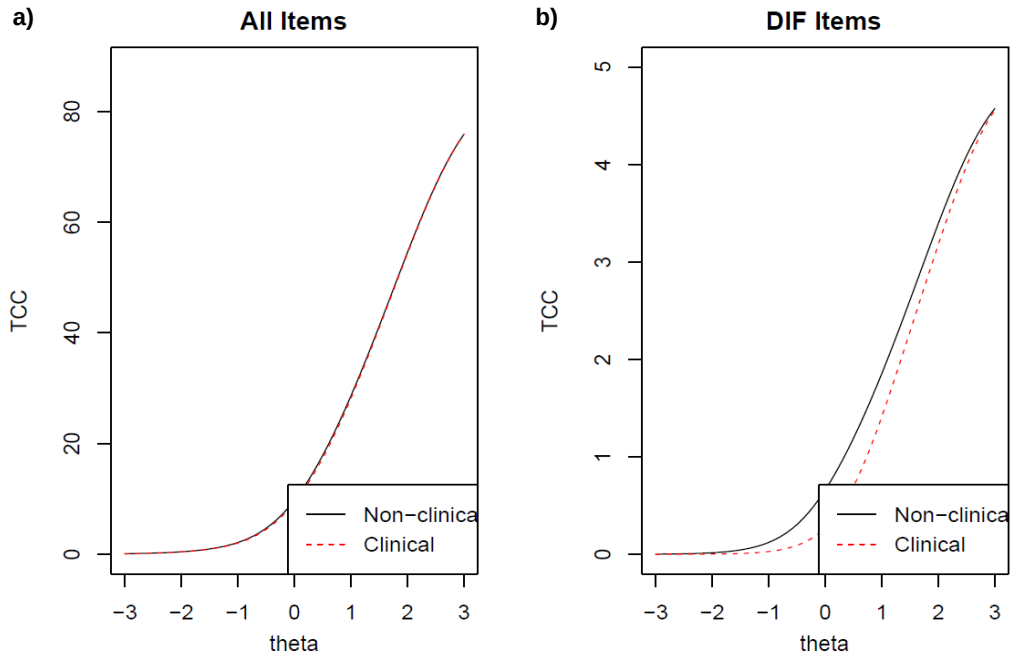

Figure S9. Impact of DIF items on the TCCs. **a)** TCCs of all items (both items with and without DIF) for the non-clinical and clinical groups; **b)** TCCs for the subset of items found to have DIF between the two population groups. At the overall test level, these curves suggest a minimal difference in the total expected score across the paranoia (theta) spectrum between the non-clinical and clinical groups.

5.8. Population DIF analysis: Item 5 plots

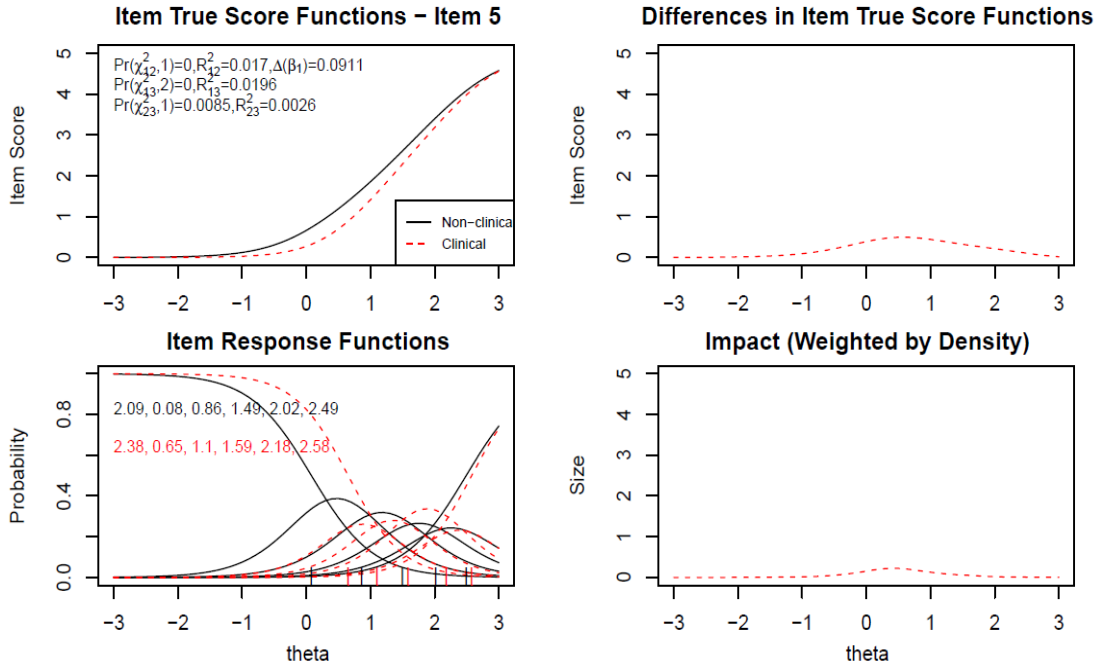

Figure S10. Graphical display of the item '*People are trying to embarrass me in class on purpose*' showing a uniform DIF with respect the population group. **a)** ICCs for both population groups; **b)** Item response functions based on population-specific item parameters; **c)** Absolute difference between the ICCs for both groups, indicating the difference is at moderate levels of paranoia (theta); **d)** Absolute difference between the ICCs weighted by the score distribution for the focal group (i.e. clinical sample), indicating minimal impact.

References

Hu, Li-tze, & Bentler, PM. (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

Rosseel, Y. (2012). Lavaan: An *R* package for structural equation modelling. *Journal of Statistical Software,* 48(2), 1-36. http://www.jstatsoft.org/v48/i02/