bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A psychometric evaluation of the 12-item EPQ-R neuroticism scale in 384,183 UK Biobank participants using item response theory (IRT)

Sarah Bauermeister[1] and John Gallacher[1]

[1]Department of Psychiatry and on behalf of Dementias Platform UK, Warneford Hospital, University of Oxford, Oxford, OX3 7JX

Sarah Bauermeister (corresponding author) sarah.bauermeister@psych.ox.ac.uk

John Gallacher john.gallacher@psych.ox.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background

Neuroticism has been described as a broad and pervasive personality dimension or 'heterogeneous' trait measuring components of mood instability; worry; anxiety; irritability; moodiness; self-consciousness; sadness and irritabililty. Consistent with depression and anxiety-related disorders, increased neuroticism places an individual vulnerable for other unipolar and bipolar mood disorders. However, the measurement of neuroticism through a self-report scale remains a challenge. Our aim was to identify psychometrically efficient items and inform the inclusion of redundant items across the 12-item EPQ-R Neuroticism scale (S. B. Eysenck, Eysenck, & Barrett, 1985) using Item Response Theory (IRT).

Methods

The 12-item binary EPQ-R Neuroticism scale was evaluated by estimating a two-parameter (2-PL) IRT model on data from 384,183 UK Biobank participants aged 39 to 73 years. Post-estimation mathematical assumptions were computed and all analyses were processed in STATA SE 15.1 (StataCorp, 2018) on the Dementias Platform UK (DPUK) Data Portal (Bauermeister et al., Preprint).

Results

A plot of θ values (Item Information functions) showed that most items clustered around the mid-range where discrimination values ranged from 1.34 to 2.27. Difficulty values for

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

individual item θ scores ranged from -0.14 to 1.25. A Mokken analysis suggested a weak to medium level of monotonicity between the items, no items reach strong scalability (H=0.35-0.47). Systematic item deletions and rescaling found that an 8-item scale is more efficient and reliable with information ranging from 1.43 to 2.36 and strong scalability (H=0.43-0.53). A 3-item scale is highly discriminatory but offers a narrow range of person ability (difficulty). A logistic regression differential item function (DIF) analysis exposed significant gender item bias functioning uniformly across both all versions of the scale.

Conclusions

Across 384,183 UK Biobank participants the 12-item EPQ-R neuroticism scale exhibited psychometric inefficiency with poor discrimination at the extremes of the scale-range. High and low scores are relatively poorly represented and uninformative suggesting that high neuroticism scores derived from the EPQ-R are a function of cumulative mid-range values. The scale also shows evidence of gender item bias and future scale development should consider the former and, selective item deletions and validation of new items to increase scale informativeness and reliability.

Keywords

Item Response Theory; IRT; neuroticism; psychometric; EPQ-R; UK Biobank; epidemiology

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Neuroticism has been described as a broad and pervasive personality dimension which influences far beyond its own limited definition (Costa & McCrae, 1987). Operationally, it has been defined as a personality trait assessed by items referencing to instances of worry; anxiety; irritability; moodiness; self-consciousness; sadness and irritabililty (Costa & McCrae, 1980, 1992; Lahey, 2009). The NEO-PI (Neuroticism-Extraversion-Openess Personality Inventory) operationalises neuroticism as a combination of individual behavioural traits which may also be measured as isolated components of mood state e.g., anxiety; hostility; depression; self-consiousness; impulsiveness and vulnerability (Costa & McCrae, 1987). Also defined as a 'heterogeneous' trait with significant overlap with depression and anxiety, neuroticism places an individual vulnerable for other unipolar and bipolar mood disorders (Lahey, 2009). Moreover, increased levels of neuroticism places an individual vulnerable to other neurotic disorders, psychological distress and 'emotional instability' (Birley et al., 2006). There is also consistent research suggesting a positive relationship between neuroticism and negative effect (Rusting, 1998) notwithstanding neurotism is essentially a dimension of negative effect (Watson & Clark, 1984). Eysenck has further argued that neuroticism is a direct reaction to the autonomic nervous system (H. J. Eysenck, 1967, 1994), findings supported where increased neuroticism was correlated with tolerance to a highly stressed environment, suggesting a habituation relationship with everyday stressors (Farrington & Jolliffe, 2001; LeBlanc, Ducharme, & Thompson, 2004).

Eysenck's attempts to define neuroticism and evaluate the measurement items persisted and an original version of the Eysenck neuroticism scale became a component of the

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Maudsley Medical Questionnaire (Faulwasser & Kittlaus, 1973). Assessment outcomes of this scale were reported in the Manual for the Maudsley Personality Inventory (MPI) where gender differences were found across the psychiatric patients and soldiers, on whom the data were derived (Francis, 1993). Later versions of the MPI were revised to remove gender-specific items although to our knowledge, details of their removal are not available. The revised neuroticism scale became a component of the Eysenck Personality Questionnaire (EPQ-R: S. B. Eysenck et al., 1985) and thereby exists as a culmination of attempts to select the relevant items through multiple revisions of the MPI. Although the EPQ-R neuroticism scale is reported to have been developed through clinical judgement and, multiple cluster and factor analyses, it is suggested that reasons for acceptance or rejection of items were complex, unclear and not 'objectified' (H. J. Eysenck & Eysenck, 1976; Francis, 1993).

The aforementioned process by which items in the EPQ-R neuroticism scale were chosen, known as classical test theory (CTT), whilst widely used, has a bias towards identifying closely associated items as being informative and is opaque to the individual item contribution or person ability. Indeed, it is suggested that the EPQ-R neuroticism scale lacks items to identify respondents who would normally endorse items at the extreme ends of the trait continuum, e.g. high vs. low neuroticism (Birley et al., 2006). Furthermore, the scale has shown to maintain gender-specific items, females consistently scoring higher (Allsop, Eysenck, & Eysenck, 1991; S. B. Eysenck et al., 1985), a difference which has been reported cross-culturally (H. J. Eysenck & Eysenck, 1982) and across the age range (e.g., S. B. Eysenck & Abdel-Khalek, 1989).

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We investigated the psychometric efficiency of the 12-item EPQ-R neuroticism scale - hereafter 'EPQ-R' (S. B. Eysenck et al., 1985) as a widely used measurement of neuroticism. We applied item response theory (IRT) to psychometrically evaluate the EPQ-R using data from UK Biobank (Sudlow et al., 2015), a large population study which assessed neuroticism at baseline. Our expectation was that the large sample size would provide valuable item-level information for assessing the informativeness of individual items and overall psychometric reliability of the scale which may have important implications in clinical settings and for epidemiological research.

Methods

Participants

The UK Biobank is a large population-based prospective cohort study of 502,664 participants. Invitations to participate in the UK Biobank study were sent to 9.2 million community-dwelling persons in the UK who were registered with the UK National Health Service (NHS) aged between 39 and 73 years. A total of 502,655 respondents elected to participate, a response rate of 5.5%. Ethical approval was granted to Biobank from the Research Ethics Committee - REC reference 11/NW/0382 (Sudlow et al., 2015).

Procedure

Assessments took place at 22 centres across the UK where participants completed an informed consent and undertook comprehensive mental health, cognitive, lifestyle, biomedical and physical assessments. The selection of mental health assessments were completed on a touchscreen computer, including the EPQ-R where participants were

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

required to answer, 'yes', 'no', 'I don't know' or 'I do not wish to answer' in response to the 12 questions: 'Does your mood often go up and down?'; 'Do you ever feel just miserable for no reason?'; 'Are you an irritable person?'; 'Are your feelings easily hurt?'; 'Do you often feel fed-up?'; 'Would you call yourself a nervous person?'; 'Are you a worrier?'; 'Would you call yourself tense or highly strung?'; 'Do you worry too long after an embarrassing experience?'; 'Do you suffer from nerves?'; 'Do you often feel lonely?'; 'Are you often troubled by feelings of guilt?'

IRT model

For these binary response data a 2 parameter logistic (2-PL) IRT model was appropriate:

$$P\left(X_i = 1 \middle| \theta, \beta_i, \alpha_i\right) = \frac{\exp(\alpha_i\,(\theta - \beta_i\,))}{1 + \exp(\alpha_i\,(\theta - \beta_i\,))}$$

The dependent variable is the dichotomous response (yes/no), the independent variables are the person's trait level, theta (θ) and item difficulty (β_i). The independent variables combine accumulatively and the item's difficulty is subtracted from θ. That is, the ratio of the probability of success for a person on an item to the probability of failure, where a logistic function provides the probability that solving any item (i) is independent from the outcome of any other item, controlling for person parameters (θ), and item parameters. The 2-PL model includes two parameters to represent the item properties (difficulty and discrimination) in the exponential form of the logistic model.

For each item, an item response function (IRF) may be calculated which calibrates the responses of an individual against each item. A calibrated standardised score for trait

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

severity θ is returned and may be plotted as an item characteristic curve (ICC) along a standardised scale with a mean of 0 (Figure 1). From the ICC two parameters may be estimated. The first is the value of θ at which the likelihood of item endorsement is 0.5, interpreted as 'expressed trait severity'. The second is the slope of the curve from the point at which the likelihood of item endorsement is 0.5, interpreted as 'expressed item discrimination' i.e., the ability to discriminate between greater and lesser severity scores. The IRF may also be expressed as an item information curve (IIF) which displays the relationship between severity and discrimination (Figure 2). The apex of the curve for any IIC indicates the value of θ at which there is maximum discrimination. By convention, scales expressing a range of θ values are more informative than those with items clustering around a single value and items with a discrimination of score of >1.7 are considered informative, although lower values are considered contributory within context (Baker, 2001). Statistical assumptions underlying the IRT principles of scalability, unidimensionality and item-independence are examined. UK Biobank data for this analysis (application 15008) were uploaded onto the Dementias Platform UK (DPUK) Data Portal (Bauermeister et al., Preprint) and analysed using STATA SE 15.1 (StataCorp, 2018).

Results

Sample

Of the 502,655 participants, 502,591 provided neuroticism scores at baseline. 48 requested their records to be withdrawn; a further 22, 608 reported a present or past neurological condition and were excluded. Participants with missing data points totalling

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

95,752 were excluded and the number of participants included in these analyses were 384,183 (207,320 female), aged 39-73 years (*M*=56.32 years; *SD*=8.07 years).

IRT analysis

A 2-PL IRT model was estimated whereby difficulty and discrimination parameters were extracted (Table 1). The discrimination (item-information) parameters across the scale range between 1.34 and 2.27, the item measuring 'Does your mood often go up and down?' exhibits the highest level of discrimination at 2.27, suggesting that this 'mood' question possesses the highest amount of information synonymous with the neurotic trait. In contrast, the item 'Are you an irritable person?', 1.34, is the lowest, and below the suggested recommended level of 1.7 for an ideal discrimination level for items measuring trait values (Baker, 2001). The items, 'Are you a worrier?'; 'Do you suffer from nerves'; 'Do you ever feel just miserable for no reason?'; 'Do you often feel fed-up?' and 'Would you call yourself tense or highly strung' also have discrimination values of above 1.7.

The difficulty parameter functions as a probability scale with the item position on θ indicating the probability value of a respondent endorsing an item. Figure 3 shows the item characteristic curves (ICCs) for each of the items, presenting both the steepness of the discrimination curve and position of the difficulty value on the θ continuum. For example, for the item 'Does your mood often go up and down?', there is a 50% probability that someone with a θ of 0.22 (someone who does experience neurotic trait characteristics) would endorse this item, therefore it is considered an item characteristic of neuroticism, albeit low. On contrary, for the item ''Are you a worrier?", there is a 50%

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

chance of someone with a Θ of -0.13 endorsing this item, therefore, someone who does not experience neurotic trait charateristics.

Additional item discrimination is available by graphing the IIF curves (see Figure 4). The IIF curves thereby display the relationship between difficulty and discrimination, and an important feature of this graph is also the position on the continuum from which the point is drawn perpendicular from the apex of each item curve. The items which have their maximum curvature positioned along the Θ continuum in the positive half provide information about the neurotic trait when there is an endorsement (presence) of the trait characteristic. For example, the item 'Do you often feel lonely?' is an endorsement of neuroticism if a respondent endorses it, as its apex is positioned in positive Θ and is more likely to be endorsed by someone with a higher difficulty level of neuroticism (1.42) than a person endorsing the item 'Does your mood often go up and down?' which is also positioned in the positive Θ but has a lower difficulty value (0.22). Therefore, although the 'mood' item has the highest discrimination value (see previous), it does not provide sufficient information about respondents who possess a high level (presence) of the trait (+1 or +2) or a low level (absence) of the trait (-1 or -2), instead it provides the most information for respondents who possesses an average (Θ=0) to a minimal amount of the neuroticism trait. A further item for which the apex is also is placed just beyond the average trait Θ is 'Do you worry too long after an embarrassing experience?' (0.16), suggesting that respondents with just an above average amount of neuroticism might endorse this item but the item does not actually possess high level of information about the trait, and could be endorsed by someone with just an average amount, or no neurotic traits. Furthermore, the discrimination value of this item is also very low (see Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The item which possesses the least trait characteristic information is the item, 'Are you an irritable person?', Although the IIF curve apex is positioned over a positive ϴ (0.96), and may be endorsed by a respondent possessing an amount of the trait characteristic, the discrimination value is low (1.34).

In summary, the overall pattern of item distribution across the ϴ continuum suggests that across the 12-item EPQ-R neuroticism scale there are no items which measure an extreme level of neurotic trait characteristics or an extreme level of non-neurotic trait characteristics. It also suggests that the questions are mostly measuring the neurotic trait characteristics which have a higher probability of endorsement by individuals who are experiencing a minimal level of neuroticism rather than an average amount or none.

Reliability

In IRT, the information from the IIF for each item may be combined into a test information function (TIF) which provides an overall indication of how reliable the overall scale performs across all variables (Figure 5). Reliability is thereby calculated at multiple point values of ϴ along the continuum. The TIF graph suggests that there is reliable information to differentiate respondents who possess an average to just above average amount of neurotic traits however, there is little reliable information to differentiate the absence of neurotic trait characteristics. At ϴ=-1, there is virtually no reliable information that can be obtained from the scale items and likewise, at ϴ=2. Reliability of an IRT scale may be defined at different points of ϴ with the mean of ϴ fixed at 0 and the variance at 1, facilitating identification of the model and reliability for all points along the ϴ continuum,

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distinguishing respondents according to specific values of ϴ (Thissen, 2000). Here, our TIF suggests that there is a lot of reliable information to differentiate respondents who possess just above an average amount of trait information (ϴ=1) however, the range is narrow beyond this value (see Table 2). The reliability figures suggest that at the neutral position of ϴ=0, the reliability of the 2-PL IRT is good at 0.87 (Kline, 2005). However, further along the continuum towards positive ϴ (greater neuroticism), reliability increases to ϴ=1 (0.88), then decreases ϴ=2 (0.76), ϴ=3 (0.44) and ϴ=4 (0.14) suggesting that the highest reliability of measuring the neurotic trait is at normal or a minimal amount of neuroticism, ϴ=0 or 1. Thereafter, reliability reduces so that the extreme end of the continuum, ϴ=3 or 4, is no longer reliably measured. The figures for negative ϴ suggest lower reliability of the scale to measure absence of the trait with ϴ=-1 (0.71) and all remaining reliability measures of ϴ on the negative side of the continuum are below acceptable reliability.

Statistical assumptions

1. Item independence

A correlation analysis assessed initial item independency and all items were significantly correlated (p <.000) but the majority of values were lower than 0.50, suggesting basic local item independence. A residual coefficient matrix was computed after a single-factor model was estimated, the outcome showed that no residuals were too highly correlated, i.e., R >0.20, (Yen, 1993) and all were within acceptable limits, suggesting item independence.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. Monotonicity

A Mokken analysis produced a Loevinger H coefficient (Sijtsma & Molenaar, 2002) measuring the scalable quality of items expressed as a probability measure, independent of a respondent's Θ. These coefficients ranged between 0.35 and 0.47 (Table 3), suggesting a weak (H=0.3-0.4) to moderate (H=0.4-0.5) monotonicity, no items reached strong scalability (H≥0.5) (Sijtsma & Molenaar, 2002).

3. Unidimensionality

A principal component analysis (PCA) suggested that a single major factor is responsible for 54% of the variance and a second factor responsible for 39% of the variance, above the suggested 20% proportion indicating a single major factor is being measured (Reeve et al., 2007). A post-IRT estimation model measure of unidimensionality was also computed using a semi-partial correlation controlling for Θ. This analysis provides individual item variance contribution after adjusting for all the other variables including Θ. It demonstrates the relationship between local independence and unidimensionality, reflecting a conservative assessment whereby the desired R^2 should ideally be zero or as close to zero as possible (De Mars, 2010). Most items were 0.01 and the remaining were 0.02, suggesting questionable unidimensionality. To our knowledge, there is still no standardised cut-off criterium for assessing this value (i.e., how close to zero all items should be across a scale).

IRT revised analysis

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To assess a revised scale, items were systematically removed from the scale according to discrimination value with the lowest discrimating item removed first ('Are you an irritable person?', 1.34) before the IRT 2-PL model was re-estimated with the remaining items and the process repeated, removing the lowest discrimating item, below 1.7. In order of removal, the items systematically removed thereafter were: 'Do you often feel lonely?'; 'Are you often troubled by feelings of guilt?'; 'Do you worry too long after an embarrassing experience?'.

Statistical assumptions were computed on the revised scale of 8 items (Table 4) and importantly a Mokken analysis suggests improved scalability (monotonicity) compared to the full 12-item scale with two items reaching values >0.50 (Table 5). Reliability across the scale is marginally improved compared to the full scale suggesting redundancy of the removed items (Table 6). Acceptable metrics for unidimensionality and item independence were achieved for this revised scale. The ICC and IIF for the revised 8-item scale are presented in Figures 6 and 7 where improved item information is presented.

A further item reduction was explored to investigate a minimal scale for ascertaining high positive discrimination of latent trait, rather than item balance across the scale. After systematic item-removal, three items remained which possessed high discrimination and positive difficulty values, 'Does your mood often go up and down?' (3.38; 0.20); 'Do you ever feel just miserable for no reason?' (2.76; 0.26) and 'Do you often feel fed-up?' (2.89; 0.34) (Table 7). A Mokken analysis suggests that scalability is moderate to good (H≥4-5) (Table 8), a semi-partial correlation analysis controlling for ϴ showed all values <0.20 and residual coefficient correlations were all 0.00. Reliability was almost comparable to

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the 8-item scale (Table 9). The ICC and IIF graphs suggest the three-item scale may present an efficient, alternative and highly informative scale, however, the scale is too narrow in range for detecting presence of neurotic trait characteristics beyond average θ (Figures 8 and 9).

Differential-Item Functioning (DIF) Analysis

To investigate gender differences in item functioning, a logistic DIF analysis was conducted across all three versions of the scale with gender as the observed group. A uniform and nonuniform DIF assessed whether specific items favoured one group (male vs. female) over the other for all values of the latent trait (uniform) or just selected values of the latent trait (nonuniform). The output of these analyses are presented in Table 10 where evidence of significant uniform DIF for gender was found across all three versions, suggesting that the scale showed evidence of gender DIF across the items.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In a large population cohort of 384,183 adults aged 39-73 years, limitations in the range and reliability of item trait characteristics were found across the EPQ-R when a 2PL IRT model was estimated. Our findings suggest that the EPQ-R is inefficient with poor discrimination at the extreme ends of the scale-range, such that high and low scores are relatively poorly represented and uninformative. A reliability plot overlaid by the standard error of measurement also suggests poor reliability at the extremes of the scale score and that high neuroticism scores derived from the EPQ-R are a function of accumulative mid-range values. In a revised 8-item version of the scale, greater item-discrimination and reliability was found across the scale suggesting that selected items within the 12-item version are redundant. A 3-item version was explored but although this scale possessed items of high discrimination and scalability, range was narrow and lacked reliability beyond normal trait values. A DIF analysis with gender as a group outcome suggests the scale exhibits significant gender differential item functionting across all versions of the scale.

To our knowledge, this is the first study to conduct a comprehensive psychometric scale assessment applying IRT to the EPQ-R on such a large population. Furthermore, although the assumption values and parameter output of the 12-item IRT calibration were mostly acceptable according to established psychometric standards, an examination of individual items suggests that there were items of low discrimination and the scale could benefit from revisions based on psychometric methodologies such as those used here, and as evidenced in the scale-revision analysis.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

It is of fundamental importance that health measurement scales are reliable and valid measures of the construct of interest. Utilising psychometric methodologies to analyse psychosocial and health related outcomes has important implications for assessing longitudinal change both in clinical settings and epidemiological research. An IRT analysis provides item-level information and scale characteristics through the computation of post-estimation assumptions, and estimation of an individual Ѳ latent metric which predicts individual Ѳ scores on the fitted IRT model. This Ѳ metric may then be used as a latent construct in assessing longitudinal change (Acock, 2016) which may be a more reliable measure compared to a single summated score (Lu, 2005). Furthermore, it is also suggested that using Ѳ in longitudinal studies, over the summated score, may be preferable reducing overestimation of the repeated measure variance and underestimation of the between-person variance which is avoided if an IRT model is implemented (Gorter, Fox, & Twisk, 2015).

A further advantage of utilising psychometric methodologies in an epidemiological context is that IRT extends the opportunity to utilise, computer adaptive testing (CAT) for both scale development and for efficient test delivery. During CAT administration, Ѳ is automatically computed in response to the trait (Ѳ) of the respondent and it is therefore not necessary to present the full range of items as the response scale is adaptive to the individual, the items underlying the trait and a stopping rule (Wainer et al., 2014). The potential to reduce a scale so that only the most reliable and informative questions are presented to participants is essential in clinical settings and for epidemiological research. This is important for individuals who are older or who have co comoborbid

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mental health or mood disorders. Moreover, focused, reliable and user-friendly scales in a research setting increases user satisfaction, reduces participant burden and maintains long-term participant retention.

Participants who display or possess the extreme trait characteristics are rare, however, the potential should exist for this eventuality, but many scales are simply not adequately designed to do so (Acock, 2016). Moreover, previous research suggests that both the 12 and 23-item EPQ-R neuroticism scales may have reduced power to discriminate between low and high scoring individuals (Birley et al., 2006), we found evidence of this in the 12-item scale. It is important in both clinical and research settings that scales are designed to measure across the trait spectrum and this is possible if scales are developed using psychometric methodologies such as those described here and elsewhere (e.g., de Ayala, 2009; Streiner, Norman, & Cairney, 2015).

Conclusions

The 12-item neuroticism EPQ-R scale lacked item reliability and neurotic trait-specific information at the extreme ends of the neurotic continuum when a 2-PL IRT model was estimated. In a secondary analysis, a systematic item-elimination and 2-PL model re-estimation procedure was followed and it was found that an 8-item scale possessed items with higher levels of item information and reliability. This study suggests that the 12-item EPQ-R scale could benefit from item revisions and updating with both existing item deletions and validation of replacement items which consider gender item bias.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Strengths of this study were the large population cohort available for a comprehensive IRT analysis and the psychometric methodologies which were applied to the data.

Acknowledgements

This is a DPUK supported project with all analyses conducted on the DPUK Data Portal, constituting part 1 of DPUK Application 0169.

The Medical Research Council supports DPUK through grant MR/L023784/2

Authors' Contributions

SB and JG conceptualised the idea. SB analysed and interpreted the data, and wrote the manuscript. JG edited and proofread the manuscript. Both authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Acock, A. C. (2016). *A Gentle Introduction to Stata* (5th ed.). Texax, USA: A Stata Press Publication.

Allsop, J., Eysenck, H. J., & Eysenck, S. B. G. (1991). Machiavellianism as a component in psychoticism and extraversion. *Personality and Individual Differences, 12*, 29-41.

Baker, F. B. (2001). *The basics of item response theory*. Orignal work published in 1985 http://echo.edres.org:8080/irt/baker/final.pdf: College Park, DM: ERIC Clearinghouse on Assessment and Evaluation.

Bauermeister, S., Orton, C., Thompson, S., Barker, R. A., Bauermeister, J. R., Ben-Shlomo, Y., . . . Gallacher, J. E. (Preprint). Data Resource Profile: The Dementias Platform UK (DPUK) Data Portal. *BioRxiv*. doi:10.1101/582155

Birley, A. J., Gillespie, N. A., Heath, A. C., Sullivan, P. F., Boomsma, D. I., & Martin, N. G. (2006). Heritability and nineteen-year stability of long and short EPQ-R neuroticism scales. *Personality and Individual Differences, 40*(4), 737-747. doi:10.1016/j.paid.2005.09.005

Costa, P. T., Jr., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *J Pers Soc Psychol, 38*(4), 668-678.

Costa, P. T., Jr., & McCrae, R. R. (1987). Neuroticism, somatic complaints, and disease: is the bark worse than the bite? *J Pers, 55*(2), 299-316.

Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653-665.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. USA: The Guildford Press.

De Mars, C. (2010). *Item Response Theory*. New York, USA: Oxford University Press.

Eysenck, H. J. (1967). *The biological basis of personality*. London: Springfield, III: Charles C. Thomas.

Eysenck, H. J. (1994). *Personality; biological foundation. the neurophysiology of indivdual difference*. New York: Academic Press.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality*. London: Hodder & Stoughton.

Eysenck, H. J., & Eysenck, S. B. G. (1982). Recent advances in the cross-cultural study of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality development* (pp. 41-69). Hillsdale, USA.: Erlbaum.

Eysenck, S. B., & Abdel-Khalek, A. M. (1989). A cross-cultural study of personality: egyptian and english children. *Int J Psychol, 24*(1-5), 1-11. doi:10.1080/00207594.1989.10600028

Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences, 6*, 21-29.

Farrington, D., & Jolliffe, D. (2001). Personality and Crime. In P. B. B. N. J. Smelser (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (1st ed.). USA: Elsevier.

Faulwasser, H., & Kittlaus, H. (1973). [Economy of the Maudsley Medical Questionnaire (MMQ)]. *Psychiatr Neurol Med Psychol (Leipz), 25*(5), 276-281.

Francis, L. J. (1993). The Dual Nature of the Eysenckian Neuroticism Scales - a Question of Sex-Differences. *Personality and Individual Differences, 15*(1), 43-59. doi:Doi 10.1016/0191-8869(93)90040-A

Gorter, R., Fox, J. P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol, 15*, 55. doi:10.1186/s12874-015-0050-x

Kline, P. (2005). *The Principles and Practice of Structural Equation Modeling* (2nd ed.). USA: The Guildford Press.

Lahey, B. B. (2009). Public health significance of neuroticism. *Am Psychol, 64*(4), 241-256. doi:10.1037/a0015309

LeBlanc, J., Ducharme, M. B., & Thompson, M. (2004). Study on the correlation of the autonomic nervous system responses to a stressor of high discomfort with personality traits. *Physiol Behav, 82*(4), 647-652. doi:10.1016/j.physbeh.2004.05.014

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Lu, I. R. R. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling-a Multidisciplinary Journal, 12*(2), 263-277. doi:DOI 10.1207/s15328007sem1202_5

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Group, P. C. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care, 45*(5 Suppl 1), S22-31. doi:10.1097/01.mlr.0000250483.85507.04

Rusting, C. L. (1998). Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. *Psychol Bull, 124*(2), 165-196.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). London: Sage Publications.

StataCorp, L. (2018). Stata SE 15.1. College Station, TX, USA: StataCorp LLC. Retrieved from www.stata.com

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health Measurement Scales* (5th ed.). Great Britain: Oxford University Press.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., . . . Collins, R. (2015). UK Biobank: An Open Access Resource for Indentifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *12*(3). doi:10.1371/journal.pmed.1001779

Thissen. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized Adaptive Testing: A primer* (pp. 159-184). London: Lawrence Erlbaum: Lawrence Erlbaum.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2014). *Computerized Adaptive Testing: A Primer* (2nd ed.). New York, USA: Routledge.

Watson, D., & Clark, L. A. (1984). Negative affectivity: the disposition to experience aversive emotional states. *Psychol Bull, 96*(3), 465-490.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

bioRxiv preprint doi: https://doi.org/10.1101/741249; this version posted August 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables

Table 1. 2-PL IRT model item parameters for the 12-item scale

Item	Parameter	B	SE	Z	95% CI
Mood go up and down?	Discrimination	2.27	0.01	217.05	(2.25 2.29)
	Difficulty	0.22	0.00	86.03	(0.21 0.22)
Feelings easily hurt?	Discrimination	1.60	0.01	225.49	(1.59 1.62)
	Difficulty	-0.13	0.00	-43.10	(-0.13 -0.12)
Are you a worrier?	Discrimination	1.85	0.01	223.93	(1.83 1.87)
	Difficulty	-0.13	0.00	-47.27	(-0.14 -0.12)
Suffer from nerves?	Discrimination	1.85	0.01	205.39	(1.83 1.87)
	Difficulty	1.04	0.00	263.46	(1.03 1.05)
Feel miserable no reason?	Discrimination	1.97	0.01	223.01	(1.95 1.98)
	Difficulty	0.29	0.00	106.41	(0.29 0.30)
Often feel fed-up?	Discrimination	2.09	0.01	219.81	(2.07 2.11)
	Difficulty	0.37	0.00	135.88	(0.36 0.38)
Tense or highly strung?	Discrimination	2.04	0.01	198.01	(2.02 2.06)
	Difficulty	1.24	0.00	292.74	(1.24 1.25)
Often feel lonely?	Discrimination	1.47	0.01	193.73	(1.46 1.49)
	Difficulty	1.42	0.01	250.93	(1.41 1.44)

An irritable person?	Discrimination	1.34	0.01	206.98	(1.33 1.36)
	Difficulty	0.96	0.00	213.31	(0.95 0.97)
A nervous person?	Discrimination	1.66	0.01	203.48	(1.65 1.68)
	Difficulty	1.18	0.01	260.18	(1.17 1.19)
Worry embarrassing experience?	Discrimination	1.45	0.01	221.60	(1.44 1.47)
	Difficulty	0.15	0.00	46.84	(0.14 0.15)
Troubled feelings of guilt?	Discrimination	1.54	0.01	215.34	(1.53 1.56)
	Difficulty	0.86	0.00	220.30	(0.85 0.87)

Note: Item names truncated for brevity, see text; B=standardised beta coefficients; SE=standard error

$p < .000$ for all B values.

Table 2. Reliability for values of ϴ from a 2-PL IRT model fit for the 12-item scale

ϴ	TIF	TIF SE	Reliability
-3	1.10	0.95	0.09
-2	1.51	0.81	0.34
-1	3.42	0.54	0.71
0	7.92	0.36	0.87
1	8.03	0.35	0.88
2	4.13	0.49	0.76
3	1.78	0.75	0.44

Note: TIF=Test Information Function; SE=standard error

Table 3. Loevinger H coefficients for the 12-item scale

Item	H
Mood go up and down?	0.46
Feelings easily hurt?	0.44
Are you a worrier?	0.47
Suffer from nerves?	0.43
Feel miserable no reason?	0.43
Often feel fed-up?	0.44
Tense or highly strung?	0.47
Often feel lonely?	0.39
An irritable person?	0.35
A nervous person?	0.41
Worry embarrassing experience?	0.39
Troubled feelings of guilt?	0.39

Note: Item names truncated for brevity, see text.

Table 4. 2-PL IRT model item parameters for the 8-item scale

Item	Parameter	B	SE	Z	95% CI
Mood go up and down?	Discrimination	2.36	0.01	194.99	(2.34 2.39)
	Difficulty	0.21	0.00	84.97	(0.21 0.22)
Feelings easily hurt?	Discrimination	1.43	0.01	211.92	(1.42 1.44)
	Difficulty	-0.14	0.00	-43.10	(-0.14 -0.13)
Are you a worrier?	Discrimination	1.76	0.01	208.06	(1.74 1.76)
	Difficulty	-0.14	0.00	-47.27	(-0.14 -0.13)
Suffer from nerves?	Discrimination	1.93	0.01	188.28	(1.91 1.95)
	Difficulty	1.03	0.00	260.11	(1.01 1.03)
Feel miserable no reason?	Discrimination	2.02	0.01	205.06	(2.00 2.04)
	Difficulty	0.29	0.00	105.39	(0.28 0.29)
Often feel fed-up?	Discrimination	2.07	0.01	203.28	(2.05 2.09)
	Difficulty	0.37	0.00	134.24	(0.36 0.38)
Tense or highly strung?	Discrimination	2.03	0.01	186.64	(2.01 2.05)
	Difficulty	1.25	0.00	285.95	(1.24 1.26)
A nervous person?	Discrimination	1.72	0.01	192.42	(1.71 1.74)
	Difficulty	1.16	0.00	258.26	(1.15 1.17)

Note: Item names truncated for brevity, see text; B=standardised beta coefficients; SE=standard error

Table 5. Loevinger H coefficients for the 8-item scale

Item	H
Mood go up and down?	0.48
Feelings easily hurt?	0.43
Are you a worrier?	0.48
Suffer from nerves?	0.50
Feel miserable no reason?	0.45
Often feel fed-up?	0.46
Tense or highly strung?	0.53
A nervous person?	0.47

Note: Item names truncated for brevity, see text.

Table 6. Reliability for values of Θ from a 2-PL IRT model fit for the 8-item scale

Θ	TIF	TIF SE	Reliability
-3	1.12	0.94	0.11
-2	1.54	0.81	0.35
-1	3.09	0.57	0.68
0	6.33	0.40	0.84
1	5.78	0.42	0.83
2	2.75	0.60	0.64
3	1.32	0.87	0.24

Note: TIF=Test Information Function; SE=standard error

Table 7. 2-PL IRT model item parameters for the 3-item scale

Item	Parameter	B	SE	Z	95% CI
Mood go up and down?	Discrimination	3.38	0.03	117.28	(3.32 3.44)
	Difficulty	0.20	0.00	84.97	(0.20 0.20)
Feel miserable no reason?	Discrimination	2.76	0.02	163.35	(2.73 2.79)
	Difficulty	0.26	0.00	112.50	(0.26 0.27)
Often feel fed-up?	Discrimination	2.89	0.02	156.62	(2.85 2.92)
	Difficulty	0.34	0.00	141.52	(0.33 0.34)

Note: Item names truncated for brevity, see text; B=standardised beta coefficients; SE=standard error

$p < .000$ for all B values

Table 8. Loevinger H coefficients for the 3-item scale

Item	H
Mood go up and down?	0.57
Feel miserable no reason?	0.54
Often feel fed-up?	0.56

Note. Item names truncated for brevity, see text.

Table 9. Reliability for values of Θ from a 2-PL IRT model fit for the 3-item scale

Θ	TIF	TIF SE	Reliability
-3	1.00	0.99	0.00
-2	1.03	0.98	0.03
-1	1.58	0.80	0.37
0	6.88	0.38	0.85
1	3.38	0.54	0.70
2	1.16	0.93	0.13
3	1.01	0.99	0.01

Note: TIF=Test Information Function; SE=standard error

Table 10. Logistic regression differential item function (DIF) analysis across 12, 8 and 3-item scales

Item	12-item scale		8-item scale		3-item scale	
	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform	Uniform
Mood go up and down?	28.44***	2130.71***	88.31***	2424.25***	3.20	1350.22***
Feelings easily hurt?	69.76***	1145.94***	112.62***	5213.95***		
Are you a worrier?	89.29***	2370.92***	60.35***	2391.80***		
Suffer from nerves?	564.98***	66.20***	0.02	41.49***		
Feel miserable no reason?	35.49***	1145.94***	27.45***	1119.77***	4.25*	4265.38***
Often feel fed-up?	48.48***	1191.26***	55.20***	1262.96***	53.94***	725.93***
Tense or highly strung?	49.41***	141.80***	121.15***	155.65***		
Often feel lonely?	21.92***	168.78***				
An irritable person?	0.01	7822.22***				
A nervous person?	564.48***	3188.26***	955.25***	3670.95***		
Worry embarrassing experience?	5.94*	1289.73***				
Troubled feelings of guilt?	38.33***	1873.03***				

Note: Item names truncated for brevity, see text.

$*p < .05$; $***p < .000$

Figures

Figure 1. Item Characteristic Curve (ICC) graph

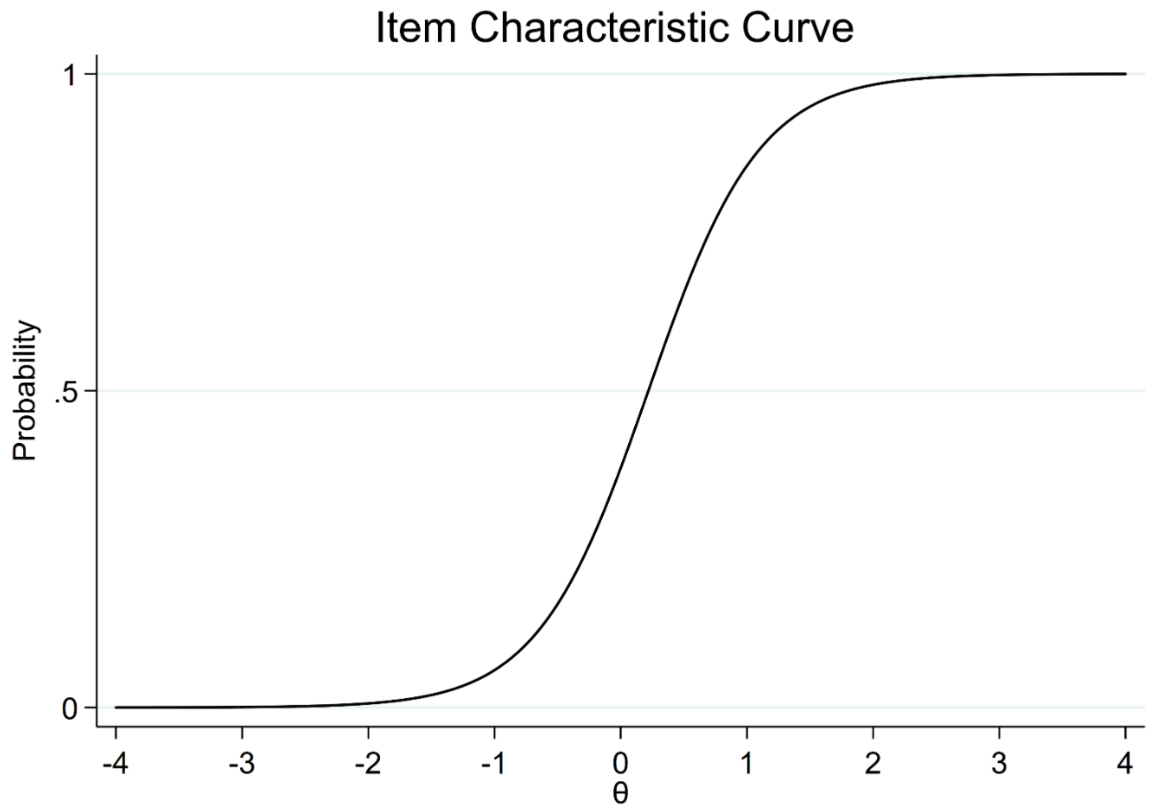

Figure 2. Item Information Function (IIF) graph

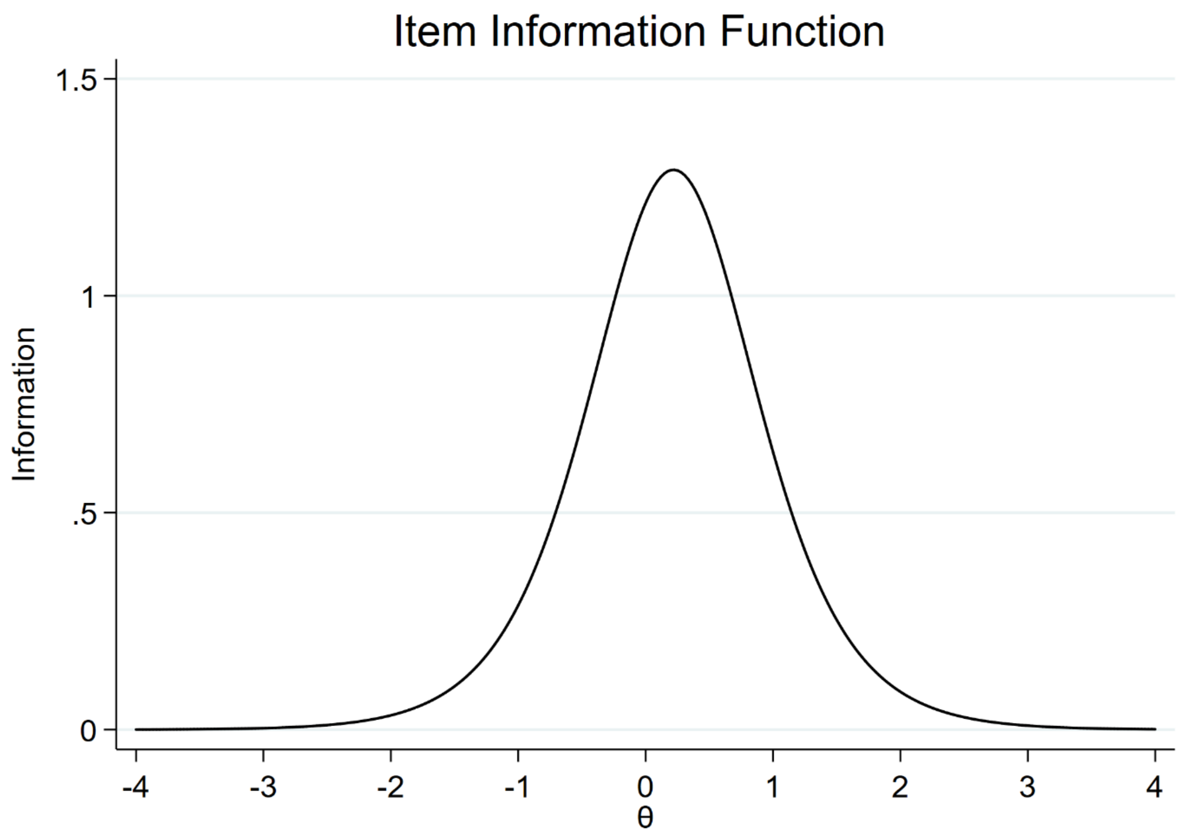

Figure 3. ICC graph for the 12-item scale

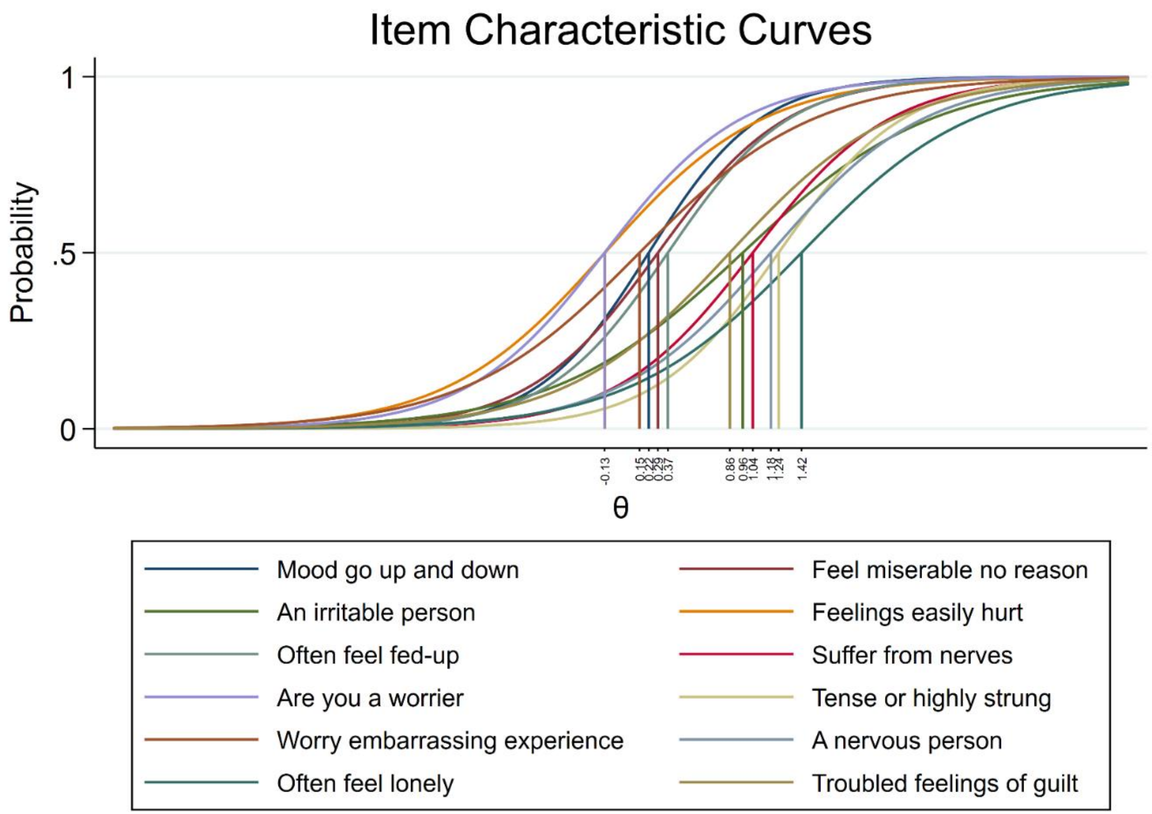

Figure 4. IIF graph for the 12-item scale

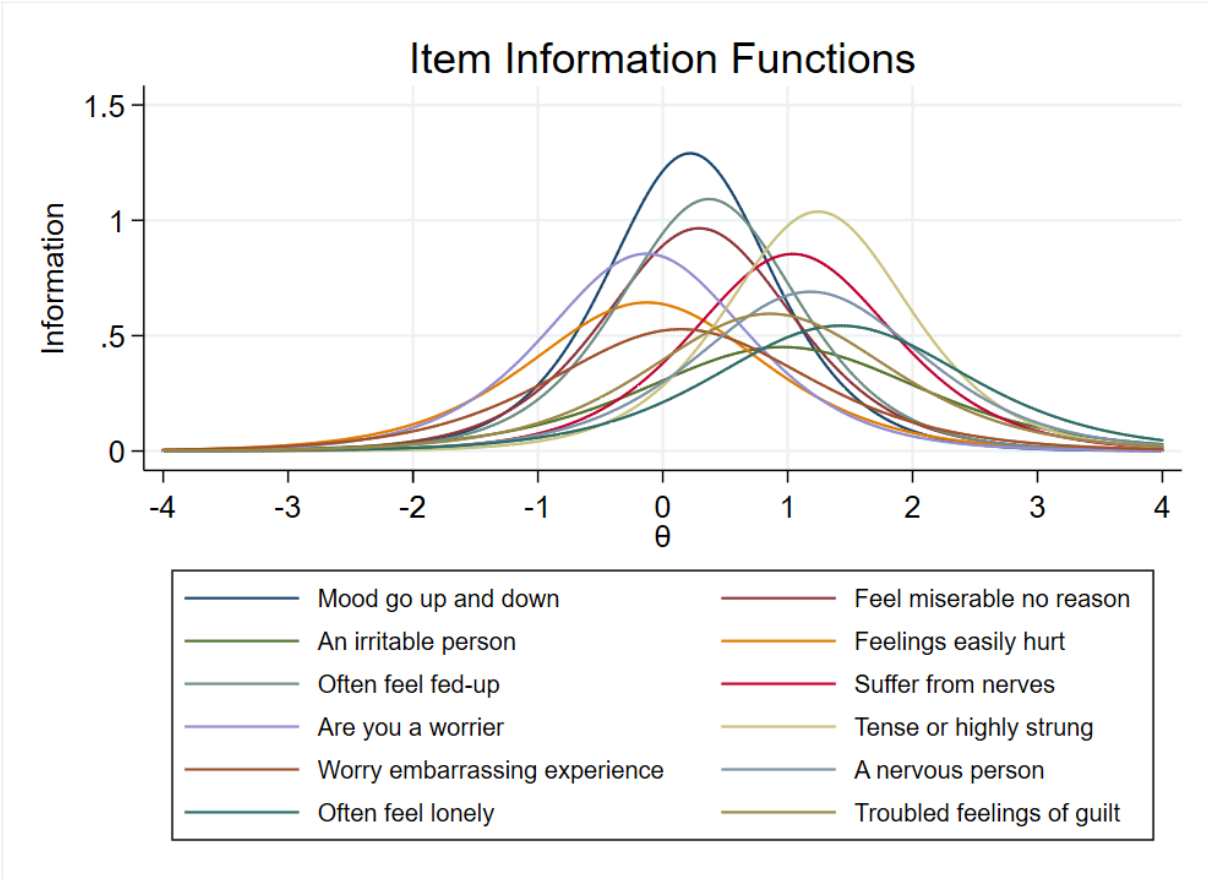

Figure 5. TIF graph for the 12-item scale

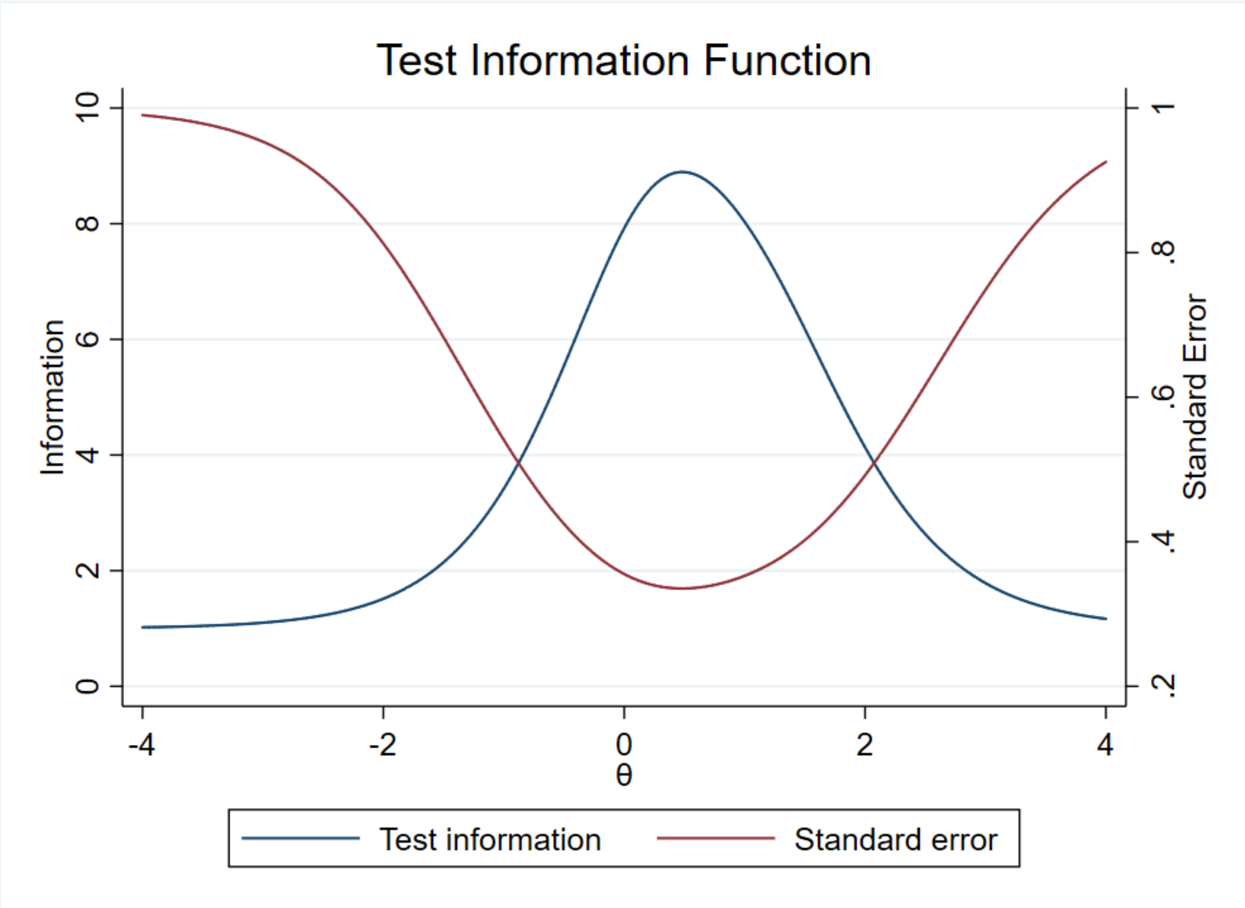

Figure 6. ICC graph for the 8-item scale

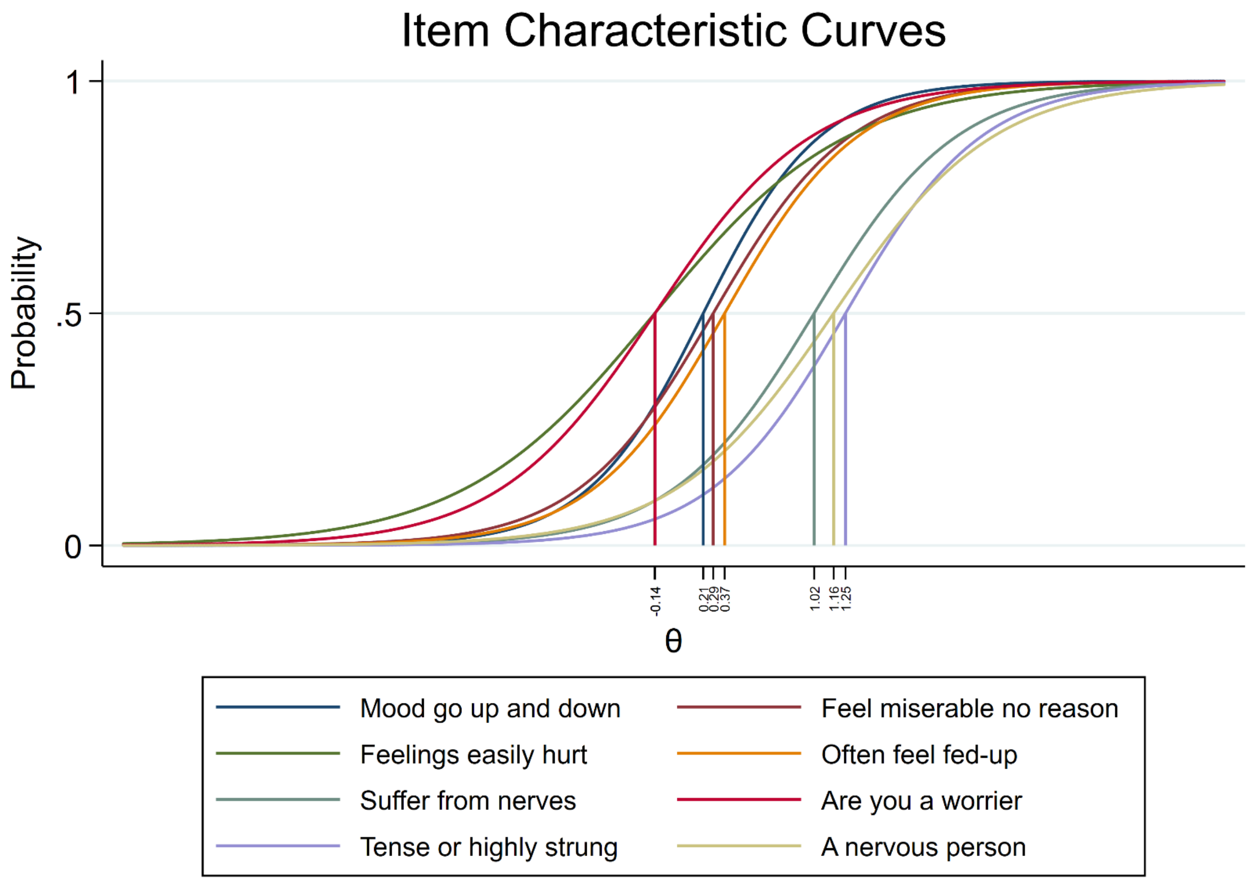

Figure 7. IIF for the 8-item scale

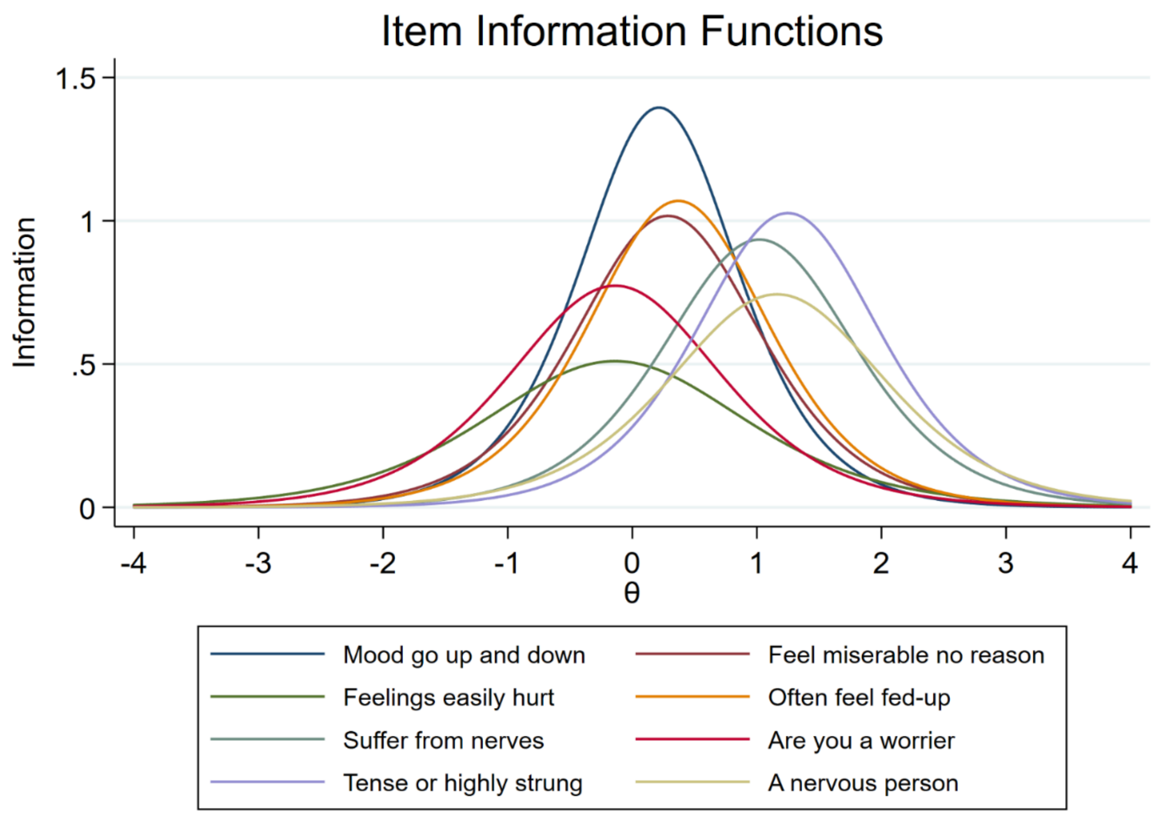

Figure 8. ICC for the 3-item scale

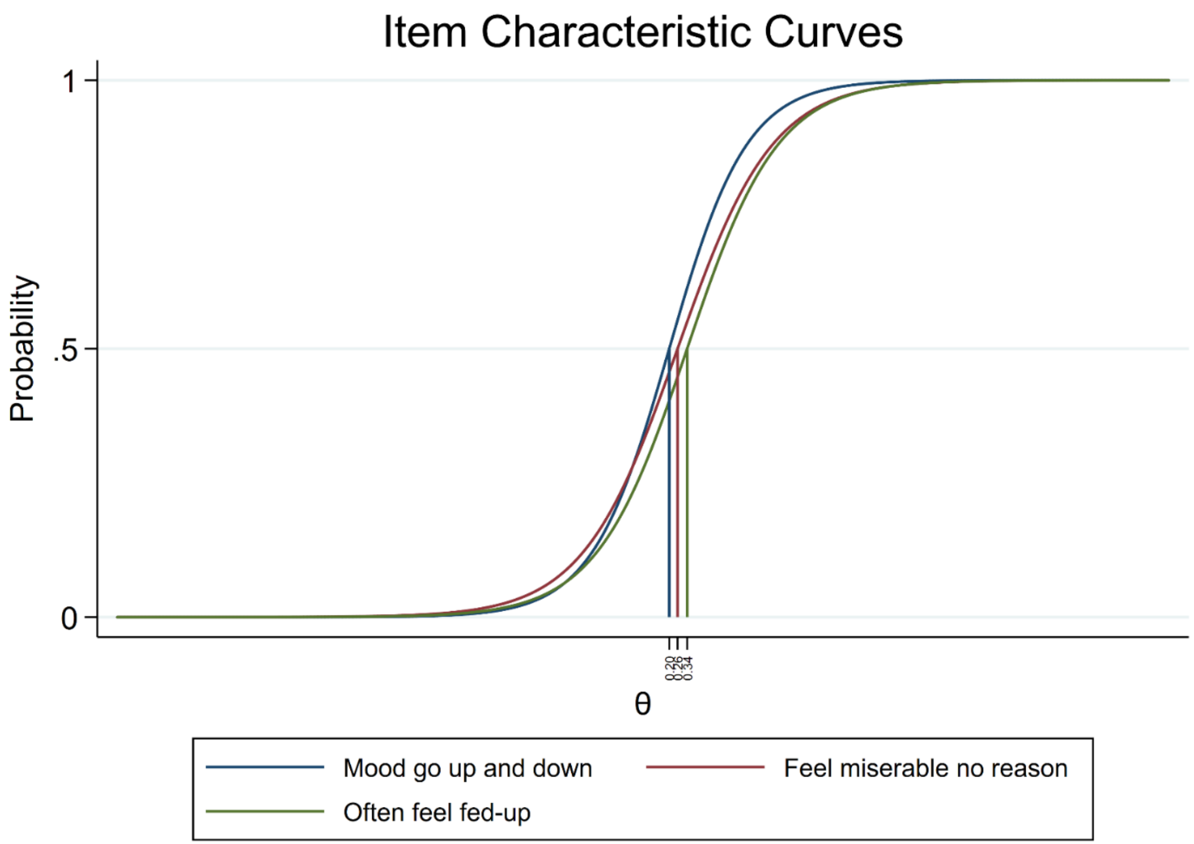

Figure 9. IIF for the 3-item scale