Bauermeister & Gallacher: Neuroticism

A psychometric evaluation of the 12-item EPQ-R neuroticism scale in
502,591 UK Biobank participants using item response theory (IRT)

Sarah Bauermeister¹ and John Gallacher¹

¹Department 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

1 **Abstract**

2 **Background**

3 Neuroticism has been described as a broad and pervasive personality dimension or
4 ‘heterogeneous’ trait measuring components of mood instability such as worry; anxiety;
5 irritability; moodiness; self-consciousness; sadness and irritability. Consistent with
6 depression and anxiety-related disorders, increased neuroticism places an individual
7 vulnerable for other unipolar and bipolar mood disorders. However, the measurement
8 of neuroticism remains a challenge. Our aim was to identify psychometrically efficient
9 items and inform the inclusion of redundant items across the 12-item EPQ-R
10 Neuroticism scale using Item Response Theory (IRT).

11

12 **Methods**

13 The 12-item binary EPQ-R Neuroticism scale was evaluated by estimating a two-
14 parameter (2-PL) IRT model on data from 502,591 UK Biobank participants aged 37 to
15 73 years ($M = 56.53$ years; $SD = 8.05$), 54% female. Models were run listwise ($n =$
16 401,648) and post-estimation mathematical assumptions were computed. All analyses
17 were conducted in STATA 16 SE on the Dementias Platform UK (DPUK) Data Portal.

18

19 **Results**

20 A plot of θ values (Item Information functions) showed that most items clustered
21 around the mid-range where discrimination values ranged from 1.34 to 2.28. Difficulty
22 values for individual item θ scores ranged from -0.13 to 1.41. A Mokken analysis

23 suggested a weak to medium level of monotonicity between the items, no items reach
24 strong scalability ($H=0.35-0.47$). Systematic item deletions and rescaling found that an
25 7-item scale is more efficient and with information (discrimination) ranging from 1.56
26 to 2.57 and stronger range of scalability ($H=0.47-0.52$). A 3-item scale is highly
27 discriminatory but offers a narrow range of person ability (difficulty). A logistic
28 regression differential item function (DIF) analysis exposed significant gender item bias
29 functioning uniformly across all versions of the scale.

30

31 **Conclusions**

32 Across 401,648 UK Biobank participants, the 12-item EPQ-R neuroticism scale exhibited
33 psychometric inefficiency with poor discrimination at the extremes of the scale-range.
34 High and low scores are relatively poorly represented and uninformative suggesting
35 that high neuroticism scores derived from the EPQ-R are a function of cumulative mid-
36 range values. The scale also shows evidence of gender item bias and future scale
37 development should consider the former along with item deletions.

38

39 **Keywords**

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

42 **Background**

43 Neuroticism has been described as a broad and pervasive personality dimension with
 44 influences beyond its own limited definition (1). Operationally, it has been defined as a
 45 personality trait assessed by items referencing to instances of worry; anxiety;
 46 irritability; moodiness; self-consciousness; sadness and irritability (2-4). The NEO-PI
 47 (Neuroticism-Extraversion-Openness Personality Inventory) operationalises neuroticism
 48 as a combination of individual behavioural traits which may also be measured as
 49 isolated components of mood state e.g., anxiety; hostility; depression; self-consciousness;
 50 impulsiveness and vulnerability (1). Also defined as a ‘heterogeneous’ trait possessing
 51 significant overlap with depression and anxiety, neuroticism places an individual
 52 vulnerable for other unipolar and bipolar mood disorders (3). Moreover, increased
 53 levels of neuroticism places an individual vulnerable to other neurotic disorders,
 54 psychological distress and ‘emotional instability’ (5). There is also consistent research
 55 suggesting a positive relationship between neuroticism and negative affect (6)
 56 notwithstanding neuroticism essentially existing as a dimension of negative affect (7).
 57 Eysenck has further argued that neuroticism is a direct reaction to the autonomic
 58 nervous system (8, 9), findings supported finding increased neuroticism correlated with
 59 tolerance to a highly stressed environment, suggesting a habituation relationship with
 60 everyday stressors (10, 11).

61

62 Eysenck’s attempts to define neuroticism and evaluate the measurement items thereof
 63 resulted in an original version of the Eysenck neuroticism scale existing as a component
 64 of the Maudsley Medical Questionnaire (12). Assessment outcomes of this scale were
 65 reported in the Manual for the Maudsley Personality Inventory (MPI) where gender

66 differences were found across the psychiatric patients and soldiers, on whom the data
67 were derived (13). Later versions of the MPI were revised to remove gender-specific
68 items although to our knowledge, details of the method of their removal are not
69 available. The revised neuroticism scale became a component of the Eysenck
70 Personality Questionnaire (EPQ-R: 14) and thereby exists as a culmination of attempts
71 to select the relevant items through multiple revisions of the MPI. Although the EPQ-R
72 neuroticism scale is reported to have been developed through clinical judgement and,
73 multiple cluster and factor analyses, reasons for acceptance or rejection of items were
74 complex, unclear and are not 'objectified' (13, 15).

75

76 Using factor analysis and correlations for item deletion whilst widely used, have a bias
77 towards identifying closely associated items as being informative and is opaque to the
78 individual item contribution or person ability. The process is commonly known as
79 classical test theory (CTT) whereby a summated score is computed from individual item
80 scoring. The EPQ-R neuroticism scale has been found to lack items to identify
81 respondents who would normally endorse items at the extreme ends of the trait
82 continuum, e.g. high vs. low neuroticism (5). Furthermore, the scale maintains gender-
83 specific items, females consistently scoring higher (14, 16), a difference which has been
84 reported cross-culturally (17) and across the age range (e.g., 18).

85

86 We investigated the psychometric efficiency of the 12-item EPQ-R neuroticism scale -
87 hereafter 'EPQ-R' (14) as a widely used measurement of neuroticism. We applied item
88 response theory (IRT) to psychometrically evaluate the EPQ-R using data from UK
89 Biobank (19), a large population study which assessed neuroticism at baseline. Our

expectation was that the large sample size and balanced gender ratio (54% female) would provide valuable item-level information for assessing the informativeness of individual items and overall psychometric reliability of the scale. This assessment may have important implications in clinical settings and for epidemiological research where it is widely utilised.

Methods

Participants

The UK Biobank is a large population-based prospective cohort study of 502,665 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 37 and 73 years. A response rate of 5.5% was recorded. Ethical approval was granted to Biobank from the Research Ethics Committee - REC reference 11/NW/0382 (19). The analysis here was applied to the whole cohort of 502,591 participants after withdrawals were considered.

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 12-item EPQ-R neuroticism scale (14) where participants were 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?'. The
 aim of the study was to investigate informativeness and reliability of the 12-item scale,
 and to explore revised versions using psychometric methodologies such as Item
 Response Theory (IRT).

IRT model

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

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

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 severity θ is returned and may be plotted as an item characteristic curve (ICC) along a standardised scale with a mean of 0 (Figure 1a). 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 1b). 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 (20). 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 (21) and analysed using STATA SE 16.1 (22)

Figure 1. Item Characteristic Curve (ICC) and Item Information Function (IIF) graphs

Results

Sample

The whole sample available after withdrawals were considered was 502,591 UK Biobank participants aged 37 to 73 years ($M = 56.53$ years; $SD = 8.05$), 54% female. Models were run listwise and the number of participants included in analyses were: 401,648 (12-item); 434,693 (7-item); 473,940 (3-item).

IRT analysis

A 2-PL IRT model was estimated whereby difficulty and discrimination parameters were computed (Table 1). The discrimination (item-information) parameters across the scale range between 1.34 and 2.28. The item measuring ‘Does your mood often go up and down?’ exhibits the highest level of discrimination at 2.28, 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 (20). 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.

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

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 2 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 0.21 (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% chance of someone with a θ of -0.13 endorsing this item, therefore, someone who does not experience neurotic trait characteristics.

Figure 2. ICC graph for the 12-item scale

Additional item discrimination is available by graphing the IIF curves (see Figure 3). The IIF curves thereby display the relationship between difficulty (trait level) and discrimination (information), 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 level of neuroticism (1.41) 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 (0.21). 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 to +4) or a low level

(absence) of the trait (-1 to -4), instead it provides the most information for respondents who possesses an average ($\theta=0$) to a minimal amount of the neuroticism trait (see Table 1). The item which possesses the least trait characteristic discrimination is the item, 'Are you an irritable person?', Although the IIF curve apex is positioned over a positive θ (0.95), and may be endorsed by a respondent possessing an amount of the trait characteristic, the discrimination value is low (1.34).

Figure 3. IIF graph for the 12-item scale

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 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 to no level of neuroticism ($\theta = -0.13$ to 1.41).

Reliability

In IRT, reliability may be calculated at multiple point values of θ along the continuum rather than a single reliability score as in CTT. Reliability is 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, distinguishing respondents according to specific values of θ (23). For the 12-item scale there is reliable information to differentiate respondents who possess no or just above an average amount of trait

information ($\alpha=0$; 0.87 and $\alpha=1$; 0.88), considered very good for reliability. However, reliability then decreases ($\alpha=2$; 0.76 and $\alpha=-1$; 0.71) suggesting that the highest reliability of measuring the neurotic trait is at normal or a minimal amount of neuroticism, $\alpha=0$ or 1. Thereafter, reliability reduces so that the extreme end of the continuum, $\alpha=3$; 4; -2; -3; -4, is no longer reliably measured (See Table 2).

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

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, requested after estimation of a single-factor model showed that no residuals were too highly correlated, $R > 0.20$ (24), suggesting basic item independence.

2. Monotonicity

A Mokken analysis produced a Loevinger H coefficient (25) which measures the scalable quality of items, expressed as a probability measure, independent of a respondent's α . 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 ($H \geq 0.5$) (25).

248

249

250 **Table 3. Mokken analysis with Loevinger H coefficients for the 12-item scale**

251

252 3. Unidimensionality

253 A principal component analysis (PCA) shows that a single major factor is responsible
 254 for 36% of the variance and a second factor responsible for 11% of the variance, the
 255 difference of which is above the suggested 20% indicating a single major factor is being
 256 measured (26). A post-IRT estimation model measure of unidimensionality was also
 257 computed using a semi-partial correlation controlling for η . This analysis provides
 258 individual item variance contribution after adjusting for all the other variables including
 259 η . It demonstrates the relationship between local independence and unidimensionality,
 260 reflecting a conservative assessment whereby the desired R^2 should ideally be zero or
 261 as close to zero as possible (27). Items ranged between 0.01 and 0.02, suggesting
 262 unidimensionality. To our knowledge, there is still no standardised cut-off criterium for
 263 assessing this value (i.e., how close to zero all items should be across a scale).

264

265 **IRT revised analysis**

266 To assess a revised scale, items were systematically removed from the scale according
 267 to discrimination value with the lowest discriminating item removed first ('Are you an
 268 irritable person?', 1.34) whereafter a 11-item 2-PL IRT model was estimated with the
 269 remaining items and the process repeated, removing the lowest discriminating item,
 270 below 1.7. In order of removal, the items systematically removed thereafter were: 'Do

271 you often feel lonely?'; 'Are you often troubled by feelings of guilt?'; 'Do you worry too
272 long after an embarrassing experience?' and 'Are your feelings easily hurt?' at which
273 stage the 7 remaining items were maintained as most were > 1.70 on 434,693
274 individuals.

275

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

285

286 **Table 4. 2-PL IRT model item parameters for the 7-item scale**

287

288 **Table 5. Mokken analysis with Loevinger H coefficients for the 7-item scale**

289

290 **Table 6. Reliability for values of α from a 2-PL IRT model fit for the 7-item scale**

291

292 **Figure 4. ICC graph for the 7-item scale**

293

294 **Figure 5. IIF graph for the 7-item scale**

295

296

297 Further item reduction was explored to investigate a ‘minimal’ scale. After systematic
 298 item-removal, three items remained when the 2-PL was estimated on sample of 473,940
 299 individuals. The scale parameters suggest those items which possessed high
 300 discrimination and positive difficulty values, ‘Does your mood often go up and down?’
 301 (3.44; 0.14); ‘Do you ever feel just miserable for no reason?’ (2.79; 0.22) and ‘Do you
 302 often feel fed-up?’ (2.92; 0.28) (Table 7). A Mokken analysis suggests that scalability is
 303 strong ($H \geq 0.50$) across all items (Table 8), a semi-partial correlation analysis
 304 controlling for η showed all values were 0.00. Reliability is only good at $\eta = 0$
 305 suggesting this scale is only reliable to measure those with an average trait (Table 9).
 306 The ICC and IIF graphs suggest the three-item scale may present an efficient, alternative
 307 and highly informative scale, however, the scale is narrow in range and does not
 308 possess items measuring neurotic traits above or below average η , at the extreme ends
 309 of the trait spectrum (Figure 6).

310

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

312

313 **Table 8. Mokken analysis with Loevinger H coefficients for the 3-item scale**

314

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

Figure 6. ICC and IIF graphs for the 3-item scale

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 over the other (male vs. female) 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.

Table 10. Logistic regression differential item function (DIF) analysis across 12, 7 and 3-item scales

Discussion

In a large population cohort of 502,591 adults aged 37-73 years, limitations in the range and reliability of item trait characteristics were found across the 12-item EPQ-R neuroticism scale when a 2PL IRT model was estimated listwise on 401,648 individuals. Our findings suggest that the 12-item scale 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 function analysis also suggests there is poor reliability at the extremes of the scale score and high neuroticism scores derived from the EPQ-R are a function of accumulative mid-range values. In a revised 7-item version of the scale, greater item-discrimination and reliability was found across the scale suggesting that selective items within the 12-item version are redundant. A further reduced 3-item version was investigated but although this scale possesses items of high discrimination and scalability, range is very narrow and lacks reliability beyond an average trait value. A DIF analysis with gender as a group outcome suggests the scale exhibits significant gender differential item functioning 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 presented here, and as evidenced in the scale-revision analysis.

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 further computation of post-estimation assumptions including the estimation of an individual

latent metric predictive of individual θ scores on the fitted IRT model. This θ metric may then be used as a latent construct in assessing longitudinal change (28) which may be a more reliable measure compared to a single summated score (29). Furthermore, it is also suggested that using an IRT derived θ in longitudinal studies, over the summated score, may be preferable with reducing overestimation of the repeated measure variance and underestimation of the between-person variance (30).

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, θ is automatically computed in response to the trait (θ) of the respondent and it is therefore not necessary to present the full range of items as the response scale is adaptive to individual performance (trait level), the items underlying the trait and a stopping rule (31). 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 to consider when working with individuals who are older or who have co morbid psychiatric 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 (28). Moreover, previous research suggests that both the 12 and 3-item EPQ-R neuroticism scales may have reduced power to discriminate

between low and high scoring individuals (5); 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., 32, 33).

Conclusions

The 12-item neuroticism EPQ-R scale lacks item reliability and neurotic trait-specific information at the extreme ends of the neurotic continuum when a 2-PL IRT model is estimated. A secondary analysis suggests that systematic item-elimination and re-estimation of the 2-PL model produces a 7-item 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 including item deletions and validation of replacement items which consider gender item bias. 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.

Declarations

Ethics approval and consent to participate:

Analysis of secondary data only with ethical approval in place from source cohort, UK Biobank Research Ethics Committee - REC reference 11/NW/0382.

Consent for publication:

407 SB and JG give full consent for publication

408

409 **Availability of data and materials:**

410 The dataset(s) supporting the conclusions of this article is(are) available in the
 411 Dementias Platform UK (DPUK) Data Portal repository,
 412 <https://portal.dementiasplatform.uk/>.

413

414 **Competing interests:**

415 SB and JG declare no competing interests

416

417 **Funding:**

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

419

420 **Authors' contributions:**

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

424

425 **Acknowledgements:**

426 This is a Dementias Platform UK (DPUK) supported project with all analyses conducted
 427 on the DPUK Data Portal, constituting part 1 of DPUK Application 0169.

428

429 **Abbreviations**

430	DPUK	Dementias Platform UK
431	EPQ-R	Eysenck Personality Questionnaire-Revised
432	ICC	Item Characteristic Curve
433	IIF	Item Information Function
434	IRF	Item Response Function
435	IRT	Item Response Theory
436	PCA	Principal Component Analysis

437

References

1. Costa PT, Jr., McCrae RR. Neuroticism, somatic complaints, and disease: is the bark worse than the bite? *J Pers.* 1987;55(2):299-316.
2. Costa PT, Jr., McCrae RR. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *J Pers Soc Psychol.* 1980;38(4):668-78.
3. Lahey BB. Public health significance of neuroticism. *Am Psychol.* 2009;64(4):241-56.
4. Costa PT, Jr., McCrae RR. Four ways five factors are basic. *Personality and Individual Differences.* 1992;13(6):653-65.
5. Birley AJ, Gillespie NA, Heath AC, Sullivan PF, Boomsma DI, Martin NG. Heritability and nineteen-year stability of long and short EPQ-R neuroticism scales. *Personality and Individual Differences.* 2006;40(4):737-47.
6. Rusting CL. Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. *Psychol Bull.* 1998;124(2):165-96.
7. Watson D, Clark LA. Negative affectivity: the disposition to experience aversive emotional states. *Psychol Bull.* 1984;96(3):465-90.
8. Eysenck HJ. *The biological basis of personality.* London: Springfield, III: Charles C. Thomas; 1967.
9. Eysenck HJ. *Personality; biological foundation. the neurophysiology of individual difference.* New York: Academic Press; 1994.
10. Farrington D, Jolliffe D. *Personality and Crime.* In: N. J. Smelser PBB, editor. *International Encyclopedia of the Social & Behavioral Sciences.* 1st ed. USA: Elsevier; 2001.
11. LeBlanc J, Ducharme MB, Thompson M. Study on the correlation of the autonomic nervous system responses to a stressor of high discomfort with personality traits. *Physiol Behav.* 2004;82(4):647-52.
12. Faulwasser H, Kittlaus H. [Economy of the Maudsley Medical Questionnaire (MMQ)]. *Psychiatr Neurol Med Psychol (Leipz).* 1973;25(5):276-81.

Bauermeister & Gallacher: Neuroticism

- 464 13. Francis LJ. The Dual Nature of the Eysenckian Neuroticism Scales - a Question of Sex-
465 Differences. *Personality and Individual Differences*. 1993;15(1):43-59.
- 466 14. Eysenck SB, Eysenck HJ, Barrett P. A revised version of the psychoticism scale.
467 *Personality and Individual Differences*. 1985;6:21-9.
- 468 15. Eysenck HJ, Eysenck SBG. Psychoticism as a dimension of personality. London: Hodder &
469 Stoughton; 1976.
- 470 16. Allsop J, Eysenck HJ, Eysenck SBG. Machiavellianism as a component in psychoticism and
471 extraversion. *Personality and Individual Differences*. 1991;12:29-41.
- 472 17. Eysenck HJ, Eysenck SBG. Recent advances in the cross-cultural study of personality. In:
473 Spielberger CD, Butcher JN, editors. *Advances in personality development*. Hillsdale, USA.:
474 Erlbaum; 1982. p. 41-69.
- 475 18. Eysenck SB, Abdel-Khalek AM. A cross-cultural study of personality: egyptian and
476 english children. *Int J Psychol*. 1989;24(1-5):1-11.
- 477 19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open
478 Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and
479 Old Age. 2015;12(3).
- 480 20. Baker FB. The basics of item response theory. Original work published in 1985
481 <http://echo.edres.org:8080/irt/baker/final.pdf>; College Park, DM: ERIC Clearinghouse on
482 Assessment and Evaluation; 2001.
- 483 21. Bauermeister S, Orton C, Thompson S, Barker RA, Bauermeister JR, Ben-Shlomo Y, et al.
484 Data Resource Profile: The Dementias Platform UK (DPUK) Data Portal. *BioRxiv*. Preprint.
- 485 22. StataCorp L. Stata SE 16.1: StataCorp LLC; 2019 [Available from: stata@stata.com.
- 486 23. Thissen. Reliability and measurement precision. In: Wainer H, editor. *Computerized*
487 *Adaptive Testing: A primer*. London: Lawrence Erlbaum: Lawrence Erlbaum; 2000. p. 159-84.
- 488 24. Yen WM. Scaling performance assessments: Strategies for managing local item
489 dependence. *Journal of Educational Measurement*. 1993;30:187-213.

Bauermeister & Gallacher: Neuroticism

- 490 25. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. London:
491 Sage Publications; 2002.
- 492 26. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric
493 evaluation and calibration of health-related quality of life item banks: plans for the Patient-
494 Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl
495 1):S22-31.
- 496 27. De Mars C. Item Response Theory. New York, USA: Oxford University Press; 2010.
- 497 28. Acock AC. A Gentle Introduction to Stata. 5th ed. Texax, USA: A Stata Press Publication;
498 2016.
- 499 29. Lu IRR. Embedding IRT in structural equation models: A comparison with regression
500 based on IRT scores. Struct Equ Modeling. 2005;12(2):263-77.
- 501 30. Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal
502 questionnaire data analysis in medical research. BMC Med Res Methodol. 2015;15:55.
- 503 31. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ, Steinberg L, et al. Computerized
504 Adaptive Testing: A Primer. 2nd ed. New York, USA: Routledge; 2014. 335 p.
- 505 32. de Ayala RJ. The Theory and Practice of Item Response Theory. USA: The Guildford
506 Press; 2009.
- 507 33. Streiner DL, Norman GR, Cairney J. Health Measurement Scales. 5th ed. Great Britain:
508 Oxford University Press; 2015.

509

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	2.28	0.01	222.03	(2.26 2.30)
	Difficulty	0.21	0.00	84.64	(0.21 0.22)
Feelings easily hurt?	Discrimination	1.60	0.01	230.62	(1.59 1.62)
	Difficulty	-0.13	0.00	-44.94	(-0.14 -0.12)
Are you a worrier?	Discrimination	1.85	0.01	229.06	(1.83 1.87)
	Difficulty	-0.13	0.00	-49.27	(-0.14 -0.12)
Suffer from nerves?	Discrimination	1.85	0.01	210.41	(1.83 1.87)
	Difficulty	1.03	0.00	268.82	(1.03 1.05)
Feel miserable no reason?	Discrimination	1.97	0.01	228.19	(1.95 1.99)
	Difficulty	0.29	0.00	106.76	(0.28 0.29)
Often feel fed-up?	Discrimination	2.09	0.01	244.94	(2.07 2.11)
	Difficulty	0.36	0.00	135.93	(0.36 0.37)
Tense or highly strung?	Discrimination	2.05	0.01	203.01	(2.03 2.07)
	Difficulty	1.23	0.00	299.42	(1.22 1.24)
Often feel lonely?	Discrimination	1.47	0.01	198.86	(1.46 1.49)
	Difficulty	1.41	0.01	256.50	(1.40 1.42)
An irritable person?	Discrimination	1.34	0.01	211.83	(1.33 1.35)
	Difficulty	0.95	0.00	217.40	(0.95 0.96)

Bauermeister & Gallacher: Neuroticism

A nervous person?	Discrimination	1.67	0.01	208.65	(1.65 1.69)
	Difficulty	1.17	0.00	266.14	(1.16 1.18)
Worry embarrassing experience?	Discrimination	1.45	0.01	226.58	(1.44 1.46)
	Difficulty	0.15	0.00	48.19	(0.14 0.15)
Troubled feelings of guilt?	Discrimination	1.54	0.01	220.37	(1.53 1.56)
	Difficulty	0.86	0.00	224.84	(0.85 0.86)

Note: Item names truncated for brevity, see text; B=standardised beta coefficients; SE=standard error
 $p < .000$ for all B values.

517

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

θ	TIF	TIF SE	Reliability
-4	1.02	0.99	0.02
-3	1.10	0.95	0.09
-2	1.52	0.81	0.34
-1	3.43	0.54	0.71
0	7.96	0.35	0.87
1	8.04	0.35	0.88
2	4.11	0.49	0.76
3	1.77	0.75	0.44
4	1.16	0.93	0.14

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

518

519

Table 3. Mokken analysis with 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

520

Table 4. 2-PL IRT model item parameters for the 7-item scale

Item	Parameter	B	SE	Z	95% CI
Mood go up and down?	Discrimination	2.57	0.01	195.10	(2.54 2.59)
	Difficulty	0.19	0.00	82.07	(0.18 0.19)
Are you a worrier?	Discrimination	1.56	0.01	215.41	(1.54 1.57)
	Difficulty	-0.16	0.00	-57.07	(-0.17 -0.16)
Suffer from nerves?	Discrimination	1.79	0.01	196.09	(1.77 1.81)
	Difficulty	1.04	0.00	265.19	(1.03 1.05)
Feel miserable no reason?	Discrimination	2.15	0.01	205.06	(2.00 2.04)
	Difficulty	0.29	0.00	105.39	(0.28 0.29)
Often feel fed-up?	Discrimination	2.17	0.01	209.88	(2.17 2.21)
	Difficulty	0.34	0.00	136.47	(0.34 0.35)
Tense or highly strung?	Discrimination	1.98	0.01	194.87	(1.96 2.00)
	Difficulty	1.25	0.00	297.66	(1.24 1.26)
A nervous person?	Discrimination	1.70	0.01	199.42	(1.69 1.72)
	Difficulty	1.15	0.00	269.35	(1.15 1.16)

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

$p < .000$ for all B values.

525

526

Table 5. Mokken analysis with Loevinger H coefficients for the 7-item scale

Item	H
Mood go up and down?	0.50
Are you a worrier?	0.50
Suffer from nerves?	0.49
Feel miserable no reason?	0.47
Often feel fed-up?	0.47
Tense or highly strung?	0.52
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 7-item scale

θ	TIF	TIF SE	Reliability
-4	1.01	1.00	1.01
-3	1.04	0.98	0.04
-2	1.24	0.90	0.20
-1	2.38	0.65	0.58
0	6.24	0.40	0.84
1	5.78	0.42	0.83
2	2.82	0.60	0.65
3	1.37	0.85	0.27
4	1.06	0.97	0.06

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	3.44	0.03	128.71	(3.39 3.49)
	Difficulty	0.14	0.00	74.48	(0.14 0.15)
Feel miserable no reason?	Discrimination	2.79	0.02	182.49	(2.76 2.82)
	Difficulty	0.22	0.00	103.80	(0.21 0.22)
Often feel fed-up?	Discrimination	2.92	0.02	173.92	(2.89 2.96)
	Difficulty	0.28	0.00	133.95	(0.28 0.28)

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

$p < .000$ for all B values

535

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

Item	H
Mood go up and down?	0.58
Feel miserable no reason?	0.55
Often feel fed-up?	0.57

536

Note. Item names truncated for brevity, see text.

537

538

539

540

541

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

θ	TIF	TIF SE	Reliability
-4	1.00	1.00	0.00
-3	1.00	1.00	0.00
-2	1.03	0.98	0.03
-1	1.66	0.76	0.40
0	7.39	0.37	0.86
1	3.09	0.57	0.68
2	1.13	0.94	0.11
3	1.01	1.00	0.01
4	1.00	1.00	0.00

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

546

Table 10. Logistic regression differential item function (DIF) analysis across 12, 7 and 3-item scales

Item	12-item scale		8-item scale		3-item scale	
	Nonuniform	Uniform	Nonuniform	Uniform	Nonuniform	Uniform
	Chi2	P	Chi2	P	Chi2	P
Mood go up and down?	29.34***	2268.50***	76.58***	1518.61***	5.10*	1623.93***
Feelings easily hurt?	74.44***	5355.91***				
Are you a worrier?	92.87***	2502.24***	66.89***	4628.85***		
Suffer from nerves?	12.62***	74.00***	11.22	31.54***		
Feel miserable no reason?	35.49.08***	1152.08***	26.20***	2426.29***	16.19***	5131.04***
Often feel fed-up?	49.72***	1267.51***	119.35***	659.99***	53.50***	931.90***
Tense or highly strung?	51.49***	152.88***	70.14***	22.37***		
Often feel lonely?	22.35***	149.19***				
An irritable person?	0.04	8256.58***				
A nervous person?	604.41***	3286.63***	792.74***	3174.96***		
Worry embarrassing experience?	3.81*	1385.12***				
Troubled feelings of guilt?	38.38***	1991.85***				

547 Note: Item names truncated for brevity, see text.

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

Figures

Figure 1. Item Characteristic Curve (ICC) and Item Information Function (IIF) graphs

554

Figure 2. ICC graph for the 12-item scale

555

556

557

Figure 3. IIF graph for the 12-item scale

Figure 4. ICC graph for the 7-item scale

Figure 5. IIF graph for the 7-item scale

Figure 6. ICC and IIF graphs for the 3-item scale

Bauermeister & Gallacher: Neuroticism

579