Final project for Statistical Methods in Psychometrics

Michaela Cichrová, Kristína Sakmárová

January 2024

1 Data

The DASS 42 depression anxiety stress scale test is designed to evaluate the severity of mental disorder symptoms associated with anxiety, stress and depression. The test has 42 items in total, each item is associated with one of the three measured constructs, i.e. with depression, anxiety or stress. Each item has 4 possible answers:

- 1 = Did not apply to me at all
- 2 = Applied to me to some degree, or some of the time
- 3 = Applied to me to a considerable degree, or a good part of the time
- 4 = Applied to me very much, or most of the time

Our dataset consists of data collected on volunteers with an on-line version of the DASS 21 in 2017 - 2019. The volunteers took the test for free to get personalised results and could participate in a short research survey at the end of the DASS test. This dataset includes those who agreed to complete the research survey and also answered yes to the question "Have you given accurate answers and may they be used for research?". There were 39775 respondents in total. Example of a question is shown in the figure 1.

The items included are:

1. I found myself getting upset by quite trivial things.
2. I was aware of dryness of my mouth.
3. I couldn't seem to experience any positive feeling at all.
4. I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).
5. I just couldn't seem to get going.
6. I tended to over-react to situations.
7. I had a feeling of shakiness (eg, legs going to give way).
8. I found it difficult to relax.
9. I found myself in situations that made me so anxious I was most relieved when they ended.
10. I felt that I had nothing to look forward to.
11. I found myself getting upset rather easily.
12. I felt that I was using a lot of nervous energy.
13. I felt sad and depressed.
14. I found myself getting impatient when I was delayed in any way (e.g. elevators, traffic lights, being kept waiting).
15. I had a feeling of faintness.

16. I felt that I had lost interest in just about everything.
17. I felt I wasn't worth much as a person.
18. I felt that I was rather touchy.
19. I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.
20. I felt scared without any good reason.
21. I felt that life wasn't worthwhile.
22. I found it hard to wind down.
23. I had difficulty in swallowing.
24. I couldn't seem to get any enjoyment out of the things I did.
25. I was aware of the action of my heart in the absence of physical exertion (e.g. sense of heart rate increase, heart missing a beat).
26. I felt down and blue.
27. I found that I was very irritable.
28. I felt I was close to panic.
29. I found it hard to calm down after something upset me.
30. I feared that I would be thrown by some trivial but unfamiliar task.
31. I was unable to become enthusiastic about anything.
32. I found it difficult to tolerate interruptions to what I was doing.
33. I was in a state of nervous tension.
34. I felt I was pretty worthless.
35. I was intolerant of anything that kept me from getting on with what I was doing.
36. I felt terrified.
37. I could see nothing in the future to be hopeful about.
38. I felt that life was meaningless.
39. I found myself getting agitated.
40. I was worried about situations in which I might panic and make a fool of myself.
41. I experienced trembling (e.g. in the hands).
42. I found it difficult to work up the initiative to do things.

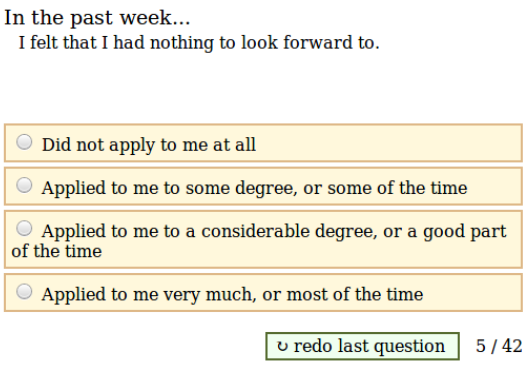

Figure 1: Demonstration of the questionnaire item responses.

These responses are stored in variable ending with letter A (e.g. Q1A). Also recorded was the time taken in milliseconds to answer that question (E) and that question's position in the survey (I). Other variables recorded were: time spent on the introduction/landing page (in seconds), time spent on all the DASS questions (should be equivalent to the time elapsed on all the individual questions combined), time spent answering the rest of the demographic and survey questions.

On the next page was a generic demographics survey with many different questions. The Ten Item Personality Inventory was administered (see Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528.):

- TIPI1 Extraverted, enthusiastic.
- TIPI2 Critical, quarrelsome.
- TIPI3 Dependable, self-disciplined.
- TIPI4 Anxious, easily upset.
- TIPI5 Open to new experiences, complex.
- TIPI6 Reserved, quiet.
- TIPI7 Sympathetic, warm.
- TIPI8 Disorganized, careless.
- TIPI9 Calm, emotionally stable.
- TIPI10 Conventional, uncreative.

The TIPI items were rated "I see myself as:".... such that:

- 1 = Disagree strongly
- 2 = Disagree moderately
- 3 = Disagree a little
- 4 = Neither agree nor disagree
- 5 = Agree a little
- 6 = Agree moderately
- 7 = Agree strongly

The following items were presented as a check-list and subjects were instructed "In the grid below, check all the words whose definitions you are sure you know":

- VCL1 boat
- VCL2 incoherent
- VCL3 pallid
- VCL4 robot
- VCL5 audible
- VCL6 cuivocal
- VCL7 paucity
- VCL8 epistemology
- VCL9 florted

- VCL10 decide
- VCL11 pastiche
- VCL12 verdid
- VCL13 abysmal
- VCL14 lucid
- VCL15 betray
- VCL16 funny

A value of 1 is checked, 0 means unchecked. The words at VCL6, VCL9, and VCL12 are not real words and will be used as a validity check in the next section. Other information recorded includes: education, what type of area did the person live at as a child, gender, whether English is the natural language, age, hand used to write (Right, Left, Both), religion (Agnostic, Atheist, Buddhist, Christian - Catholic, Christian - Mormon, Christian - Protestant, Christian - Other, Hindu, Jewish, Muslim, Sikh, Other), sexual orientation (Heterosexual, Bisexual, Homosexual, Asexual, Other), race (Asian, Arab, Black, Indigenous Australian, Native American, White, Other), whether the person voted in a national election in the past year, marital status, how many children did the person's mother have, major at the university. The following values were derived from technical information: ISO country code of where the user connected from, screensize (device with small screen, device with big screen), how the user found the test (from the front page of the site hosting the survey, from google, other or unknown).

Out of all these variables, we will work with only few chosen ones in the next sections. We also need to add a criterion variable, which will be used in the validity section. As mentioned above, items VCL6, VCL9 and VCL12 are made-up words and we will use them as validity check. New binary variable criterion is defined as follows: $\text{criterion} = \mathbb{1}\{VCL6 = 0, VCL9 = 0, VCL12 = 0\}$.

The table below includes basic characteristics for age and total score of the respondents. As we can see, there are some spurious age values in the data (maximum age 1998). Therefore, for the 7 respondents, whose age exceeded 100 years, we replaced the dubious values with NA. We can see the basic characteristics after this arrangement in the table as well. As for the total score, the higher the values, the bigger issues with depression, stress and anxiety had the respondent. There may have been some respondents with none of these problems according to the minimum value of total score.

	Min	Max	Mean	1. quartile	Median	3. quartile	Standard deviation
Age	13.00	1998.00	23.61	18.00	21.00	25.00	21.58
Corrected age	13.00	99.00	23.40	18.00	21.00	25.00	8.58
Total score	42.0	168.0	100.3	77.0	100.0	123.0	30.03

Table 1: Sample characteristics for age and total score.

The next table summarizes some of the categorical variables in the dataset. Vast majority of the respondents were females. Both high school education and university degree were approximately equally represented. Most of the respondents have never been married and English was not native language for more than half of the respondents.

Variable	Group	Counts	Relative counts
gender	Male	8789	22.1%
	Female	30367	76.35%
	Other	552	1.39%
	Missing	67	0.17%
education	Less than high school	4066	10.22%
	High school	15066	37.88%
	University degree	15120	38.01%
	Graduate degree	5008	12.59%
	Missing	515	1.29%
married	Never	34131	85.81%
	Currently	4357	10.95%
	Previously	1092	2.75%
	Missing	195	0.49%
English natural language	Yes	14380	36.15%
	No	25343	63.72%
	Missing	52	0.13%
Criterion	Yes	34583	86.95%
	No	5192	13.05%

Table 2: Representation of factor variables.

The figure below shows graphically characteristics for total score from table 2.

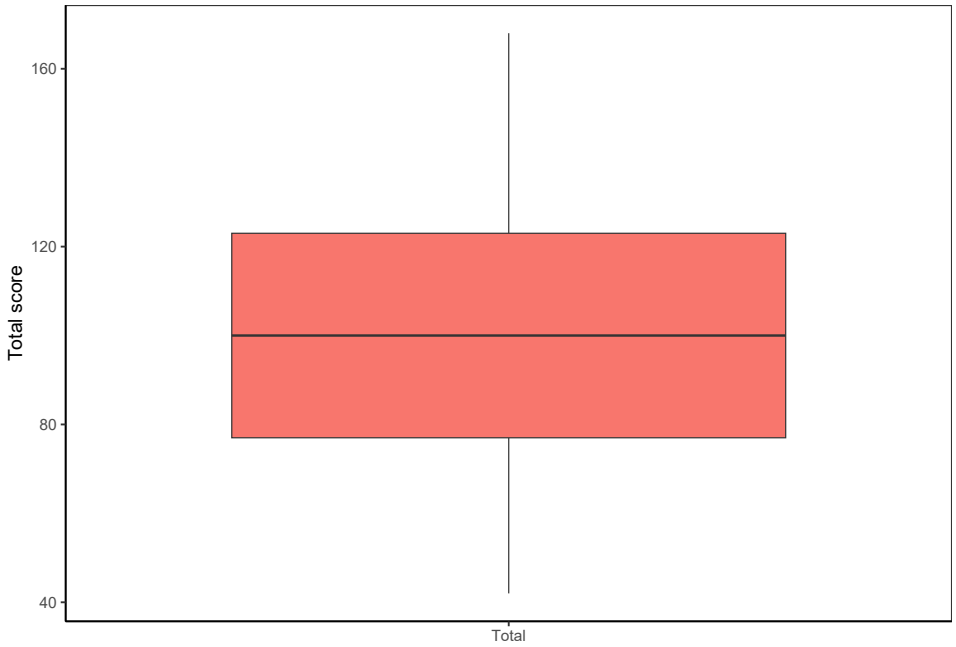

Figure 2: Box-plot for total score for each participant.

As we can see on the next figure, normal density approximates the frequencies of total score relatively well.

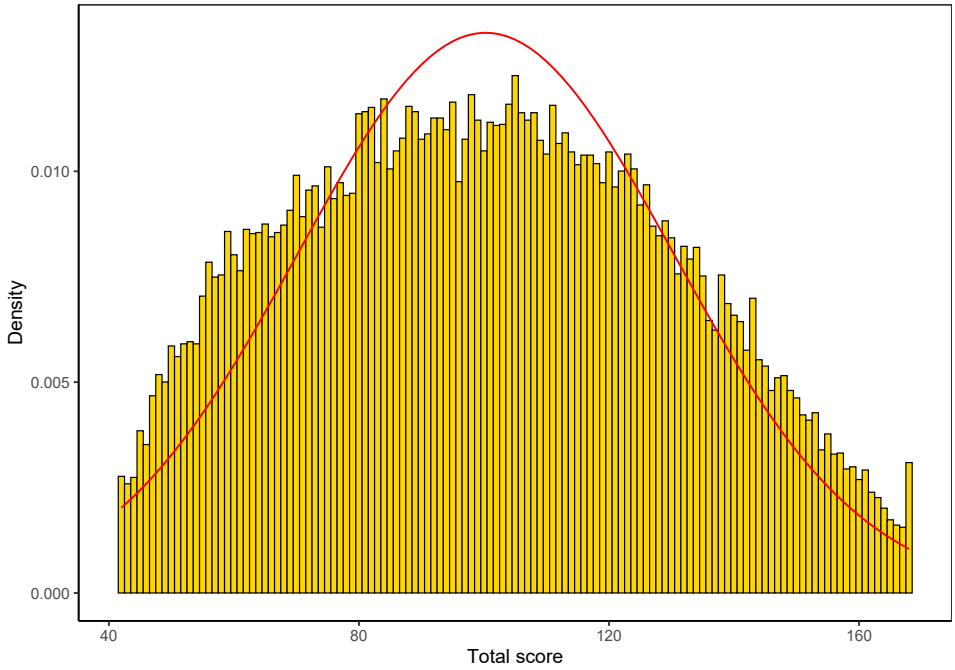

Figure 3: Histogram for the total score with fitted normal distribution density function.

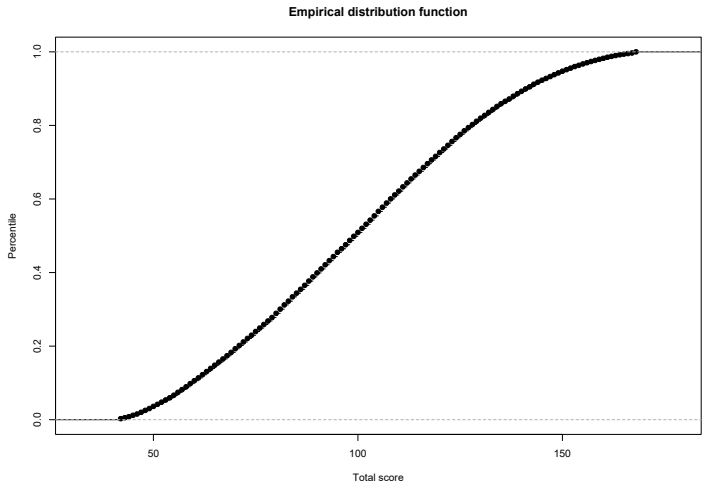

Figure 4: The empirical cumulative distribution function for total scores.

Each item was presented one at a time in a random order for each new participant along with a 4 point rating scale asking the user to indicate how often that had been true of them in the past week. The data can be found at "$https://openpsychometrics.org/\_rawdata/$".

The total score for each respondent was calculated as the sum of scores from all questions. For each respondent we calculated Z-scores, T-scores, percentiles and success rate. No other manipulation with the data was needed.

For the first respondent the z-score was 1.42, which means, that the score of this respondent was 1.42 sd above the mean score. The t-score for this same respondent is 64.23, the success rate 85.11 and percentile 91. All of these quantities indicate relatively high total score of the respondent and we can assume that there are some problems with either depression, anxiety or stress.

2 Test validity

Test intends to measure depression, anxiety and stress. The test scores should quantify some of the symptoms of depression, anxiety and stress. Each item should refer to one of the constructs. As the questions are pretty straightforward, we can't think of another use of the items other than in this type of the test. To check the validity of test scores based on test content we would need to assure that the questions lead to characterization of the symptoms of depression, stress and anxiety and could use data on whether some experts deem the items important or not. The symptoms for stress, depression and anxiety can be be found on the following links:

- *https://www.mind.org.uk/information-support/types-of-mental-health-problems/stress/signs-and-symptoms-of-stress/*
- *https://www.mind.org.uk/information-support/types-of-mental-health-problems/depression/symptoms/*
- *https://www.mind.org.uk/information-support/types-of-mental-health-problems/anxiety-and-panic-attacks/symptoms/*

To assess criterion validity we could compare the results before and after some treatment. However, the study design doesn't cover this scenario. We can check the concurrent validity comparing the boxplots for total score according to values of the criterion variable criterion defined in the previous section. As we can see, the boxplots don't differ that much so we can't consider this variable as good validity check.

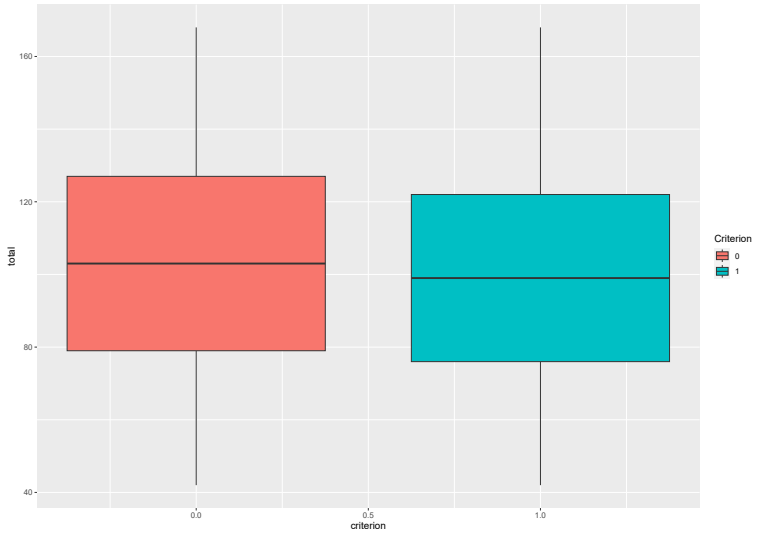

Figure 5: Boxplot of total score by criterion variable.

To check the convergent validity, we would need more data, e.g. whether some stress/anxiety/depression issues run in the family or whether the person has already struggled with some of the aforementioned problems. The discriminant validity was checked via correlations with family size, religion, English as native language and area where the person lived as a child. Total score wasn't correlated with none of the aforementioned variables, therefore shows good discriminant validity. These results and corresponding boxplots can be found in the R code. We don't have the data needed to obtain predictive validity. To asses incremental validity we could check whether the 42 DASS test provides additional information compared to 21 DASS test. To obtain evidence of the validity of test scores based on internal structure we first calculate and plot polychoric correlations between the items. The following figure shows positive or nearly zero correlations between every pair of the items.

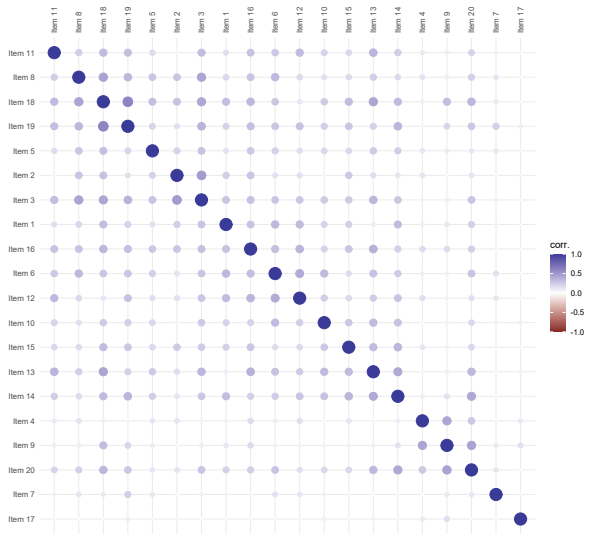

Figure 6: Correlation heat map of items on stress, depression and anxiety.

On the dendogram bellow, the y-axis is a measure of closeness of either individual data points or clusters. Based on the plot we get some basic idea, which items are closer to each other.

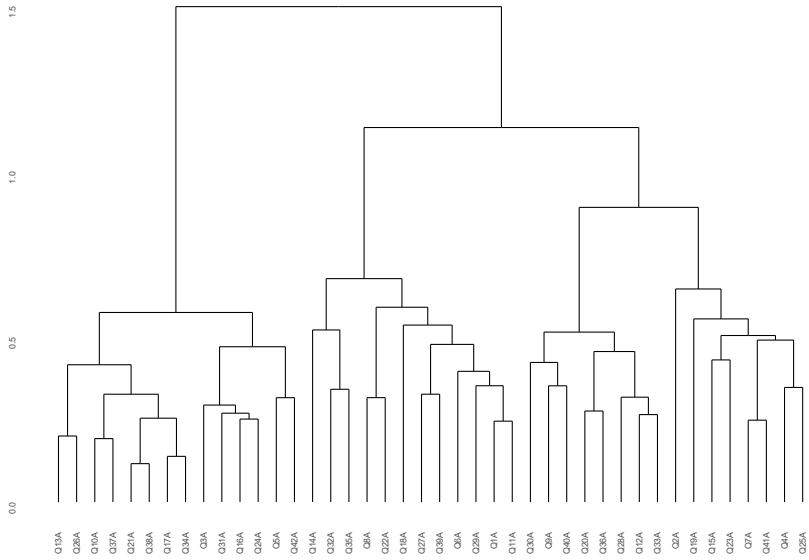

Figure 7: Dendogram of items on stress, depression and anxiety.

Better image on number of factors offers scree plot. Based on the graph and for better interpretability, we will choose 3 factors for the factor analysis. According to *Lovibond, S.H. & Lovibond, P.f. (1995). Manual for the Depression anxiety Stress Scales. (2nd Ed) Sydney: Psychology Foundation.* these factors should represent anxiety, depression and stress. Scores of Depression, Anxiety and Stress are calculated by summing the scores for the relevant items. The depression scale items are 3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42. The anxiety scale items are 2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41. The stress scale items are 1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39 (See attached document). From the exploratory factor analysis and the three bar-plots we would assign items 3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38 and 42 to the first factor and denote this factor as depression. We would assign items 1, 6, 8, 9, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39 and 40 stress to the second factor and this should correspond to stress and items 2, 4, 7, 15, 19, 20, 23, 25, 28, 30, 36 and 41 to the third one and denote it as anxiety. We can see, that the results from exploratory factor analysis differ only minimally. These factors were used in the confirmatory factor analysis. The path diagram graphically summarizes the fitted model from confirmatory factor analysis.

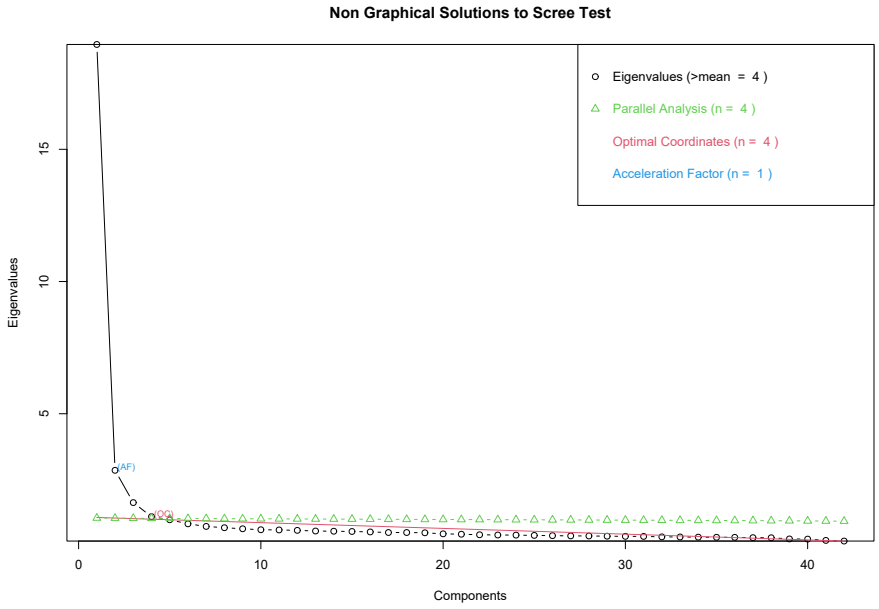

Figure 8: Scree plot on the DASS dataset.

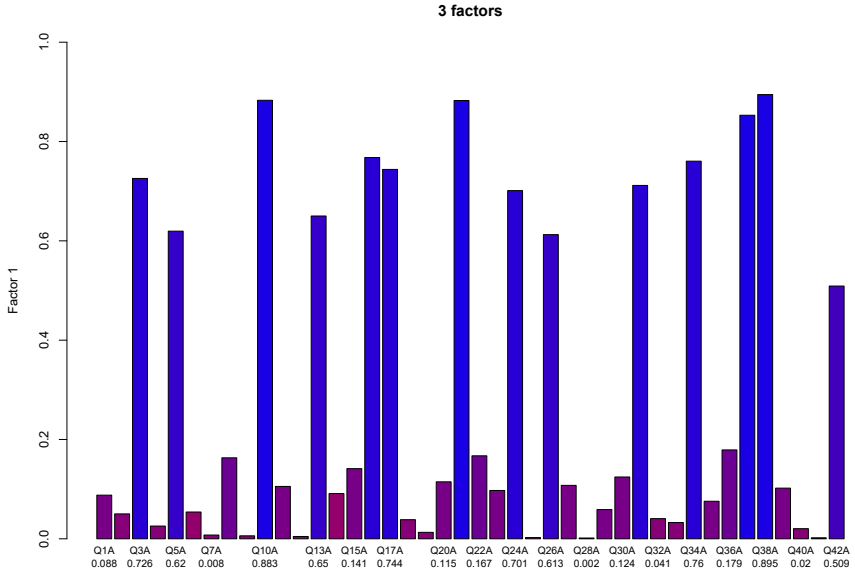

Figure 9: Bar-plot of correlations of items with factor 1.

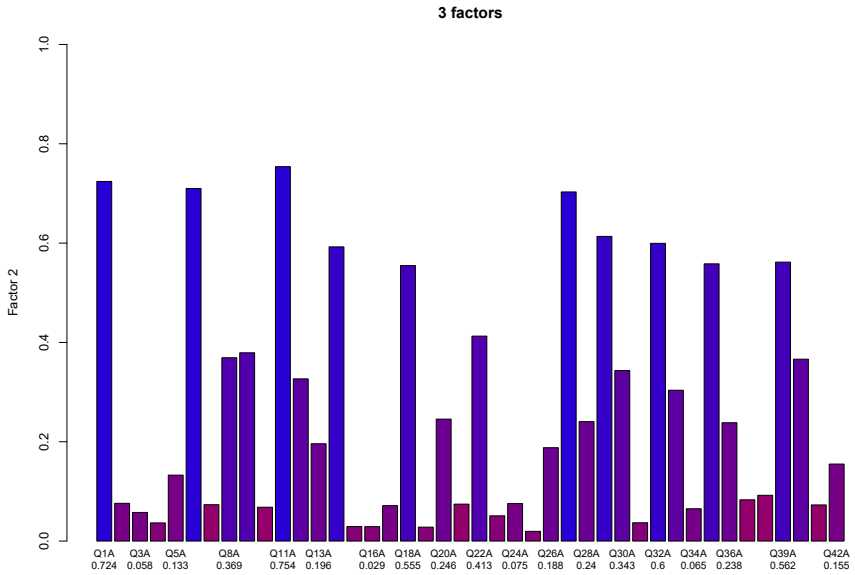

Figure 10: Bar-plot of correlations of items with factor 2.

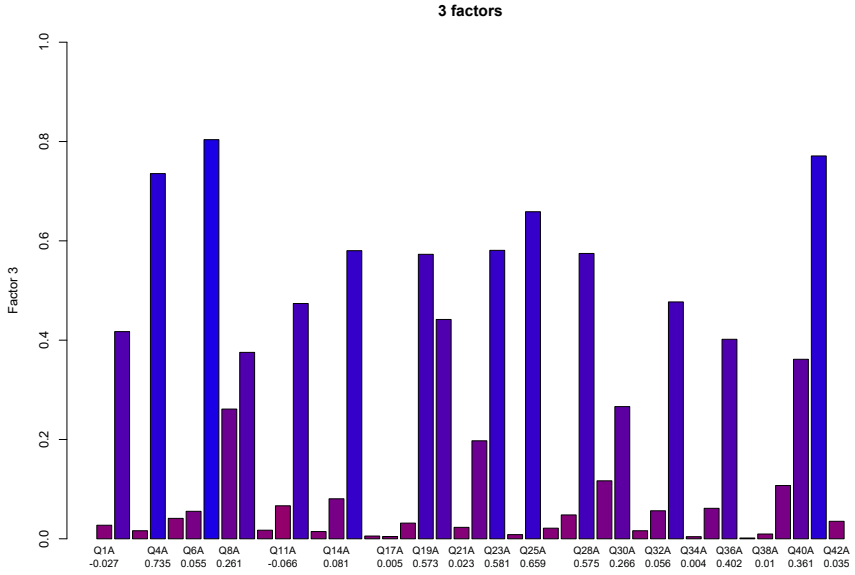

Figure 11: Bar-plot of correlations of items with factor 3.

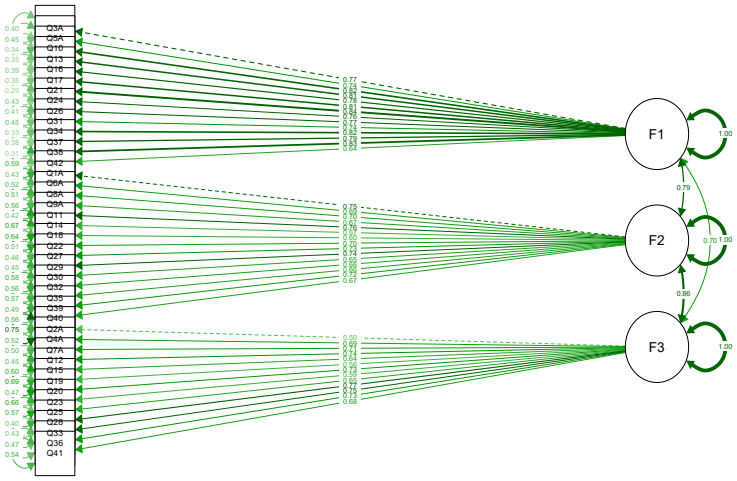

Figure 12: Path diagram of the three-factor CFA model for the Stress, Anxiety and Depression domains of the DASS dataset.

3 Test reliability

The dataset has a Crohnbach's alpha coefficient estimate of 0.9699 (95% CI (0.9695, 0.9703)). Based on Cicchetti's cut-off values, the reliability classifies as excellent. The high value is due to the large number of items and the fact that all items have just 4 possible outcomes. According to the Spearman-Brown formula, doubling the number of items would lead to Crohnbach's alpha estimate 0.985. The minimum number of items needed to get reliability of 0.9 is 39. Other reliability measures provide similar results, e.g., the first-second split-half coefficient is 0.9774, and the even-odd split-half coefficient is 0.9754, indicating strong internal consistency. In contrast, the reliability estimated using one-way ANOVA, which evaluates consistency by assessing the variance among group means rather than correlating test halves, is slightly lower at 0.9665. Additionally, the McDonald's omega estimate is 0.9705, which is a more suitable measure of internal consistency than Cronbach's alpha, as it takes into account the possible multidimensionality.

4 Item analysis

Based on Table 3, the most endorsed item is Q23, "I had trouble swallowing." This might be due to the fact that swallowing issues can also be related to other medical issues rather than solely to mental distress and might not be as frequently related to stress or anxiety as Q34, "I felt I was pretty worthless," which is one of the least endorsed items.

In the assessment of item discrimination, several items displayed notably lower discrimination indices, suggesting they might be less effective in differentiating between levels of the underlying latent variables. These items are item 2 "I was aware of dryness of my mouth," item 19 "I perspired noticeably (e.g., hands sweaty) in the absence of high temperatures or physical exertion," item 23 "I had difficulty in swallowing," item 25 "I experienced breathing difficulty (e.g., excessively rapid breathing, breathlessness in the absence of physical exertion)," and item 41 "I just couldn't seem to get going."

	dif.	RIR	RIT	ULI	alphaDrop		dif.	RIR	RIT	ULI	alphaDrop
Q1	2.62	0.69	0.71	0.42	0.97	Q22	2.34	0.68	0.69	0.40	0.97
Q2	2.17	0.46	0.49	0.30	0.97	Q23	1.56	0.52	0.54	0.25	0.97
Q3	2.23	0.70	0.72	0.42	0.97	Q24	2.44	0.69	0.71	0.42	0.97
Q4	1.95	0.60	0.62	0.36	0.97	Q25	2.18	0.56	0.59	0.35	0.97
Q5	2.52	0.69	0.71	0.43	0.97	Q26	2.66	0.72	0.74	0.45	0.97
Q6	2.54	0.62	0.65	0.38	0.97	Q27	2.61	0.68	0.69	0.41	0.97
Q7	1.92	0.60	0.63	0.36	0.97	Q28	2.22	0.69	0.71	0.43	0.97
Q8	2.48	0.69	0.70	0.42	0.97	Q29	2.65	0.69	0.71	0.43	0.97
Q9	2.67	0.64	0.66	0.40	0.97	Q30	2.39	0.64	0.66	0.40	0.97
Q10	2.45	0.70	0.72	0.47	0.97	Q31	2.38	0.67	0.69	0.40	0.97
Q11	2.80	0.70	0.72	0.43	0.97	Q32	2.45	0.61	0.63	0.36	0.97
Q12	2.43	0.68	0.69	0.42	0.97	Q33	2.41	0.69	0.71	0.42	0.97
Q13	2.78	0.75	0.77	0.48	0.97	Q34	2.63	0.73	0.74	0.50	0.97
Q14	2.58	0.51	0.54	0.32	0.97	Q35	2.30	0.61	0.63	0.35	0.97
Q15	1.83	0.58	0.60	0.32	0.97	Q36	2.27	0.70	0.72	0.45	0.97
Q16	2.52	0.69	0.71	0.45	0.97	Q37	2.37	0.67	0.69	0.45	0.97
Q17	2.66	0.72	0.74	0.49	0.97	Q38	2.39	0.71	0.73	0.50	0.97
Q18	2.48	0.55	0.57	0.34	0.97	Q39	2.45	0.68	0.69	0.40	0.97
Q19	1.95	0.49	0.52	0.30	0.97	Q40	2.65	0.64	0.66	0.42	0.97
Q20	2.32	0.68	0.70	0.45	0.97	Q41	1.97	0.57	0.60	0.34	0.97
Q21	2.35	0.72	0.74	0.50	0.97	Q42	2.68	0.61	0.63	0.37	0.97

Table 3: Item difficulty (dif.), RIR, RIT, ULI and alphaDrop for the 42 test items.

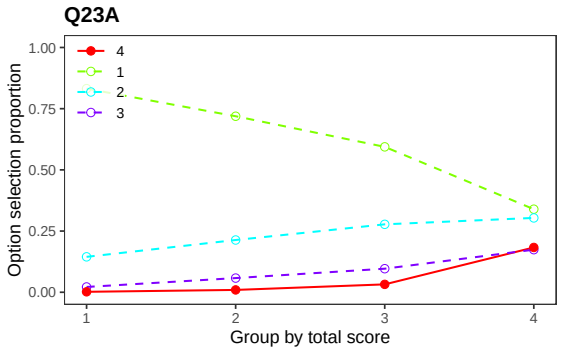

Figure 13: Empirical ICC for the item Q23.

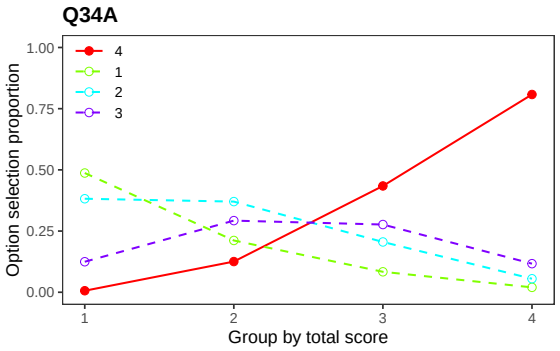

Figure 14: Empirical ICC for the item Q34.

To fit a regression model, for simplicity we redefine the items as binary (1 = 3, 4; 0 = 1, 2). Based on the two parameter regression model (by using IRT parametrization) we get estimates $a = 1.66$ and $b = 1.62$ for the item 34, and $a = 2.18$ and $b = -0.11$ for the item 23. That means, that the z-score needed to answer 3 or 4 with 50% probability (the difficulty of the item) is 1.62 for item 34, and -0.11 for item 23. The change of log odds associated with one-unit increase of the total score is 1.66 for question 34, 2.18 for question 23. Note that this model does not take into account the multidimensional structure of the data, rather it assumes some collective underlying latent trait "mental distress". The following models are more appropriate since they take into account the multidimensional structure.

5 Item response theory models

Now we focus on an appropriate IRT model that can model our multivariate ordinal items. As we have sufficient sample size and we are interested in modeling the ability level needed to provide a response category higher than a certain threshold, e.g. "Applied to me to some degree, or some of the time", we use the multidimensional graded response model, with an equation

$$\pi_{ik} = P(Y_{pi} \geq k|\boldsymbol{\theta_p}) = \frac{\exp(\beta_{0ik} + \boldsymbol{\beta_{1i}\theta_p})}{1 + \exp(\beta_{0ik} + \boldsymbol{\beta_{1i}\theta_p})}, \tag{1}$$

where $k = 2, \ldots, 4$, $i = 1, \ldots, 42$, and $p = 1, \ldots, 39775$, and $\pi_{i1} = 1$.

The coefficients for the first 10 items can be found in Table 4, the estimation was done using Marginal Maximum Likelihood estimation method.

	β_{1i1}	β_{1i2}	β_{1i3}	β_{0i2}	β_{0i3}	β_{0i4}
Q1	-1.51	0.81	1.86	3.19	0.10	-1.99
Q2	-0.78	0.68	0.18	0.71	-0.76	-1.81
Q3	-2.28	-0.15	0.90	1.69	-1.18	-3.11
Q4	-1.46	1.44	0.09	0.36	-1.64	-3.21
Q5	-1.98	0.06	0.77	2.34	-0.21	-1.97
Q6	-1.08	0.90	1.44	2.42	-0.09	-1.85
Q7	-1.60	1.68	0.02	0.33	-1.89	-3.51
Q8	-1.49	0.89	0.86	2.17	-0.30	-2.00
Q9	-1.29	1.19	0.71	2.54	0.23	-1.42
Q10	-2.80	-0.51	0.82	2.20	-0.43	-2.37

Table 4: Coefficients of the graded response model (3) for items Q1 – Q10.

The vectors $\boldsymbol{\beta_{1i}}$ represent the item-specific slopes and β_{0ik} are ordered category-specific intercepts.

For better visualization, Rasch-type models are of interest. An important assumption of this class of models is unidimensionality. As for the three tested latent traits - depression, anxiety, and stress — the evaluation is conducted separately using 14 different questions for each trait, and also the confirmatory factor analysis is aligned with the division of questions into three such subsets, we consider three separate Rasch models. Since we have ordinal data, Partial Credit Risk model will be used. The model equation is

$$\log\left(\frac{\pi_{pik}}{\pi_{pi(k-1)}}\right) = \theta_p - b_{ik}, \quad k = 2, \ldots, 4. \tag{2}$$

The following analysis was conducted for the latent trait depression. The Figure 20 is the Wright (person-item) map. We can see the depression latent trait distribution as well as the item thresholds of adjacent category locations, i.e. the points where the probability of responding in either of the two adjacent categories is the same. Items with higher thresholds are considered more difficult because they require a higher level of the latent trait (depression) being measured to endorse a higher response category. Conversely, items with lower thresholds are considered easier because they require a lower level to endorse a higher response category. The most endorsed seems item 13 "I felt sad and depressed."

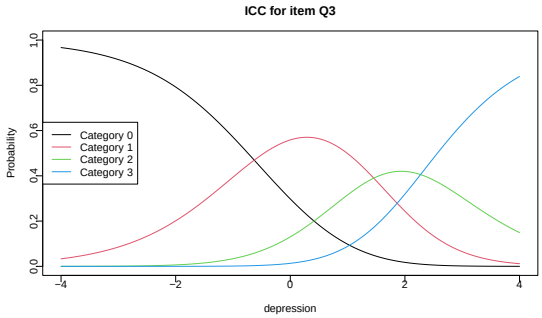

Figure 16: Item characteristic plot for the item Q3. The categories are indexed lowest to highest.

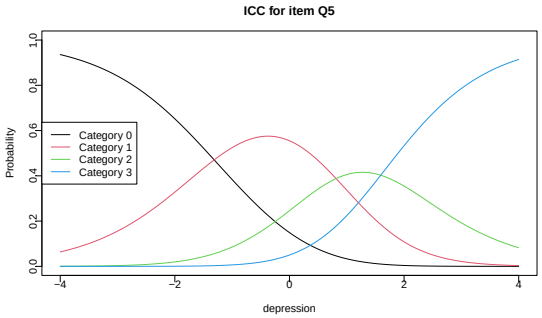

Figure 17: Item characteristic plot for the item Q5.

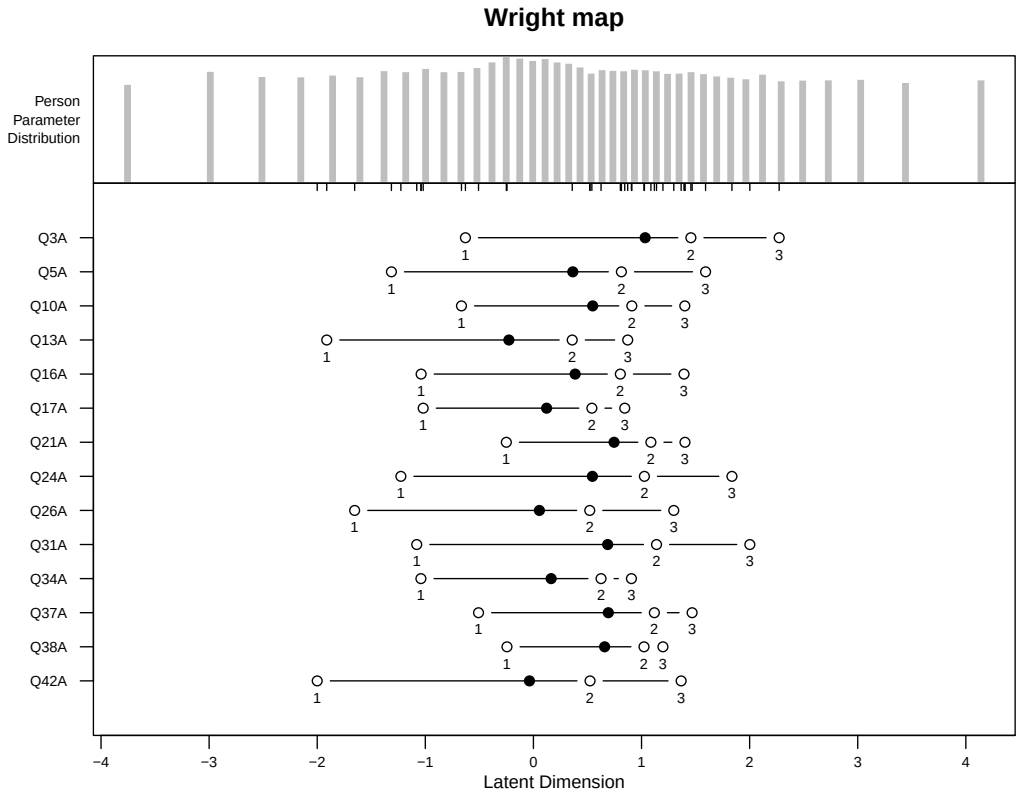

Figure 15: Wright map.

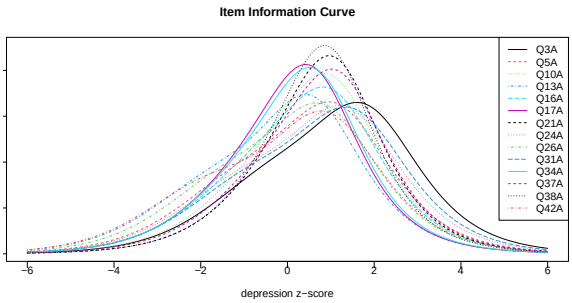

Figure 18: Item Information Curve plot.

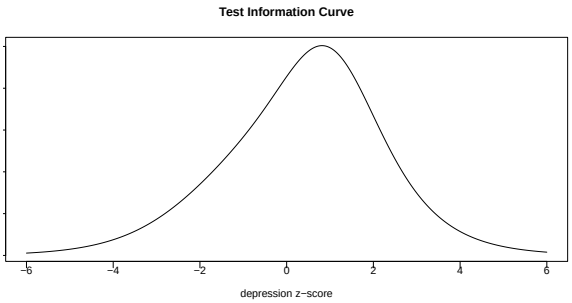

Figure 19: Test Information Curve plot.

Item	Threshold 1	SE	Threshold 2	SE	Threshold 3	SE
Q3A	-0.63	0.01	1.46	0.02	2.27	0.02
Q5A	-1.31	0.02	0.81	0.02	1.59	0.02
Q10A	-0.67	0.02	0.91	0.02	1.40	0.02
Q13A	-1.91	0.02	0.36	0.02	0.87	0.02
Q16A	-1.04	0.02	0.80	0.02	1.39	0.02
Q17A	-1.02	0.02	0.54	0.02	0.84	0.02
Q21A	-0.25	0.02	1.09	0.02	1.40	0.02
Q24A	-1.23	0.02	1.03	0.02	1.84	0.02
Q26A	-1.65	0.02	0.52	0.02	1.30	0.02
Q31A	-1.08	0.02	1.14	0.02	2.00	0.02
Q34A	-1.04	0.02	0.63	0.02	0.91	0.02
Q37A	-0.51	0.02	1.12	0.02	1.47	0.02
Q38A	-0.24	0.02	1.02	0.02	1.20	0.02
Q42A	-2.00	0.02	0.52	0.01	1.37	0.02

Table 5: Thresholds and their standard errors for each item

By using information scores (or Item Information Curves), the most informative for average ability level and ability level 1SD above the average has item Q17, and ability level 1SD above the average has item Q38.

The first respondent has ability estimate 1.14, with standard error 0.32 and 95% confidence interval (0.51, 1.77).

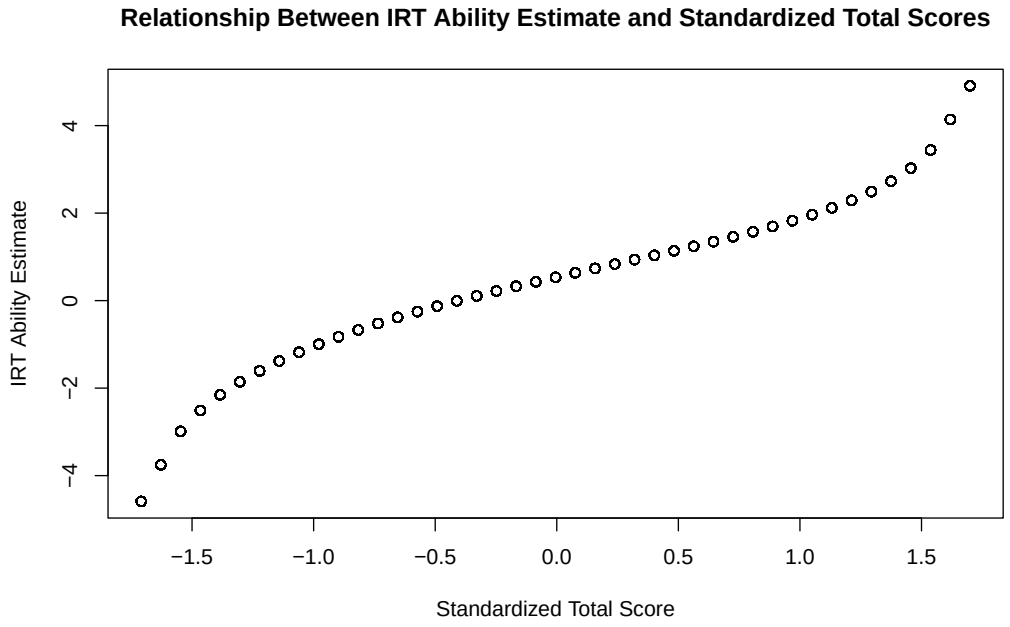

Figure 20: The relationship between ability estimates in IRT models and traditional ability estimates based on (standardized) total scores.

6 Differential item functioning

Differential Item Functioning (DIF) analysis is essential in test validation for ensuring fairness, validity, and reliability across diverse socio-economic groups. For the DASS-42, DIF analysis plays a crucial role in determining whether each item on the test performs equally well across different subgroups, such as age, gender, and cultural backgrounds. This is vital to ensure that the test accurately assesses these psychological states without bias.

We are examining whether gender is the source of DIF. Our data are ordinal, multidimensional and

we have very large sample size. Recall that for the three tested latent traits - depression, anxiety, and stress — the evaluation is conducted separately using 14 different questions for each trait. Consequently, for simplicity, the DIF analysis will be performed on each subset of questions independently. The analysis was again conducted using the cumulative graded response model, due to the ordinal item structure, with the equation:

$$\pi_{ik} = P(Y_{pi} \geq k|\theta_p, G_p) = \frac{\exp(\beta_{0ik} + \beta_{1i}\theta_p + \beta_{1i2}G_p + \beta_{3i}\theta_p G_p)}{1 + \exp(\beta_{0ik} + \beta_{1i}\theta_p + \beta_{1i2}G_p + \beta_{3i}\theta_p G_p)}, \tag{3}$$

where $k = 2, \ldots, 4$, $i = 1, \ldots, 14$, and $p = 1, \ldots, 39755$, and $\pi_{i1} = 1$. The variable θ_p corresponds to the underlying latent trait and is commonly replaced by the total score.

Almost all 42 items were marked as DIF items, which is not surprising given the very large sample size of over 39 000 respondents. DIF detection methods are known for Type I error inflation, where an increasing sample size — and therefore increasing power — leads to even very small differences being identified as statistically significant for large sample sizes. Using the McFadden's pseudo-R^2 measure, we get the maximum item's Δ pseudo-R^2, 0.003 for depression, 0.002 for anxiety and 0.006 for stress. That suggest that the relative improvement in the log-likelihood of model with gender vs. submodel without gender is for all items rather small, consequently it appears that DIF is not a large concern for any item. To confirm, a random subset of 500 respondents, a number comparable with similar published studies, was chosen and the same analysis was conducted again. The second identified no DIF items, the first model identified question 26 "I felt down and blue" and question 34 "I felt I was pretty worthless" as possible DIF items for the depression latent trait, with Δ pseudo-R^2 of 0.01 and 0.02, respectively, which again does not seem like a large model performance improvement. The last model identified item 11 "I found myself getting upset rather easily" as possible DIF item, with pseudo-ΔR^2 of 0.01. Since the DASS-42 test is well-established test in psychometrics which was carefully build and extensively studied for the last few decades, it is unsurprising that no items seem to be exhibiting DIF towards gender.

In the context of psychological test, it might be appropriate to consider also other grouping variables, such as age. As the dataset contains many grouping variables, it might be interesting to analyze DIF with respect to some rarely collected variables, such as marital status, or e.g. major at university, since the understanding and background in psychology and behavioural sciences could lead to different understanding of the questions.

7 Discussion

The dataset is rather complex, as the data are ordinal and multidimensional. There are three separate outcomes of the test, separate for depression, anxiety, and stress. That was used on several occasions in the project for clarity and simplicity of the methods. That might be a possible limitation, as these analysis do not use all the available information utilizing the correlations as multidimensional techniques do. The results are in general expected, we did not flag any items as possible DIF items based on gender, all items seem to be contributing and the validity and reliability results seem promising. That is to be expected as this test is well-established.

Another limitation is the number of respondents for DIF analysis, since for the whole number of respondents the results seem to be unusable. A possible solution is to restrict the sample size by taking a random subset, however, the issue is how to choose the sample size magnitude.

A possible limitation is also the data collection. The population consists of people interested in their mental health or already being aware of some issues, who voluntarily agreed to take a questionnaire after filling the DASS-42 test. Another results might be obtained on other populations, e.g. patients already being treated or having been treated e.g. for clinical depression and analyzing such group could be interesting data to be collected for further analyses.

8 Supplement A - Commented R code

The code includes all data manipulation, plots and results for discriminant validity (which are mentioned in the relevant chapter), functions and outputs for both exploratory and confirmatory analysis. All the computations for reliability section (such as Cronbach's alpha, Split-half coefficient, ANOVA or McDonald's Omega) are included as well as all computations for traditional item analysis, IRT models and DIF.

9 Supplement B - ShinyItemAnalysis report and datasets

Attached is a report generated from ShinyItemAnanlysis app. 2PL model was chosen as IRT model. There are several limitations to using ShinyItemAnalysis on our dataset, as we have multidimensional and ordinal data. The data were binarized (unlike in our analysis when they were left as ordinal) and treated as unidimensonal with one assumed underlying trait. As in our analysis, we might upload three separate datasets to get comparable IRT models and DIF detection, however, that would mean potential loss for the validity and reliability analysis.

Another issue is that the variable gender is not binary, but includes also options "other" and "empty". In our analysis, we were able to use all available observations and limit the range just for the DIF analysis by comparing males and females. ShinyItemAnalysis requires doing all the analysis on the same (smaller, n = 39156) range of data. The percentage difference in sample size is minimal and in theory should therefore play no role in the results comparison. In practice, however, we weren't able to run the ShinyItemAnalysis on the subset of respondents, for unknown reasons. Therefore we were forced to enclose version with all the respondents and leave out the DIF.

10 Additional analyses

Above what was required we included McDonald's omega coefficient, which is based on a factor analytic approach, in contrast to alpha, which is primarily based on the correlation between the questions. It is widely recognised as a more accurate estimate of reliability. It makes fewer assumptions; it does not assume that items are τ-equivalent or parallel. In addition, omega total is in theory an upper bound for alpha – alpha is expected to be less than omega total unless items are τ-equivalent, in which case alpha equals omega total.