







ARTICLE

A psychometric analysis of the clinical screening scale¹

 Brunna Quatrochi *¹  Gabriela dos Santos Amaral¹  Jorge Luis Bazán²  Emanuela Pap da Silva³

¹ Institute of Mathematical and Computer Sciences, University of Sao Paulo, São Carlos, SP, Brazil

² Department of Applied Mathematics and Statistics, University of Sao Paulo, São Carlos, SP, Brazil

³ University of Sao Paulo, São Carlos, SP, Brazil

* Corresponding author. Email: brunnaquatrochi@usp.br.

(Received: May 6, 2025; Revised: July 18, 2025; Accepted: August 1, 2025; Published: December 18, 2025)

Abstract

A psychometric analysis of the SQR-20 screening scale is proposed, composed of twenty dichotomous items that assess indicators of common mental disorders (CMD). Databases from the years 2021, 2022, and 2023 were used, obtained through questionnaires applied by GAPSI and Apoia USP, answered by undergraduate students from USP São Carlos. The results indicate that the test is reliable and unidimensional, and CMD scores can be obtained both through classical test theory and the two-parameter model of item response theory (IRT). Considering IRT, the results showed that the two-parameter model was the most suitable to represent the responses, highlighting that questions related to feelings of sadness and nervousness are more discriminating than physical questions. More severe items, such as suicidal thoughts, are concerning but less effective for discrimination. The comparative analysis between sociodemographic variables, such as age, gender, and course, revealed a significant impact on the well-being of students. Additionally, the weighting of responses highlighted the importance of the course in determining the difficulties faced by students. With these results, it is expected to propose more effective interventions for promoting mental health and well-being in the academic environment.

Keywords: Psychometric Analysis; SRQ-20; Clinical Analysis; University students

1. Introduction

The area of Psychometrics, a field encompassing the statistical methods used in Psychological and educational testing, has become a very important and active area of research, evident from the large body of literature that has been developed in the form of books, volumes and research papers. Mainstream statisticians also have found profound interest in the field because of its unique nature.

The field of psychometrics, which encompasses the statistical methodologies underpinning psychological and educational testing, has emerged as a vital and dynamic area of research within the statistical community. See Rao and Sinharay (2007) for details.

This prominence is reflected in the extensive body of literature that has been produced, including seminal books, edited volumes, and high-impact research papers not only on the psychometric journal, but also in the important statistical journals (see for example, Bazan *et al.* (2006), Azevedo *et al.* (2011), da Silva *et al.* (2019), de Oliveira *et al.* (2023) and Fernandes *et al.* (2024).

The unique challenges posed by psychometric research — such as modeling latent constructs, addressing measurement error, and ensuring fairness and validity in assessments — have attracted

significant attention from mainstream statisticians. These challenges often require the development and application of sophisticated statistical techniques, such as item response theory (IRT), factor analysis, structural equation modeling (SEM) and Bayesian hierarchical models, making psychometrics a rich and fertile ground for advancing statistical theory and methodology. As a result, the intersection of psychometrics and statistics continues to foster interdisciplinary collaboration and innovation, further solidifying its importance in both theoretical and applied domains.

In this work, we develop a psychometric analysis based on statistical analysis, which will be applied to an instrument used to screen for mental disorders, considering data collected annually using the SRQ-20 questionnaire applied to undergraduate students by Apoiá USP and GAPsi. The analysis includes the application of Item Response Theory (IRT) models to measure the effectiveness of these items in identifying mental disorders.

From a statistical perspective, the aim is to promote and expand statistical, mathematical, and data science methods in applied biological sciences, demonstrating how these models can be extended to other similar cases.

Presenting the methodology of psychometric analysis to the statistical community through a review article is an effective strategy to foster integration between psychometrics and statistics, providing a solid foundation for new applications and interdisciplinary collaborations. This presentation can stimulate the development of partnerships between statisticians and psychometricians, creating theoretical and practical bridges between the two areas. As a result, new approaches to common problems emerge, and existing models are improved, benefiting both statistical and behavioral sciences. Moreover, the application of advanced statistical models can enhance the quality and reliability of psychometric instruments used in psychology, mental health, and education. Statisticians can collaborate in validating these instruments, ensuring their effectiveness in identifying disorders and behaviors, contributing to the development of specific statistical methods adapted to the needs of psychology and health.

Furthermore, the statistical community can play a crucial role in ensuring that psychometric analyses are conducted robustly and rigorously, increasing the credibility and acceptance of the results. This is particularly relevant in sensitive areas, such as assessing mental distress, where the results can directly impact people's lives.

By involving statisticians, it is also possible to promote the use of advanced data analysis technologies, such as software for analyzing Item Response Theory (IRT) and Classical Test Theory (CTT). This can increase efficiency in applying psychometric models, facilitating the use of statistical resources on a large scale and expanding their reach of application.

Finally, from an applied perspective, psychometric analysis is a powerful tool for screening mental health issues in various contexts, such as schools, universities and clinical environments. Collaboration with the statistical community can further enhance these tools, making them more effective and capable of generating applicable results, contributing to more assertive and evidence-based public policies.

By presenting psychometric analysis to the statistical community, not only is a better understanding of the techniques used promoted, but new opportunities for the evolution of these methodologies are also opened. This creates a cycle of continuous innovation, benefiting both statistics and behavioral sciences and expanding their applications in multidisciplinary contexts.

This version seeks to make the text more direct, fluid and easy to read, focusing on the reasons for dissemination and interaction between the fields of psychometrics and statistics. This article is structured as follows: the characteristics of the test used for the application is discussed in Section 2.

In Section 3, we present the methodology of the psychometric analysis applied to mental health questionnaires, while in Section 4, we discuss the results obtained, including the application of measurement models, the choice of the most appropriate model, and the estimation of latent traits, followed by comparative analyses with sociodemographic variables.

2. Measuring the mental distress screening

As stated by Bazan (2018), to measure a person's ability, it is necessary to establish a measurement scale, that is, a rule with a specific metric. This rule should be applied to quantify a person's capacity in a certain area. The usual procedure involves defining a measure of ability and developing a test, which generally consists of a set of items based on specific questions. For example, we define "performance in reading written texts" as the "ability" needed to answer questions associated with reading.

In the context of mental distress screening, as opposed to psychiatric diagnosis (which must be performed exclusively by a professional), questionnaires are used to measure distress among university students, analogous to educational assessment. Questions are formulated by experts, and psychometric instruments with different properties are adopted (see Pasquali & Primi, 2003).

Among these psychometric properties, the statistical analysis of items, the evaluation of the questionnaire's dimensionality, and the use of measurement models stands out, considering that the questions serve as an indirect measure of an unobservable latent variable. The goal is to generate scores of this variable for subsequent analyses.

According to Bazan (2018), the measurement model is used to associate the observed variables (responses to items) with the latent or unobserved variables corresponding to the construct being measured. There are two main models for this relationship and obtaining scores from responses: the classical test model and the item response model. Although other models exist, these are the most consolidated.

The classical test model, also known as Classical Test Theory (CTT), is a psychometric approach aimed at predicting test responses, such as item difficulty or respondent ability. Its primary objective is to understand and improve the reliability of tests (Nunnally & Bernstein, 1994). CTT is based on three central concepts: the existence of measurement errors, the idea that error is a random variable, and the notion that, through correlations, it is possible to estimate a score representing the true score based on item responses. Thus, it is possible to obtain the score as an estimate of the true score based on the total score or sum of responses to the test items or questions.

On the other hand, item response models, known as Item Response Theory (IRT), focus on understanding whether the examinee chooses a specific item (correctly or not), rather than worrying about the total score. IRT (Andrade *et al.*, 2020) describes how a person's latent trait — what is being measured — and item characteristics (item parameters) are related to observed responses through a probabilistic model. In this case, the test score corresponds to the estimate of the latent trait, obtained by inferential statistical procedures, which also provide estimates of item parameters.

Both CTT and IRT are complementary and widely used, as in the case of the National High School Exam (ENEM). Currently, both models are mainly applied in measuring cognitive abilities.

However, in the field of psychological assessment, the use of these models is less known, especially within the statistical community. The main objective of this work is to present a statistical analysis based on both measurement models in the context of clinical test evaluation. The complete analysis, in this case, is called psychometric analysis and involves several steps before applying the

measurement model, in addition to the subsequent use of the obtained scores for further analyses.

To illustrate this methodology, we will analyze data collected through the test called Self-Report Questionnaire (SRQ-20), a screening instrument composed of 20 dichotomous items ("yes" or "no"). Developed by the World Health Organization (WHO), the SRQ-20 aims to identify indicators of common mental disorders, especially in primary care contexts.

The promotion of mental health in the university environment has become a growing concern, especially due to the increase in stress, anxiety, and other emotional difficulties among students. Factors such as academic workload, the transition to university life, and social pressures can negatively impact student well-being, affecting both academic performance and personal quality of life. In this scenario, Apoia USP and the Psychopedagogical Support Group (GAPsi) of the Institute of Mathematical and Computer Sciences (ICMC) at the University of São Paulo (USP), São Carlos campus, have played a fundamental role in supporting students.

3. Materials and Methods

The methodology follows the steps of developing and validating psychometric instruments, focusing on the application of statistical models, namely Classical Test Theory (CTT) and Item Response Theory (IRT).

3.1 Data

The dataset encompasses 823 students and contains 20 dichotomous variables ("yes" and "no") related to the SRQ-20 items. This information was obtained through a form applied to undergraduate students. The SRQ-20 is a self-applicable questionnaire composed of 20 items used to screen for symptoms of common mental disorders, such as depression, anxiety, and stress.

- I01: Do you often have headaches?
- I02: Do you have a lack of appetite?
- I03: Do you sleep poorly?
- I04: Do you get startled easily?
- I05: Do you have trembling hands?
- I06: Do you feel nervous, tense, or worried?
- I07: Do you have indigestion?
- I08: Do you have difficulty thinking clearly?
- I09: Have you been feeling sad lately?
- I10: Have you been crying more than usual?
- I11: Do you find it difficult to perform your daily activities satisfactorily?
- I12: Do you have difficulty making decisions?
- I13: Do you have difficulty at work (is your work burdensome, causing you suffering)?
- I14: Are you unable to play a useful role in your life?
- I15: Have you lost interest in things?
- I16: Do you feel like a useless person, without value?
- I17: Have you had any thoughts of ending your life?
- I18: Do you feel tired all the time?
- I19: Do you tire easily?
- I20: Do you have unpleasant sensations in your stomach?

The sample consists of undergraduate students from the University of São Paulo (USP), São Carlos

campus, who responded to the annual questionnaire from the Psychopedagogical Support Group (GAPsi). The sample included students from different courses, age groups, genders, and years of study, ensuring the diversity necessary for a robust analysis. The data were anonymized to ensure the confidentiality of the respondents.

The SRQ-20 was chosen because it is a validated and widely used instrument in screening for emotional symptoms. The questionnaire items cover dimensions such as somatic symptoms, depressive symptoms, and anxiety symptoms, with binary responses (yes/no). The total score is calculated by summing the affirmative responses, and the higher the score, the greater the probability that the respondent presents symptoms of psychic distress (Silveira *et al.*, 2022).

3.2 Procedures for Psychometric Analysis

3.2.1 Item Classical Analysis, Reliability: Classical Test Theory (CTT)

In this study, we initially developed a classical item analysis with the main statistics for the questionnaire items. See for example Pettersson & Turkheimer (2010). Additionally, we obtained Cronbach's Alpha estimates of the questionnaire, commonly used to analyze its reliability and the contribution of the questions to it.

Classical item analysis is a method for evaluating the quality of test items using Classical Test Theory (CTT). It is used to identify effective, ambiguous, or misleading items and to improve the overall quality of a test. This analysis is a simple and intuitive approach that uses descriptive statistics to evaluate the effectiveness of each item, focusing on the mean and standard deviation. More details in Nunnally & Bernstein (1994).

Some researchers identified items with values less than 0.30 as difficult items and those greater than 0.70 as easy items. An item that is too easy and items that are too difficult may need to be reviewed to achieve a valid test. A good test should have a variety of items, from easy, moderate to difficult items according to Wright and Stone (1979).

In the case of dichotomous or binary items, the greatest variability is obtained when the mean is 0.5. In this case, the variation is 0.25 and the standard deviation is 0.5. Greater variability is desirable so that the item can capture a wider range of responses.

In this study, we used Ordinal Cronbach's Alpha, which, unlike traditional Cronbach's Alpha, employs the tetrachoric correlation matrix instead of the Pearson covariance matrix. This approach was chosen because the calculation of conventional Cronbach's Alpha assumes that the data are continuous, which does not correspond to our case since the data are dichotomous. When this assumption is violated, as occurs with categorical data, the Pearson covariance matrix may present distortions, compromising the estimate of the alpha coefficient (Gadermann *et al.*, 2012).

Cronbach's alpha is a coefficient that signals higher reliability of scores obtained from a set of items when closer to 1. Since the questionnaire data are based on ordinal response scales, as demonstrated by Zumbo *et al.* (2007), ordinal alpha is more appropriate than traditional Cronbach's Alpha. Nunnally & Bernstein (1994) suggest that a minimum value of 0.8 is recommended in applied research. Additionally, the correlation of the item with the rest of the questions is provided, as well as the Cronbach's alpha of the questionnaire excluding a specific item to determine if removing the item improves the questionnaire. The goal is to create tests with the fewest number of items that still provide the desired level of reliability.

3.2.2 Dimensionality Assessment

Dimensionality analysis is a process that seeks to determine whether a set of test items measures a

single characteristic, also known as one-dimensionality. One-dimensionality refers to the idea that students' mental health is influenced by a single symptom or latent trait.

Multiple factors in a test do not rule out unidimensionality. Carmines & Zeller (1979) argue that a test is unidimensional if the first factor explains at least 40% of the variance. Orlando *et al.* (2000) suggest that even with multiple factors, if over 80% of questions have factor loadings above 0.35, the test can still be considered unidimensional and suitable for item response theory analysis. These criteria are essential for evaluating test unidimensionality.

Thus, to assess one-dimensionality, we will apply the factor analysis technique (Revelle, 2022) considering the items that use the tetrachoric correlation matrix and consider the following specific criteria to determine one-dimensionality: a) the first factor explains at least 40% of the variance and b) the percentage of items with factor loadings above 0.35 on the first factor is above 80%.

3.2.3 Response Analysis: Item Response Theory

Item Response Theory (IRT) will be applied to provide a more detailed analysis of item performance and the latent ability of respondents. IRT will allow modeling the probability of an affirmative response to an item as a function of the student's level of mental distress. For this analysis the models with 1, 2, and 3 parameters (See Andrade *et al.*, 2000, Pasquali & Primi, 2003 and Bazan *et al.*, 2010) will be applied. The more general model considers the following item parameters:

- Difficulty parameter (b): Indicates the level of mental distress necessary to respond affirmatively to the item.
- Discrimination parameter (a): Measures the item's ability to differentiate between individuals with different levels of mental distress.
- Guessing parameter (c): Refers to the probability of a respondent marking an item.

IRT will be used to identify which items of the SRQ-20 have the highest discriminative power and what levels of difficulty are associated with each one from the chosen model. Based on this model, latent traits will be estimated, which are underlying characteristics of a person not directly observable, such as abilities or attitudes ('*escoretri*'), and we will test their normality to conduct comparison tests with sociodemographic variables.

3.2.4 Comparative Analysis: Sociodemographic Variables and Mental Health

The final stage of the methodology involves a comparative analysis between the score obtained from the latent traits and sociodemographic variables, such as gender, age, and course, to explore possible associations between these factors and questions about mental distress. Non-parametric statistical tests Kruskal-Wallis and Wilcoxon will be used to investigate whether there are significant differences between groups by considering weight that adjust the proportions of respondents to reflect the expected population distribution, as Lumley & Scott (2013) did to develop tests comparing two samples under complex sampling.

To the implementation of the methodology considering the steps 3.2.1 and 3.2.3 described above could be consulted Amaral *et al.* (2025).

4. Results and Discussion

The methodology follows the steps of developing and validating psychometric instruments, focusing on the application of statistical models, namely Classical Test Theory (CTT) and Item Response Theory (IRT). To reproduce the results could be consulted Amaral *et al.* (2025).

4.1 Classical analysis

The following table presents a detailed analysis of the SQR-20 items, including descriptive statistics and reliability indicators. In the first section, measures such as mean, standard deviation (SD), and coefficient of variation (CV) are displayed, providing insights into the dispersion and variability of participants' responses. The second section presents item reliability statistics if it is removed, including the adjusted Cronbach's alpha coefficient, the average correlation between items and the corrected item-total correlation (r.drop). These data are fundamental to assessing the internal consistency of the instrument and identifying items that may affect its overall reliability.

As proposed, Cronbach's ordinal alpha will be calculated using the tetrachoric correlation matrix (Zumbo *et al.*, 2007). For dichotomous data, it is possible to employ the polychoric or tetrachoric functions to obtain this matrix and, from it, calculate the usual Cronbach's alpha. In this case, the coefficient obtained considering all items was 0.94.

Initially, we performed a traditional classical item analysis with the main statistical metrics: the mean, standard deviation (SD), and coefficient of variability (CV) for each item. See Table 1.

Table 1. Descriptive and item reliability statistics of the SQR-20 items in a sample of 823 students

Item	Descriptive Statistics			Item Omitted Statistics		
	Mean	SD	CV (%)	Raw_Alpha	Average_r	r.drop
I01	0.43	0.5	114.32	0.94	0.43	0.53
I02	0.35	0.48	137.85	0.94	0.43	0.53
I03	0.61	0.49	80.42	0.94	0.43	0.55
I04	0.41	0.49	119.56	0.94	0.44	0.45
I05	0.3	0.46	152.8	0.94	0.44	0.45
I06	0.86	0.35	40.12	0.93	0.42	0.75
I07	0.35	0.48	136.38	0.93	0.43	0.58
I08	0.53	0.5	94.5	0.93	0.42	0.74
I09	0.63	0.48	75.98	0.93	0.42	0.78
I10	0.32	0.47	147.24	0.93	0.43	0.59
I11	0.72	0.45	62.13	0.93	0.42	0.71
I12	0.65	0.48	72.63	0.93	0.43	0.57
I13	0.41	0.49	120.16	0.93	0.42	0.69
I14	0.33	0.47	141.64	0.93	0.42	0.65
I15	0.61	0.49	80.83	0.93	0.42	0.68
I16	0.43	0.5	115.75	0.93	0.42	0.7
I17	0.19	0.39	208.55	0.93	0.42	0.68
I18	0.68	0.47	69.15	0.93	0.42	0.67
I19	0.72	0.45	62.13	0.93	0.42	0.69
I20	0.42	0.49	118.67	0.93	0.43	0.59

Upon analyzing the results of the questions (Table 1) with the highest means, such as "Do you feel nervous, tense, or worried?" (item I06), "Do you find it difficult to perform your daily activities satisfactorily?" (item I11), "Do you have difficulty making decisions?" (item I12), and "Do you tire easily?" (item I19), it is evident that these symptoms are present in a significant proportion of the research participants. This suggests a notable prevalence of anxiety, difficulties in daily functioning, and fatigue in the studied sample, pointing to potential mental health issues. On the other hand, when considering the questions with the lowest means, such as "Do you have trembling hands?" (item I05) and "Have you had any thoughts of ending your life?" (item I17), we observe a relatively lower incidence of these symptoms in the sample. However, it is important to note that even these less prevalent symptoms should not be underestimated, as they may indicate serious physical and mental health problems, such as neurological disorders and suicidal ideation.

The standard deviations revealed considerable heterogeneity in the participants' reports, especially in questions like "Do you sleep poorly?" (item I03), "Do you feel nervous, tense, or worried?" (item I06), "Have you been feeling sad lately?" (item I09), "Do you find it difficult to perform your daily activities satisfactorily?" (item I11), "Do you have difficulty making decisions?" (item I12), and "Do you tire easily?" (item I19). This suggests that although these symptoms are common in the sample, the severity and frequency vary considerably among individuals. This diversity of experiences highlights the importance of personalized approaches in the diagnosis and treatment of mental health problems.

Next, we will consider the *raw_alpha*, *average_r*, and *r.drop* statistics, which will be detailed throughout the analysis. See Table 1.

First, we will examine the Cronbach's alpha of the questionnaire with the exclusion of specific items, observing whether the alpha value increases, remains the same, or decreases. This value is reported as *raw_alpha*. When items such as "Do you often have headaches?" (I01), "Do you have a lack of appetite?" (I02), "Do you sleep poorly?" (I03), "Do you get startled easily?" (I04), and "Do you have trembling hands?" (I05) are excluded, the alpha value remains at 0.94, indicating that the test's reliability is not affected. Although these items do not improve reliability, they are retained due to their relevance to psychological assessment.

In contrast, the exclusion of other items, such as "Have you been feeling sad lately?" (I09), "Have you been crying more than usual?" (I10), "Do you feel like a useless person?" (I16), and "Have you had any thoughts of ending your life?" (I17), reduces the alpha from 0.94 to 0.93. These items are essential for the study and therefore must be retained.

Next, we analyzed the *average_r statistic*, which represents the average of the correlations between an item and the others. All items presented values above 0.42, indicating a significant relationship between them.

Finally, we considered *r.drop*, which measures the correlation of an item with the total score without it. The items with the lowest correlations were "Do you get startled easily?" (I04) and "Do you have trembling hands?" (I05), with values of 0.45, suggesting a moderate correlation with the other items.

The items with the highest correlations were "Do you feel nervous?" (I06), "Do you have difficulty thinking clearly?" (I08), "Have you been feeling sad?" (I09), "Do you find it difficult to perform daily activities?" (SRQ11), "Do you feel like a useless person?" (I16), and "Do you feel tired all the time?" (I18), with correlations ranging between 0.70 and 0.75. These items have a high correlation with the others, indicating that they measure similar characteristics.

Overall, the results demonstrate good internal consistency in the questionnaire.

4.2 Dimensionality Assessment

Initially, we proposed a one-factor analysis implemented with the *fa* function of the *psych* package. Then, we also tested whether it is possible to propose two factors for the test. Results of one and two-factor analyses, and measures of adequacy of factorial scores are shown on Tables 2 and 3.

Table 2. Results of the one- and two-factor analysis, and adequacy measures of the factor scores

Item	One-Factor Solution			Two-Factor Solution			
	MR1	h2	com	MR1	MR2	h2	com
I01	0.54	0.29	1	0.25	0.57	0.39	1.4
I02	0.54	0.29	1	0.41	0.35	0.29	2
I03	0.57	0.32	1	0.43	0.36	0.32	1.9
I04	0.46	0.21	1	0.27	0.4	0.24	1.8
I05	0.46	0.21	1	0.28	0.39	0.23	1.8
I06	0.78	0.6	1	0.55	0.56	0.62	2
I07	0.59	0.35	1	0.2	0.76	0.62	1.1
I08	0.76	0.58	1	0.61	0.46	0.58	1.9
I09	0.82	0.67	1	0.77	0.33	0.71	1.3
I10	0.61	0.38	1	0.48	0.37	0.37	1.9
I11	0.74	0.55	1	0.75	0.23	0.62	1.2
I12	0.59	0.35	1	0.5	0.32	0.35	1.7
I13	0.72	0.52	1	0.64	0.35	0.53	1.6
I14	0.69	0.47	1	0.76	0.14	0.6	1.1
I15	0.71	0.5	1	0.73	0.21	0.57	1.2
I16	0.73	0.54	1	0.75	0.22	0.62	1.2
I17	0.71	0.51	1	0.69	0.27	0.55	1.3
I18	0.75	0.57	1	0.57	0.49	0.57	2
I19	0.69	0.48	1	0.56	0.4	0.48	1.8
I20	0.59	0.35	1	0.11	0.91	0.83	1

Table 3. Results of the one- and two-factor analysis, and adequacy measures of the factor scores

Measures	One Factor	Two Factors	
	MR1	MR1	MR2
Eigenvalues SS	8.74	6.14	3.92
Proportion of Variance	0.44	0.31	0.2
Cumulative Variance	0.44	0.31	0.5
Explained Proportion After Rotation	-	0.61	0.39
Cumulative Proportion After Rotation	-	0.61	1
Average Item Complexity	1	1.5	
RMSR	0.08	0.05	
Fit Based on Off-Diagonal Values	0.97	0.99	
Correlation of Scores with Factors	0.97	0.95	0.94
Multiple R-Squared of Scores with Factors	0.95	0.9	0.89
Minimum Correlation of Factorial Scores	0.89	0.8	0.77

We observe that the results of the factor analysis using the unweighted least squares method (*minres*) (Harman & Jones, 1966) indicate that a single factor, named MR1, is sufficient to explain the variance of data. This factor explains 44% of the total variance, with an average item complexity equal to 1; additionally, all items have factor loadings above 0.25. The model fit is evaluated as excellent, with a root mean square residual (RMSR) of 0.08 and an adequate correlation between the estimated factorial scores and the extracted factors (0.97).

According to the results obtained with two factors, the first factor, MR1, presents a standardized loading of 0.61, explaining 31% of the total variance, while the second factor, MR2, has a standardized loading of 0.39, explaining 20% of the variance. Together, these two factors explain 50% of the total variance in the data. The model also suggests a good fit to the data, with an RMSR of 0.05 and a fit of 0.99 based on the off-diagonal values of the correlation matrix. Furthermore, the adequacy of the factorial scores is confirmed by the high correlation coefficients between the estimated scores and the extracted factors (0.95 for MR1 and 0.94 for MR2).

Although the two-factor solution may be interesting, based on the analysis above, we conclude that a single factor is sufficient because the first factor explains more than 40% of the variance and the items have loadings above 0.35. Therefore, the SRQ-20 questionnaire can be considered unidimensional.

4.3 Application of IRT Models

Next, we applied the one-parameter (1L), two-parameter (2L), and three-parameter (3L) item response theory models considering the *mirt* package (Chalmers, 2012) and a maximum likelihood estimation for the SQR-20 data since the questionnaire was considered unidimensional. The results of the 1L, 2L, and 3L models are presented below in Table 4.

Table 4. Table of item parameters with the mirt package for the 1L, 2L and 3L models

Item	1L	2L		3L		
	b	a	b	a	b	c
I01	0.387	1.023	0.315	1.198	0.483	0.073
I02	0.904	1.090	0.725	1.095	0.725	0.000
I03	-0.596	1.166	-0.476	1.170	-0.472	0.000
I04	0.512	0.833	0.490	0.841	0.489	0.000
I05	1.184	0.870	1.124	0.881	1.132	0.004
I06	-2.480	2.237	-1.374	2.221	-1.377	0.000
I07	0.874	1.241	0.644	1.249	0.644	0.000
I08	-0.145	2.0267	-0.0963	2.175	-0.048	0.026
I09	-0.755	2.528	-0.415	2.522	-0.408	0.000
I10	1.084	1.354	0.761	1.525	0.801	0.029
I11	-1.313	2.024	-0.769	2.112	-0.708	0.038
I12	-0.881	1.262	-0.663	1.487	-0.373	0.140
I13	0.526	1.816	0.309	1.831	0.312	0.000
I14	0.978	1.721	0.602	1.884	0.630	0.020
I15	-0.582	1.793	-0.368	1.906	-0.298	0.038
I16	0.422	1.910	0.240	1.923	0.244	0.000
I17	2.007	1.772	1.235	1.785	1.230	0.000
I18	-1.017	2.079	-0.593	2.077	-0.587	0.000
I19	-1.313	1.758	-0.818	1.746	-0.816	0.000
I20	0.491	1.199	0.362	1.205	0.365	0.000

When fitting the 1PL model, it is observed that the estimates of the difficulty parameters (b) of the SRQ questions using *mirt* package ranged from -2.480 to 2.007, corresponding respectively to items 06 and 17. In general, we can think of the difficulty parameter as referring to the degree of severity of the symptom; when negative, it means the symptom is less severe, and the more positive it is, the more severe the symptom. Thus, question 06, which deals with "Do you feel nervous, tense, or worried?" (item I06), represents a less severe symptom among students and can be considered a more common symptom among them. In contrast, question 17, which addresses "Have you had any thoughts of ending your life?" (item I17), reflects the most severe symptom and can be considered a harder symptom to be present among students.

In the 2PL model, in addition to the difficulty parameter, there is also a discrimination parameter for each question. We found that the estimates of the SRQ difficulty parameters range between -1.3737 and 1.2354, corresponding to the same questions in the 1PL model, I06 and I17, respectively, with the same interpretation. For the discrimination parameter (a), which assesses how well a question differentiates between individuals with high and low scores of mental distresses, we

found values ranging between 0.8328 and 2.5284, corresponding to questions 04 ("Do you get startled easily?" (item I04)) and 09 ("Have you been feeling sad lately?" (item I09)), respectively. According to the literature, it is expected that items minimally have a discrimination greater than 1. In this case, we found that questions 4 and 5 have lower discrimination. Meanwhile, questions 7 and 9 show higher discrimination, indicating that these items may be relevant for identifying the risk of mental distress in the SRQ-20.

Finally, in the 3PL model, the discrimination values (a) range between 0.8406 for question 04 ("Do you get startled easily?" (item I04)) and 2.5220 for question 09 ("Have you been feeling sad lately?" (item I09)), meaning that question 04 is a less severe symptom while question 09 becomes a more severe symptom among undergraduate students.

Regarding the difficulty parameter, the values range between -1.3748 and 1.2298, corresponding to questions 06 and 17, respectively. Therefore, the questions "Do you feel nervous, tense, or worried?" (item I06) and "Have you had any thoughts of ending your life?" (item I17) are important for identifying a risk case in the SRQ-20 questionnaire.

This model includes a new parameter, 'guessing,' which refers to the probability of a respondent marking an item. In the questionnaire in question, we found guessing values very close to zero ($\min = 0.000$, $\max = 0.1369$, $\text{mean} = 0.01862$), indicating that student responses do not present this parameter.

We then conducted an ANOVA test to choose the best model for the data. Results are in Table 5.

Table 5. Model comparison

Modelo	LogLike	AIC	BIC
1PL	-8.803.857	17649.72	17748.69
2PL	-8.702.742	17485.48	17674.00
3PL	-8.700.626	17521.25	17804.03

According to the results, we note that the 1PL model has a significant difference compared to the 2PL model, and the 2PL model does not have a significant difference compared to the 3PL model. By parsimony, we chose the 2PL model because it presents the lowest AIC and BIC values.

For the following analysis steps, we chose the results obtained using the 2PL model estimated with the *mirt* package.

In the scatter plot presented in Figure 1, we have the questions distributed by the values of difficulty and discrimination.

From the Figure 1, we see that the most severe items, i.e., those with greater difficulty, are questions 17 ("Have you had any thoughts of ending your life?" (item I17)) and 5 ("Do you have trembling hands?" (item I05)). The least severe are items 6 ("Do you feel nervous, tense, or worried?" (item I06)) and 19 ("Do you tire easily?" (item I19)). That is, questions related to the user's physical state are more severe.

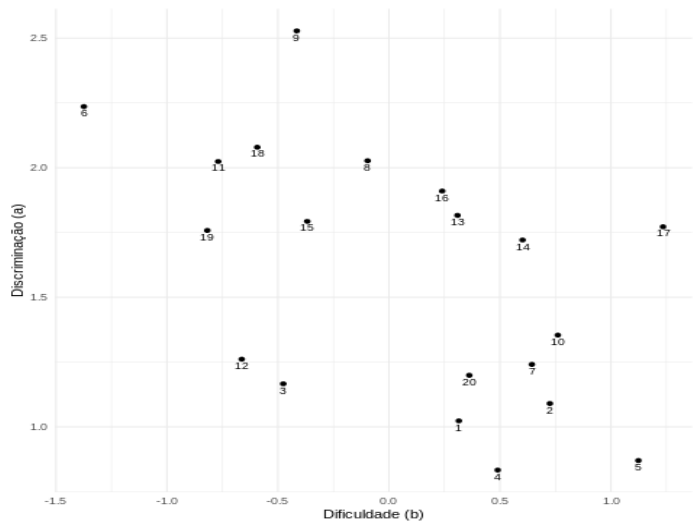


Figure 1. Scatter plot of discrimination parameter (a) and Difficulty parameter (b).

Additionally, the most discriminative (significant) items are 6 ("Do you feel nervous, tense, or worried?" (item I06)) and 9 ("Have you been feeling sad lately?" (item I09)), and the least significant are 4 ("Do you get startled easily?" (item I04)) and 5 ("Do you have trembling hands?" (item I05)). Thus, we see that questions related to emotions are more significant than those related to physical symptoms.

It is also possible to conclude that items 6 ("Do you feel nervous, tense, or worried?" (item I06)), 9 ("Have you been feeling sad lately?" (item I09)), and 19 ("Do you tire easily?") present lower severity and are more significant in the questionnaire, while items 17 ("Have you had any thoughts of ending your life?" (item I17)) and 14 ("Are you unable to play a useful role in your life?" (item I14)) are more severe and more significant.

Finally, items 12 ("Do you have difficulty making decisions?") and 3 ("Do you sleep poorly?" (item I03)) are less severe and less significant in the SRQ-20, and items 5 ("Do you have trembling hands?" (item I05)) and 10 ("Have you been crying more than usual?" (item I10)) are more severe and non-discriminatory.

Next, having obtained the item estimate results of the 2PL model for the SRQ-20 data, we estimated the latent traits underlying the response patterns considering the following code:

First, in Table 6, we show some statistics of the scores (obtained by summing the responses to the items) - SRQ Score.

Table 6. Frequency distribution of the SRQ score and descriptive statistics

Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Freq.	18	25	21	40	34	45	53	43	52	48	53	45	65	55	62	40	37	24	29	23	11

	n	Mean	SD	Median	Trimmed Mean	MAD	Min	Max	Range	Skewness	Kurtosis	SE
X1	823	9.94	5.06	10	9.95	5.93	0	20	20	-0.05	-0.88	0.18

Analyzing the statistics of the SRQ score variable in Table 6, we see that it has a symmetric distribution, as evidenced by the mean (9.94) and median (10) being very close, along with skewness near zero (-0.05). The standard deviation (5.06) and median absolute deviation (MAD) (5.93) show moderate dispersion of values around the mean. The data range is 20, varying between 0 and 20.

Negative kurtosis (-0.88) suggests a flatter distribution with thinner tails than normal. With a large sample size (823) and a small standard error (0.18), the mean is a reliable estimate. Therefore, the data have a symmetric and moderately dispersed distribution.

Now, we estimate the latent traits considering the IRT model and a scale with a mean of 50 and a standard deviation of 10 and obtain the following results showed in Table 7:

Table 7. Descriptive statistics of the IRT score

	n	Mean	SD	Median	Trimmed Mean	MAD	Min	Max	Range	Skewness	Kurtosis	SE
X1	823	49.94	8.94	49.79	49.92	9.24	30.12	70.68	40.56	0.01	-0.41	0.31

The statistics on the TRI score indicate that the distribution is symmetric and moderately dispersed. The mean (49.94) and median (49.79) are very close, suggesting symmetry, corroborated by the nearly null skewness (0.01). The standard deviation (8.94) and median absolute deviation (MAD) (9.24) point to moderate data dispersion. Negative kurtosis (-0.41) indicates a slightly flatter distribution than normal. With a large sample size (823) and a small standard error (0.31), the estimated mean is reliable and precise.

The correlation between the SRQ score and TRI score variables is 0.989168, as evidenced by the graph below, showing a strong positive linear relationship, indicating that the two variables are highly correlated. See Figure 2.

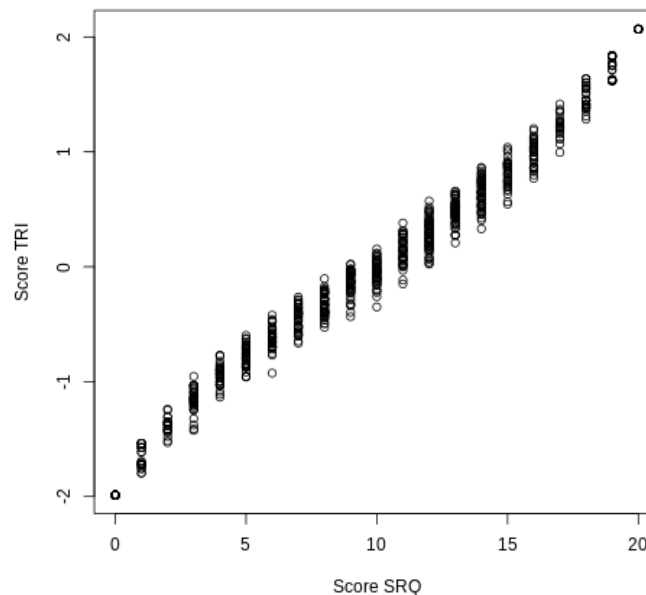


Figure 2. Relationship between scores using classical score (Score SRQ) and Score IRT.

According to the results for a given score obtained by summing the YES responses of the test (SRQ Score), we can obtain different values in the latent trait estimates of the evaluated subjects using the TRI score. This means that while the SRQ score cannot differentiate individuals with similar scores, the TRI score allows us to capture these differences based on differences in responses to the questionnaire items and their various characteristics or item parameters.

Thus, the TRI score proves to be more sensitive in identifying differences as a screening scale than the SRQ score.

Below, we present an evaluation of the normality of the TRI scores. The Shapiro-Wilk normality test result with a p-value of 0.001202 indicates that the TRI score distribution differs significantly from a normal distribution, suggesting that the data are not normally distributed. Additionally, we assessed normality graphically using the *ggqqplot* function from the *ggpubr* package, which creates a Q-Q plot or quantile-quantile plot showing the relationship between the score and the normal distribution. A 45-degree line passing through the .25 and 0.75 quantiles of the data is shown as a reference for perfect normality. Additionally, the function shows a confidence band of 95% for the quartile line using a parametric standard error.

The Q-Q plot presented in Figure 3 shows deviations at the extreme points relative to the diagonal line, and furthermore, they are not outside the confidence band, indicating that the TRI score in this application does not follow normality, consequently, on the following section we will use nonparametric test to comparative analysis (Smeeton *et al.*, 2025).

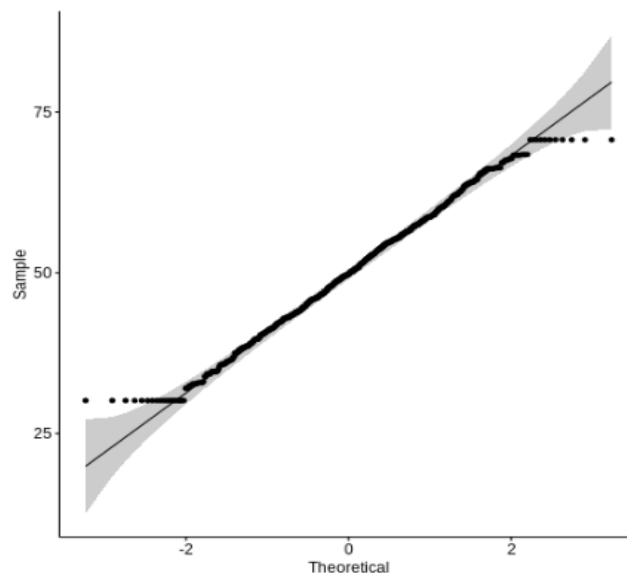


Figure 3. QQplot of the TRI score evaluating normality.

4.4 Comparative Analysis

Some sociodemographic variables such as student ethnicity, year of university entry, declared gender, and whether they live in university housing were also collected in the survey, so we were interested in comparing the TRI score according to these characteristics.

However, in the comparison between the different groups, we observed that the students who responded to the questionnaire, coming from various colleges of the São Carlos campus, do not necessarily reflect the real distribution among them in the corresponding semesters. As a result, we received a larger number of responses from smaller colleges and a reduced number of responses from larger colleges. To minimize this bias in the statistical methodology, it is common to apply weights that adjust the proportions of respondents to reflect the expected population distribution, as Lumley & Scott (2013) did to develop tests comparing two samples under complex sampling (Lumley, 2010).

To calculate the weight values in the populations of 2021, 2022, and 2023, a survey of the total population size of the campus and each college was conducted to calculate the population proportion (population size/total population). Then, the same calculation was used to find the sample proportion (sample size/total sample) of each college in each year. Finally, the weight value is given by the

population proportion/sample proportion, and to find the weighted sample size value, simply multiply the weight by the sample size. Thus, for each student in the sample, depending on the college they belonged to and the year they answered the questionnaire, a new weight variable value was determined.

Concluding the non-normality seen in the previous section, the statistical analyses were conducted using the non-parametric Wilcoxon and Kruskal-Wallis tests, according to the number of groups compared. The Wilcoxon test was applied for comparisons between two groups, while the Kruskal-Wallis test was used for comparisons involving three or more groups. In cases where the p-value obtained was less than 0.05, indicating a statistically significant difference, post hoc analyses were conducted to identify which groups showed significant distinctions.

The implementation was carried out in the *R* language, using the *svyranktest* function for the Wilcoxon and Kruskal-Wallis tests. For the post hoc comparisons, a loop was developed that iteratively calculates the Kruskal-Wallis test between the groups identified as distinct, and as mentioned earlier, the tests were conducted with sample weighting, ensuring the correction of results to reflect the population proportion. The results with weights are presented in Table 8.

Table 8. Wilcoxon tests for comparison of two groups and Kruskal-Wallis for comparison of more than two groups, using weighting

Grouping	n	Median IRT	IQR IRT	Test Statistic	p-value
Age				23.458	0.000009497 *
Up to 19	278	47.87	8.87		
20 to 25	457	50.30	12.92		
Over 25	88	23.10	11.81		
Race				7.518	4.524
White	610	49.43	11.69		
Non-white	213	49.69	12.47		
Year of Entry				13.298	0.01031 *
2019 or before	229	52.76	13.34		
2020	150	49.29	12.26		
2021	223	48.82	12.51		
2022	135	48.18	13.94		
2023	86	48.77	7.92		
Gender				27.929	0.000001087 *
Cis man	416	47.78	12.80		
Cis woman	374	51.37	11.91		
Other identities	33	53.80	14.56		
Student Housing				1.4287	1.535
No	765	49.44	12.14		
Yes	58	51.77	10.25		

When analyzing the groupings (Table 8), we identified significant differences between the Age, Year of Entry, and Gender groups, as highlighted in the table.

Next, in Table 9, we present the specific results of the comparisons between the groups of each variable, accompanied by Piecestack plots (Wu et al, 2016), which offer a clearer visualization of the data distribution within these groups (Figure 4).

Table 9. *p*-values for comparisons between groups

Category	Comparison	p-value
Age	Up to 19 vs. 20 to 25	0.0005644*
	Over 25 vs. 20 to 25	0.2091
	Over 25 vs. Up to 19	0.0008227*
Year of Entry	2020 vs. 2019 or earlier	0.1085
	2021 vs. 2019 or earlier	0.0004832*
	2022 vs. 2019 or earlier	0.03688*
	2023 vs. 2019 or earlier	0.03414*
	2021 vs. 2020	0.1026
	2022 vs. 2020	0.424
	2023 vs. 2020	0.5185
	2022 vs. 2021	0.569
	2023 vs. 2021	0.6811
	2023 vs. 2022	0.9782
Gender	Woman vs. Man	0.000003585*
	Other Identities vs. Man	0.0007636*
	Other Identities vs. Woman	0.2533

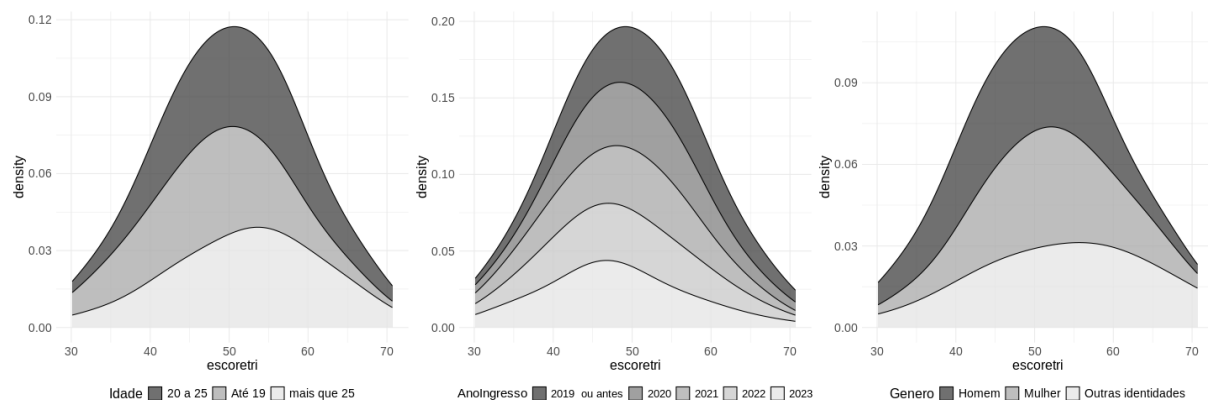


Figure 4. Density plots showing the distribution of the TRI score according to different variables, including Age, Year of Entry and Gender.

5. Conclusions

The analysis revealed that sociodemographic factors such as age, gender, and university affiliation have a significant impact on the well-being and mental health of students.

The inclusion of weights in the analyses highlighted the importance of the undergraduate course. Additionally, SRQ-20 questionnaire items showed that emotional issues, such as nervousness and sadness, are more discriminative than physical ones, underscoring the need to focus on emotional aspects when addressing psychological distress. More severe items, such as suicidal thoughts, are concerning but not necessarily more discriminative, suggesting the need for interventions that consider both the severity and relevance of symptoms. These conclusions can guide the creation of more effective strategies for promoting health and well-being in the university environment. Additionally, the findings could be further discussed by comparing them with results reported by other authors in the clinical field, such as Silveira *et al.* (2022), among others.

Finally, alternative methodologies incorporating new item response models — such as those discussed in Bazan *et al.* (2006), Azevedo *et al.* (2011), among others — could be considered in future analyses. Further methodological details are provided in Amaral *et al.* (2025), where the analysis presented here is reproduced. That work may also serve as a valuable resource for teaching psychometric analysis within the statistics community and beyond.

Acknowledgments

Work developed by Brunna Quatrochi and Gabriela Amaral, as part of the projects of the 2023-2024 and 2024-2025 editions of the Unified Scholarship Program to support the Training of Undergraduate Students at the University of São Paulo (PUB-USP) under the supervision of Jorge Luis Bazán. We are grateful to the University of São Paulo and the teams of GAPsi and Apoia USP for their support and collaboration. We also thank the editor for his comments.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: BAZAN, J. L.; **Data curation:** SILVA, E. P.; **Formal analysis:** QUATROCHI, B., AMARAL, G. S.; **Methodology:** BAZAN, J. L., QUATROCHI, B., AMARAL, G. S.; **Supervision:** BAZAN, J. L.; **Visualization:** QUATROCHI, B.; **Writing - original draft:** QUATROCHI, B., AMARAL, G. S.; **Writing - review and editing:** SILVA, E. P., BAZAN, J. L.

Data Availability Statement

The data that support the findings of this study are openly available in https://github.com/jlbazan/Analise_Psicometrica_SQR20 at <https://doi.org/10.5281/zenodo.15246854>. See Amaral *et al.* (2025).

References

1. Amaral, G., Quatrochi, B., Bazán, J. L and Pap da Silva, E. *Análise Psicométrica da escala de rastreio SQR-20 usando dados do GAPSI e Apoia USP*. Version v1.0 (2025). <https://doi.org/10.5281/zenodo.15246854>
2. Andrade, D. F., Tavares, H. R., Valle, R. C. *Teoria da Resposta ao Item. Conceitos e Aplicações*. (Associação Brasileira de Estatística, 2000).
3. Azevedo, C. L., Bolfarine, H., & Andrade, D. F. Bayesian inference for a skew-normal IRT model under the centred parameterization. *Computational Statistics & Data Analysis* **55**, 353-365 (2011).

4. Bazán J. L. Psicometria e avaliação por testes: um marco metodológico. In *Avaliação da educação: referências para uma primeira conversa* (EdUFSCar, 2018).
5. Bazán, J. L., Bolfarine, H., & Branco, M. D. A skew item response model. *Bayesian Analysis*, 861-892 (2006).
6. Bazán, J. L., Valdivieso, L. H., & Calderón, A. *Enfoque Bayesiano em modelos de teoria de resposta ao item*. (PUCP, 2010). Available in: <http://argos.pucp.edu.pe/~jlbazan/download/Reportef27.pdf>
7. Carmines, E. G. & Zeller, R. A. *Reliability and Validity Assessment* (Sage, 1979).
8. Chalmers, R., P. *mirt: A Multidimensional Item Response Theory Package for the R Environment*. *Journal of Statistical Software* **48**, 1-29 (2012).
9. da Silva, M. A., Bazán, J. L. & Huggins-Manley, A. C. Sensitivity analysis and choosing between alternative polytomous IRT models using Bayesian model comparison criteria. *Communications in Statistics-Simulation and Computation* **48**, 601-620 (2019).
10. de Oliveira, E. S., Wang, X. & Bazán, J. L. A classification model for continuous responses: Identifying risk perception groups on health-related activities. *Biometrical Journal* **65**, 2100222 (2023).
11. Fernandes, R., Bazán, J. L. & Cúri, M. A Bayesian approach for the G-DINA model. *Brazilian Journal of Probability and Statistics* **38**, 503-530 (2024).
12. Gadermann, A., Guhn, M. & Zumbo, B. Estimating Ordinal Reliability for Likert-Type and Ordinal Item Response Data: A Conceptual, Empirical, and Practical Guide. *Practical Assessment. Research & Evaluation* **17**, 1-13 (2012).
13. Harman, H. & Jones, W. Factor analysis by minimizing residuals (minres). *Psychometrika* **31**, 351-378 (1966).
14. Lumley, T. *Complex Surveys: A Guide to Analysis Using R*. (John Wiley and Sons, 2010).
15. Lumley, T., & Scott, A. J. Two-sample rank tests under complex sampling. *Biometrika* **100**, 831-842 (2013).
16. Nunnally, J. & Bernstein, I. *Psychometric Theory* (McGraw Hill, 1994).
17. Orlando, M., Sherbourne, C. D. & Thissen, D. Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment* **12**, 354- 359 (2000).
18. Pasquali, L., & Primi, R. Fundamentos da teoria da resposta ao item: TRI. *Avaliação Psicológica: Interamerican Journal of Psychological Assessment* **2**, 99-110 (2003).
19. Pettersson, E., & Turkheimer, E. Item selection, evaluation, and simple structure in personality data. *Journal of Research in Personality*, **44** 407–420 (2010). <https://doi.org/10.1016/j.jrp.2010.03.002>
20. Rao, C. R., & Sinharay, S. Handbook of Statistics. *Psychometrics* **26**, 235 (Elsevier, 2007).
21. Revelle, W. *How To: Use the psych package for Factor Analysis and data reduction*. Available in: <https://cran.r-project.org/web/packages/psychTools/vignettes/factor.pdf>, 2022)
22. Silveira, L. B., Kroeff, C. da R., Teixeira, M. A. P., & Bandeira, D. R. Uso do Self-Reporting Questionnaire (SRQ-20) para identificação de grupo clínico e predição de risco de suicídio. *Revista Psicologia e Saúde* **13**, 49–61 (2022). <https://doi.org/10.20435/pssa.v13i4.1219>
23. Smeeton, N., Spencer, N. & Sprent, P. *Applied nonparametric statistical methods*. (CRC Press, 2025).
24. Wright, B. D. & Stone, M. H. *Best test design* (Mesa Press, 1979).
25. Wu, T., Wu, Y., Shi, C., Qu, H. & Cui, W. (2016). Piecstack: Toward better understanding of stacked graphs. *IEEE Transactions on Visualization and Computer Graphics* **22**, 1640-1651 (2016).
26. Zumbo, B. D., Gadermann, Anne M. & Zeisser, C. Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. *Journal of Modern Applied Statistical Methods*, **6**, Article 4 (2007). <https://doi.org/10.22237/jmasm/1177992180>