Pontifícia Universidade Católica
do Rio de Janeiro

**Rodrigo Leão Ferreira do Nascimento**

# Applied Psychometrics: Leveraging methods and models across diverse contexts

Thesis presented to the Programa de Pós-graduação em Psicologia of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Psicologia.

Advisor: Prof. Jesus Landeira-Fernandez
Co-Advisor: Luis Flávio Chaves Anunciação

Rio de Janeiro,
January 2025

Pontifícia Universidade Católica
do Rio de Janeiro

**Rodrigo Leão Ferreira do Nascimento**

# Applied Psychometrics: Leveraging methods and models across diverse contexts

Thesis presented to the Programa de Pós-graduação em Psicologia of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Psicologia. Approved by the undersigned Examination Committee.

**Prof. Jesus Landeira-Fernandez**
Advisor
Departamento de Psicologia - PUC-Rio

**Prof. Luís Flávio Chaves Anunciação**
Departamento de Psicologia - PUC-Rio

**Prof. Thomas Eichenberg Krahe**
Departamento de Psicologia - PUC-Rio

**Prof. Leonardo Fernandes Martins**
Departamento de Psicologia - PUC-Rio

**Prof. Pedro Paulo Pires dos Santos**
UFRJ

**Profa. Jane Kaplan Squires**
UOREGON

Rio de Janeiro, January 16, 2025

**Rodrigo Leão Ferreira do Nascimento**

The author graduated in Psychology at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2016 and completed his MSc Degree in Psychology with emphasis in Neuroscience at the same University in 2018. As part of his doctorate program, he worked as visiting research scholar at the University of Oregon (United States of America) in 2024.

Bibliographic data

Nascimento, Rodrigo Leão Ferreira do

Applied psychometrics : leveraging methods and models across diverse contexts / Rodrigo Leão Ferreira do Nascimento ; advisor: Jesus Landeira-Fernandez ; co-advisor: Luis Flávio Chaves Anunciação. – 2025.
116 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Psicologia, 2025.
Inclui bibliografia

1. Psicologia – Teses. 2. Análise fatorial confirmatória. 3. Teoria de resposta ao item. 4. Análise de redes. 5. Dados faltantes. 6. Desenvolvimento humano. I. Landeira-Fernandez, Jesus. II. Anunciação, Luis Flávio Chaves. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Psicologia. IV. Título.

CDD: 150

For my godmother and grandparents.

## Acknowledgments

To my advisor Prof. J. Landeira-Fernandez, for guiding me through the scientific path since my undergraduate time at PUC-Rio.

To my co-advisor Prof. Luis Anunciação, for all the opportunities, technical advisory, and mentoring you gave me.

To Prof. Jane Squires, for hosting me and Luisa in Eugene with great kindness and care. Thank you for accepting me as a student and as part of your research team.

To Prof. Leonardo Martins, for the incredible classes and contributions to this project.

To Prof. Pedro Pires, for all the support and contributions you gave to this project.

To Prof. Thomas Krahe, for all the contributions you made to this project.

To Prof. Maracy Alves, for all the support and advice you gave me since I was an undergraduate student. It was a great pleasure to be your student.

To Prof. Roberto Cruz Moraes, for all the comments you made about this project.

To Prof. Clarissa Freitas, for all the mentoring and support you gave to me during these years.

To Prof. Emmy Uehara, for giving me the first scientific opportunity.

To the CNPq, FAPERJ and CAPES, for all scholarships and funding you provided me which made possible the present study.

To the entire community of PUC-Rio, especially, all the staff of the Department of Psychology, for all the support you provided me.

To the members of the College of Education from the University of Oregon, especially Prof. Chris Murray, for all the support and advice you gave me during my time in Eugene.

To the ASQ family, especially Prof. Diane Bricker, Dr. Jantina Clifford, and Kimberly Murphy, for all the dinners, laughs, and conversations we had in Eugene.

To the members of the Methods and Measurement Group and the Laboratório de Saúde Mental e Trabalho, for all the laughs and conversations we had together.

Alexandre, Amanda, Louise, Elias, Renato, Miguel, Juliana, Héber, Isabella, and Florencia, you are the best.

To Dr. Sylvio Guimaraes and the Clinical Psychologist Ruan Silva, for the mental health services provided to me. Your support made the difference.

To my colleagues and friends, for all the laughs, memes, and support you gave me through this endless journey. Especially to Anderson, Victor, Monique, Joana, Joã, Angela Della Gaspera, Julia, Rodolfo, Gabriel, and the Karaoke and Dota crews.

To my sensei's, colleagues, and students of martial arts, for all incredible work we have done together. It has been an honor to pursue the way of the warrior with you.

To my family, Ronaldo, Ana Lucia, Ronaldo, Ana Flora, Brunna, Débora, and Isadora, for all the love you gave me since I was a kid. I have no words to express how important you are to me.

To my wife and two cats, Luisa, Prometeu (aka Teteu), and Jonas, for choosing to share their lives with me.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

# Abstract

Leão Ferreira do Nascimento, Rodrigo; Landeira-Fernandez, Jesus (Advisor). **Applied Psychometrics: Leveraging Methods and Models Across Diverse Contexts.** Rio de Janeiro, 2025. 116 p. Doctorate Thesis – Department of Psychology, Pontifical Catholic University of Rio de Janeiro.

Psychometric studies are fundamentally important to ensure fair assessment processes in different contexts. Their use is not limited to a specific area or scientific discipline, being widely applied in fields such as Education. Regardless of the application context in which psychological characteristics are to be measured, it is necessary to employ techniques and procedures based on different psychometric traditions. Such techniques and procedures aim to generate evidence to support or refute the validity of the inferences made from the interpretations of the results produced by these instruments. Given the range of available techniques and procedures, it is crucial to contextualize the application of these different traditions, considering their substantive, methodological, and historical attributes and characteristics. In this sense, the present scientific investigation sought to focus on two distinct emphases: (1) the application of different psychometric techniques and (2) the exploration of methodological challenges, with the goal of providing evidence on both existing and newly developed psychometric instruments. To this end, the present thesis was organized into four articles. The first article aimed to gather evidence of measurement invariance for two mental health measures (depression and anxiety) in samples from three countries: Brazil, Spain, and Portugal. Overall, the results suggest the stability of the depression measure, while the invariance of the anxiety measure was found to be less stable. The second article focuses on gathering validity evidence for a recently developed instrument for assessing job penibility in the Brazilian hospital context. The findings indicate a more robust factorial structure with a single dimension, although contingent factors related to the theoretical proposition of a three-factor structure are discussed. In addition to the confirmatory investigation of the instrument's internal structure, an Item Response Theory (IRT) approach was implemented, revealing that the items, in general, presented moderate to high levels of difficulty. The third application refers to the use of Network Analysis techniques in the context of child

development in the United States. The results can be interpreted within the framework of a stable network structure comprising six communities, which aligns with the theoretical proposal of the measurement instrument. Finally, the fourth study described in this thesis aimed to present opportunities and methodological challenges in handling large amounts of missing data within IRT models, also in the context of child development. Overall, the evidence presented in this thesis highlights the importance of adopting a substantively grounded approach to the application of different psychometric techniques and procedures. This is relevant not only for the investigation of internal structure but also for the development of new instruments, particularly when seeking to understand the intricate statistical relationships that emerge across different models. Furthermore, this study sheds light on current and future methodological challenges, especially considering the growing trend toward instruments that prioritize not only the needs of developers but also the experience of respondents.

## Keywords

Confirmatory Factor Analysis; Item Response Theory; Network Analysis; Missing Data; Human Development.

# Resumo

Leão Ferreira do Nascimento, Rodrigo; Landeira-Fernandez, Jesus (Orientador). **Psicometria Aplicada: Explorando Métodos e Modelos em Contextos Diversos.** Rio de Janeiro, 2025. 116 p. Tese de Doutorado – Departmento de Psicologia, Pontifícia Universidade Católica do Rio de Janeiro.

Estudos psicométricos são de fundamental importância para garantir processos de avaliação justos em diferentes contextos. Sua utilização não está restrita a uma certa área ou disciplina científica, sendo amplamente aplicada em áreas como a Educação. Independentemente do contexto de aplicação em que se pretende mensurar tais características psicológicas, será necessário o uso de técnicas e procedimentos baseados em diferentes tradições psicométricas. Tais técnicas e procedimentos visam gerar evidências para se apoiar ou não a validade das inferências feitas a partir das interpretações dos resultados desses instrumentos. Diante dessa oferta de técnicas e procedimentos, faz-se importante colocar em perspectiva a aplicação dessas diferentes tradições, considerando suas características e atributos substantivos, metodológicos e históricos. Nesse sentido, a presente investigação científica buscou se apoiar em duas ênfases distintas: (1) aplicação de diferentes técnicas psicométricas e exploração de desafios metodológicos; (2) para oferecer evidências acerca de instrumentos psicométricos utilizados e novos. Para isso, a presente tese foi dividida em quatro artigos. O primeiro artigo buscou obter evidências de invariância de duas medidas em saúde mental (depressão e ansiedade) em amostras de três países: Brasil, Espanha e Portugal. Em linhas gerais, os resultados obtidos apontam para a estabilidade da medida de depressão, sendo a invariância menos estável na medida de ansiedade utilizada. O segundo artigo apresentado se refere à busca de evidências de validade para um inventário de avaliação da penosidade recém-criado, no contexto hospitalar brasileiro. Os achados encontrados apontam para uma estrutura fatorial mais robusta com uma dimensão, embora sejam discutidos fatores contingentes à estrutura teórica de três fatores. Além da investigação confirmatória da estrutura interna do instrumento, foi implementada uma técnica de Teoria de Resposta ao Item, que revelou níveis médios a difíceis para os itens, no geral. A terceira aplicação apresentada diz respeito ao uso de técnicas de Análise de Redes no

contexto do desenvolvimento infantil dos Estados Unidos. Os resultados obtidos podem ser interpretados sob a luz de uma estrutura de rede estável, com seis comunidades, o que vai de encontro à proposta teórica do instrumento de medida. Finalmente, o quarto estudo descrito nessa tese visou apresentar oportunidades e desafios metodológicos no tratamento de grandes quantidades de dados faltantes dentro de modelos de TRI, no contexto do desenvolvimento infantil. De modo geral, as evidências apresentadas nessa tese apontam para a importância da adoção substantivamente consubstanciada de diferentes técnicas e procedimentos psicométricos na aplicação, tanto na investigação da estrutura interna quanto no desenvolvimento de novos instrumentos, sobretudo buscando observar as intrincadas relações estatísticas verificadas entre os diferentes modelos. Além disso, o presente estudo aponta para desafios metodológicos atuais e futuros, sobretudo a partir da tendência na área de instrumentos que privilegiem não só as necessidades dos desenvolvedores, mas também a experiência dos respondentes.

## Palavras-chave

Análise Fatorial Confirmatória; Teoria de Resposta ao Item; Análise de Redes; Dados Faltantes; Desenvolvimento Humano.

## Table of Contents

## List of abbreviations

| | |
|---|---|
| **ACEs** | Adverse Childhood Experiences |
| **AIC** | Akaike Information Criterion |
| **ANOVA** | Analysis of Variance |
| **ASQ:SE** | Ages & Stages Questionnaire: Social-Emotional |
| **ASQ:SE-2** | Ages & Stages Questionnaire: Social-Emotional, 2nd Edition |
| **ASQ-3** | Ages & Stages Questionnaire, 3rd Edition |
| **BAI** | Beck Anxiety Inventory |
| **BDI-II** | Beck Depression Inventory, 2nd Edition |
| **BIC** | Bayesian Information Criteria |
| **CBO** | Brazilian Occupational Classification |
| **CFA** | Confirmatory Factor Analysis |
| **CFI** | Comparative Fit Index |
| **CI** | Confidence Interval |
| **CTT** | Classical Test Theory |
| **DF** | Degrees of Freedom |
| **DSM-III-R** | Diagnostic and Statistical Manual of Mental Disorders, 3rd edition Revised |
| **DSM-IV** | Diagnostic and Statistical Manual of Mental Disorders, 4th edition |
| **DWLS** | Diagonal Weighted Least Square |
| **EBIC** | Extended Bayesian Information Criterion |
| **EFA** | Exploratory Factor Analysis |
| **ESQ-RE** | Environmental Screening Questionnaire - Research Edition |
| **FA** | Factor Analysis |
| **FACSIMILE** | Factor Score Item Reduction with Lasso Estimator |
| **GG** | Major Groups |
| **GRM** | Graded Response Model |
| **HMMC** | Miguel Couto Municipal Hospital |
| **IAP** | Penibility Assessment Inventory |
| **ICC** | Item Characteristic Curve |
| **IPV** | Intimate Partner Violence |
| **IRT** | Item Response Theory |

| | |
|---|---|
| **KI** | Kurtosis |
| **LE** | Life Expectancy |
| **M** | Mean |
| **MAR** | Missing At Random |
| **MCAR** | Missing Completely At Random |
| **Med** | Median |
| **MGCFA** | Multi-Group Factorial Analysis |
| **MNAR** | Missing Not At Random |
| **NAEP** | National Assessment of Educational Progress |
| **OD** | Operational Definitions |
| **PISA** | Programme for International Student Assessment |
| **PMRF** | Pairwise Markov Random Field |
| **PSI-SF** | Parenting Stress Index-Short Form |
| **RMSEA** | Root Mean Square Error of Approximation |
| **SABIC** | Sample-Size Adjusted BIC |
| **SD** | Standard Deviation |
| **SDOH** | Social Determinants of Health |
| **SI** | Skewness |
| **SNAP** | Food Stamps |
| **TCLE** | Informed Consent Form |
| **TLI** | Tucker-Lewis Index |
| **US** | United States |
| **WHO** | World Health Organization |
| **WIC** | Women, Infants, and Children |

# List of figures

## Theoretical Background



## Article 1



## Article 2



## Article 3

## List of tables

### Article 1



### Article 2



### Article 3

**Article 4**

*Gate Gate Paragate Parasamgate Bodhi Svaha*

Heart Sutra

# I. THEORETICAL BACKGROUND

A famous quote, attributed to the German psychologist Herman Ebbinghaus (1850–1909), says that "Psychology has a long past, but only a short history" (Goodwin, 2022). Across his textbook about the history of this discipline, the author explains the lack of consensus among historicists of Psychology on its date of birth. Some authors say it starts in Greece, while others elect the Modern Age with René Descartes (1596-1650), albeit vast majority agree that at least the contemporary version of the psychological science starts in Germany on the second half of the nineteenth century. Since then, the science of Psychology has gained social significance and credibility within different societies worldwide. As specified by the author, its wide range of applications was one of the main reasons for the successful maintenance of Psychology as a scientific enterprise up to the 30ths of the past century, being psychological assessment one of its flagships.

Following Rust e Golombok (2008), the science of psychological assessment is named Psychometrics. It's noteworthy that one could rephrase Ebbinghaus' quote to "Psychometrics has a long past, but only a short history" given that its roots trace back to selection of talents in the Xia Dinasty (between the 21st century BC and the 17th century BC); despite its contemporary (and current) version dates to the 19th century. In which a central role was played by the eugenicist Sir Francis Galton (1822-1911) in his studies of the human intellect (Rust e Golombok, 2008). Since there, as pointed out by the authors, Psychometrics has become central to modern society. This conclusion can be easily verified given the different applications of Psychometrics in Education, Industry and Organizations, Human Development etc. (AERA, APA e NCME, 2014). As introduced in the Standards in Educational and Psychological Testing book, one of its main purposes is 'to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses' (p. 1). But, after all, what is validity?

According to the Standards in Educational and Psychological Testing (AERA, APA e NCME, 2014), the concept of validity 'refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (p. 11). It follows that validity plays a major role in the testing field. It is composed of sources of evidence based: on the test content, response processes, internal

structure, relations to other variables, and consequences of testing. Thus, to qualify a test as ‘valid or not valid’ reveals a misunderstanding about the current definition of validity, that concerns the interpretation of test scores. It implies that to provide more or less valid interpretations of test scores, it is necessary to specify a construct.

The notion of construct is commonly used in the field, but as pointed out by Boeck, De et al. (2023), it is rarely defined. In this review, the authors raise a fortunate debate about the notion of construct. According to them, the emerging dissatisfaction among psychometricians about this notion gave rise to two new trends that intended to narrow the construct-OD gap (i.e. OD refers to operational definitions). The first trend concerns *downscaling constructs*, which means to equivale it ‘to the level of effects and observables functioning as local constructs’ (p. 241). Whilst the contrasting trend toward *upscaling ODs* consists of ‘describing constructs more precisely so that ODs can be better determined for measuring these constructs’ (p. 242). Differently from both trends, the authors propose an alternative approach to this problem trying to account the complexity and variability of psychological phenomena.

Hidden in the heart of the debate underlie the consequences of the replication crisis. This recent event led to community, structural, and procedural changes in Psychology investigation (Korbmacher *et al.*, 2023; Nosek *et al.*, 2022). In line with this, Eronen e Bringmann (2021) argued that to move forward (the crisis), it is necessary to impose constraints for theories, and to improve validity studies, conducting construct-related analyses (e.g. internal structure etc.). All things considered; distinct voices seem to converge into the strengthening of substantive development combined with rigorous statistical modelling with the purpose of going to the next level in the field.

Undoubtedly, enthusiasts of the Network Model approach have been playing an important role at the debate, displaying criticisms about the Common Cause Model (limited) achievements after a century of studies in psychopathological field (Borsboom e Cramer, 2013). The latter lies in the idea that unobservable entities (e.g. Major Depression Disorder) cause observables (e.g. lack of energy, sadness etc.). According to the authors, most of the psychometrical traditions (e.g. Item Response Theory etc.) assume that part of the variance of a set of items is explained

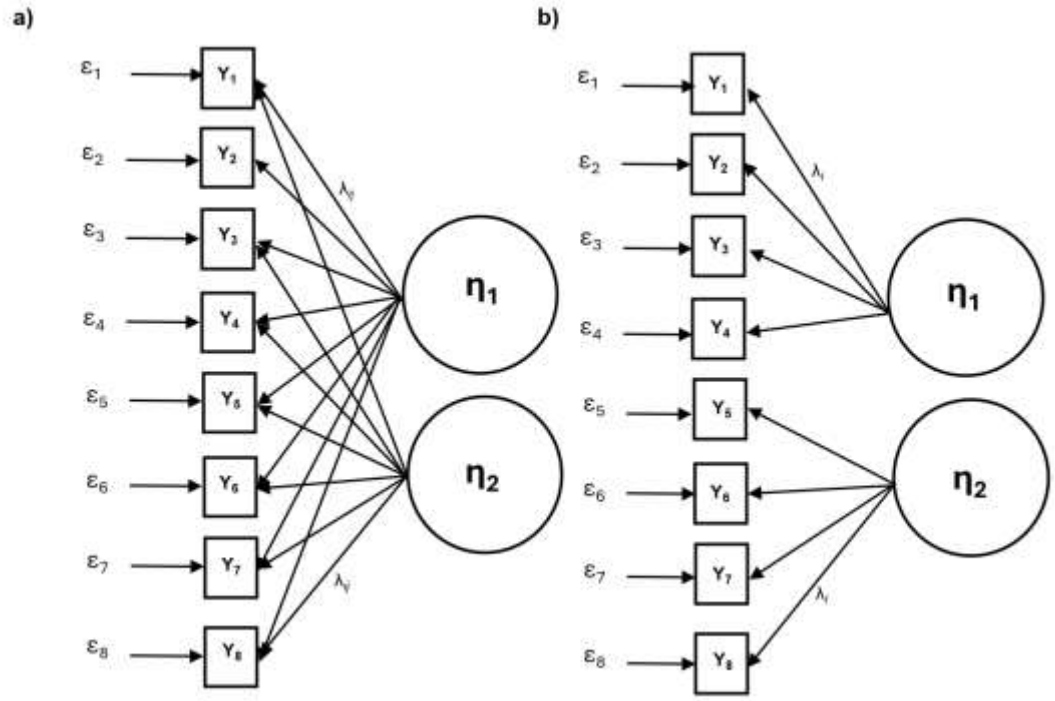
by the presence of unobservable variables (i.e. latent traits). Thus, these traditions were mirroring that paradigm. In fact, latent traits theory assumes an isomorphism among hypothetical variables (theta, $\theta$) and observable behaviors (items) which allow its indirect measurements (Pasquali, 2013). The author outlines that even the true score (tau, $\tau$) derived from the Classical Test Theory (CTT) is classified by some authors (e.g. Weiss) as equivalent to a latent trait.

CTT was the very first tradition in the field, encompassing several concepts and techniques in order to measure psychological attributes such as intelligence (Pasquali, 2013; Sartes e Souza-Formigoni, de, 2013). According to Rindskopf (2015), its main assumption is that a single underlying dimension is being measured, and that each person has a single true score on this dimension. In addition to that, each person's observed test score is also determined by some error of measurement. This assumption is usually expressed by the following equation:

$$X = T + e$$

In which $X$ represents a given person's observed test score, $T$ is the true score, and $e$ equivale to error measurement. In line with this tradition, following Brown (2015), the development of Factor Analysis (FA) by Charles Spearman (1863-1945) had a great impact among psychometricians. FA consists in a set of multivariate statistical models which aim to evaluate the dimensionality of an instrument through a few factors. Sartes e Souza-Formigoni de (2013) state that FA was widely adopted, at the 30ths of the past century, in part due to the advances proposed by Thurstone. According to Brown (2015), these advances are mainly summarized in the common factor model, 'which postulates that each indicator in a set of observed measures is a linear function of one or more common factors and one unique factor' (p. 11). The author explains that there are two main types of FA, the exploratory factor analysis (EFA) and the confirmatory factor analysis (CFA). The first one is more data-driven while the second stands on theoretical ground. In other words, it means that EFA doesn't require an explanatory model to identify the 'pattern of relationships between the common factors and the indicators' (p. 11), whilst CFA does. Figure 1 (below) represents path diagrams of a hypothetical uncorrelated two-dimensional model of depression using (a) EFA, and (b) CFA models.

Figure 1 - Path Diagram of EFA and CFA




In which each factor $\eta_i$ (eta) represents an underlying continuous variable normally distributed that predicts item ($Y_i$) common variance $\lambda_{i(j)}$ (lambda) and unique variance $\varepsilon_i$ (epsilon). With the purpose of ease reader's visualization, the $\lambda_{i(j)}$ symbols were omitted in between the border arrows.

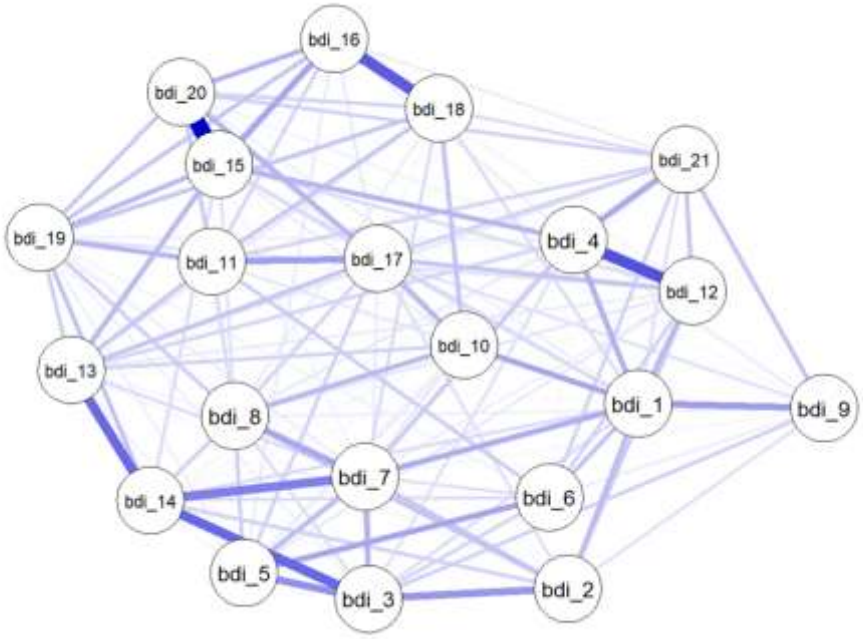
CTT and FA had limitations such as sample-dependency, and difficulty to estimate mixed-type data (Sartes e Souza-Formigoni, de, 2013). The authors follow describing how Item Response Theory (IRT) first applications stood out in the academic field at the 50ths of the past century (Sartes e Souza-Formigoni, de, 2013). Despite its origins remains to the 19$^{th}$ century on the field of statistics as a coherent methodological system instead of being a theory (Bock e Gibbons, 2021). According to the authors, IRT 'accepts the idea that a discrete behavioral response to a set task, object, or proposition (in short, to an *item*) is the expression of a stochastic mechanism that can be modeled by an unobservable random variable and a threshold' (p. 8). This relationship is usually depicted with an item characteristic curve (ICC), which denotes the probability of getting the right answer (for dichotomous items), bearing the person ability (Rindskopf, 2015). Several IRT models were developed (e.g. 2-PL, Graded Response Model etc.) to account for distinct types of data. The number of item parameters estimated depends on the

model selected. It usually encompasses the: difficulty (*b*), discrimination (*a*), and pseudo-likelihood (*c*) parameters (Bock e Gibbons, 2021). As IRT separate item parameters estimations from population ability, it surpasses CTT sample-dependency, constituting one of its main advantages. However, several studies point out that IRT model estimates are biased in the presence of large amounts of missing data (Enciso, 2016; Enders, 2013; Rose, Davier, von e Xu, 2010; Waterbury, 2019). In addition to that, IRT multidimensional modelling is still under debate given its mathematical complexity compared to unidimensional modelling (Bock e Gibbons, 2021).

Alternatively, Network Psychometrics is derived from the network approaches, which conceptualize a given phenomena as an emergent property generated through the interaction of components (Borsboom *et al.*, 2022). Oppositely to the aforementioned Common Causal Model, the Network perspective applied to psychopathological field would state that a given disorder (e.g. Major Depression) results from the causal interplay between symptoms. In other words, Major Depression Disorder would be conceptualized as a system resulting from the mutual interaction, often reciprocally reinforced, of its elements (i.e. symptoms) (Borsboom e Cramer, 2013). In his textbook, Borsboom et al. (2022) alerts to the fact that network approaches "do not require a particular physical structure, but rather require the applicability of a mode of representation, and therefore the question whether a domain 'really' is a network is often moot" (p. 12).

The development of Network theory applied to the study of psychopathology, and other areas of interest of Psychology, was accompanied by Network Psychometrical models developments. In Network models, nodes represent variables of a given phenomena and edges between nodes are statistical relations between these variables. A hypothetical depression network structure composed of 21 items of the Beck Depression Inventory 2nd Edition is depicted in figure 2.

Figure 2 - A Hypothetical Network Structure of BDI-II




In which nodes are represented by circles and their interrelationships are shown through blue edges. The blue color usually means positive associations among the nodes. Network structures like that are usually generated through algorithms that use (partial) correlations coefficients, like Pairwise Markov Random Field (PMRF).

Laying the epistemological differences aside, a few comparisons of Network Analysis, Factor Analysis and IRT models generate similar results to item estimates, as pointed out by some studies (Christensen, Golino e Silvia, 2020; Machado, Vissoci e Epskamp, 2015). These findings contribute to refreshing the relevance of substantive features on validity investigation.

# II. OBJECTIVES

In line with theoretical background, the current thesis was conceived upon two main emphases. The first emphasis is related to the application of psychometrical models to the study of different constructs. The second one consists of exploring emergent methodological challenges on the development of new tools in childhood development.

The first emphasis was addressed on three empirical studies (#1, #2, and #3), aiming at:

- To investigate the invariance of the Beck Depression Inventory-II and Beck Anxiety Inventory in a representative sample of undergraduate students from Spain, Portugal, and Brazil.
- To present preliminary psychometric evidence for the Penibility Assessment Inventory in the context of emergency services in a public health unit in the city of Rio de Janeiro.
- To explore the psychometric structure of the Environmental Screening Questionnaire – Research Edition using a network analysis approach.

The second emphasis was explored in a theoretical study (#4), aiming at:

- Providing a brief debate about the current challenges on the use of the Item Response Theory within large amounts of data in the developmental field.

# III. ARTICLE SELECTION

**Article 1**

**DO NASCIMENTO, Rodrigo Leão Ferreira; FAJARDO-BULLON, Fernando; SANTOS, Eduardo; LANDEIRA-FERNANDEZ, J.; ANUNCIAÇÃO, Luis.** Psychometric properties and cross-cultural invariance of the Beck Depression Inventory-II and Beck Anxiety Inventory among a representative sample of Spanish, Portuguese, and Brazilian undergraduate students. Published in the *International Journal of Environmental Research and Public Health*, v. 20, n. 11, p. 6009, 2023. Available in: https://doi.org/10.3390/ijerph20116009.

## Abstract

Clinical psychologists often use the Beck Depression Inventory, 2nd edition (BDI-II), and Beck Anxiety Inventory (BAI) to aid in the diagnosis of mental health issues and verify the effectiveness of treatments. Despite this common practice, studies that implement a cross-cultural design to check psychometric properties and the invariance of these scales are still scarce in the literature, which can lead to biased results that prevent comparisons among different groups. The present study investigated the internal structure of both tools and their level of invariance. From a representative sample of undergraduate students from Spain ($n$ = 1216), Portugal ($n$ = 426), and Brazil ($n$ = 315), Confirmatory Factor Analysis and Multigroup Confirmatory Factor Analysis were performed. The results revealed suitable fit indices for the two-factor structure of the BDI-II and BAI, assessed by Confirmatory Factor Analysis procedures. Additionally, the two-factor model of the BDI-II reached invariant properties at three levels, whereas the structural model of the BAI did not. Altogether, these results suggest using the BDI-II in this group in these three countries and imply that BAI scores should be interpreted cautiously.



## Keywords

measurement invariance; depression; anxiety; Multigroup Confirmatory Factor Analysis

# 1. Introduction

Depression and anxiety are clinically important disorders with a high prevalence among psychiatric conditions worldwide (GBD, 2022). With the emergence of the COVID-19 pandemic, the prevalence increased to 246 million (3153 cases per 100,000) for major depressive disorder and 374 million (4802 per 100,000) for anxiety disorders (Santomauro *et al.*, 2021). According to the World Mental Health Report (World Health Organization, 2022) (WHO Team, 2022), high- and low-income countries are affected by mental health conditions, especially women and younger groups. Different actions must be taken to address this situation, including clinical assessments that are supported by tools with adequate psychometric properties.

Among the different instruments that are used to aid the diagnosis of depression and anxiety, the Beck Depression Inventory, 2nd edition (BDI-II), and Beck Anxiety Inventory (BAI) are widely cited in the literature (Bagheri *et al.*, 2021; Ganji *et al.*, 2022; Wang e Gorenstein, 2013). In 1961, Dr. Aaron T. Beck and his colleagues developed the initial version of the BDI. This instrument underwent a reformulation in 1978. Subsequently, in 1996, the second edition of the BDI (BDI-II) was introduced, which incorporated symptoms that were outlined in the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) (Wang e Gorenstein, 2013). The current study employed factor analysis to examine the underlying structure of the BDI-II in two separate samples: outpatients and college students. The results revealed the presence of two factors in both samples. The first group of participants yielded factors that referred to “Somatic-Affective” and “Cognitive”, and the second group yielded factors that referred to “Cognitive-Affective” and “Somatic.” The authors of the study proposed that affective items, such as “Sadness” and “Crying”, that are present in the BDI-II may vary relative to the characteristics of the sample. Overall, however, the factorial solution of two positively correlated dimensions (Cognitive-Affective and Somatic) was proposed as the most fitting representation of the data (Steer e Clark, 1997).

Widespread global acceptance of the BDI-II has led to a growing interest in investigating its factorial structure, which has been found to vary depending on the type of sample and the number of factors that are considered (Dere *et al.*, 2015;

Faro e Pereira, 2020). A recent study that was conducted with a community population revisited some prominent factorial models of the BDI-II, examining solutions that ranged from two-dimensional to bi-factorial, with a further consideration of three-dimensional models. The diversity of factorial solutions that are obtained in the current study highlights the challenge of determining an optimal and generalizable model for the BDI-II (Faro e Pereira, 2020). Additionally, the complexity of the issue is further compounded by the fact that depressive symptoms appear to differ across cultures. Studies have been inconclusive with regard to whether depression is perceived as more psychological (cognitive) in nature among Western individuals compared with Eastern individuals, who tend to view it as more somatic in nature. This observation highlights the need to consider cultural variations when examining the factorial structure of the BDI-II and other measures of depression (Dere *et al.*, 2015; Dunlop *et al.*, 2020).

Anxiety symptoms are evaluated by different instruments, mainly self-report items. Nevertheless, only the seminal article by Beck, Epstein, Brown, and Steer (1988) (Beck *et al.*, 1988) achieved discriminant validity that was able to differentiate anxiety from depressive symptoms. In their study, the BAI exhibited high reliability ($\alpha = 0.92$). In 1993, Beck and Steer conducted a revision and re-examination of the evidence, proposing a two-dimensional factorial structure, referred to as "Somatic Symptoms" and "Subjective Affective and Panic Symptoms", for the instrument in question. Since then, the instrument has gained significant popularity, leading to numerous studies that investigated its factorial structure (Bardhoshi, Duncan e Erford, 2016; Chapman *et al.*, 2009; Liang, Wang e Zhu, 2018; Osman *et al.*, 2002). A review by Bardoshi and colleagues (2016) on the psychometric properties of the instrument found evidence that supports a two-factor structure, although the fit indices did not necessarily meet the recommended cut-off points in the literature. Additionally, this study revealed that the instrument had high levels of internal consistency for nonclinical samples ($\alpha = 0.91$), indicating its good reliability (Bardhoshi, Duncan e Erford, 2016). Nonetheless, there is no unanimity in the literature about the factorial composition of the BAI. Some studies suggest a general factor, whereas others propose a multidimensional model (Bardhoshi, Duncan e Erford, 2016; Magán *et al.*, 2008; Osman *et al.*, 2002; Quintão, Delgado e Prieto, 2013). Additionally, the suitability of the lexicon that is employed for BAI

items has been contested. A recent study that examined Latinos who live in the United States, for example, raised concerns about the various terms that are employed in the BAI that may be interpreted distinctively among the diverse ethnic groups that are represented within that population (Benuto *et al.*, 2020). Consequently, although the BAI and BDI have robust psychometric properties and are extensively utilized, factorial and cultural concerns continue with regard to their application.

The extensive use of these instruments with clinical and nonclinical populations in different countries has further revealed their importance but also led to concerns about measurement invariance. According to Dere et al. (2015) (Dere *et al.*, 2015), many health professionals assume they are assessing the same construct across populations in the same way. They assume that the measure that is assessed is invariant across groups. However, this assumption must be verified. There is a scarcity of studies on the psychometric properties of these instruments in developing countries (e.g., Brazil) in general (Anunciação *et al.*, 2022). Fostering further studies in this area is important. Only two studies that assessed the invariance of the BDI-II were found for Brazilian samples (Faro e Pereira, 2020; Silva, Wendt e Argimon, 2018). We found no study that assessed the invariance of the BAI for a similar group. Given the significance of the topic and ongoing debate about the factorial structure of the instruments in question, the present study investigated the initial models that were proposed by Beck using Confirmatory Factor Analysis (CFA) techniques (multigroup). The assessment of clinical models with nonclinical populations has been investigated in the literature (Abrams, Carleton e Asmundson, 2007). Furthermore, gathering evidence about the factorial structure of the instrument may be relevant to explore a vast clinical symptomatology with a vulnerable group. Thus, to explore the generalizability of the two-factor solution that was identified by the author across various cultures, we adopted the factorial model that was identified for the clinical patient sample while utilizing the BDI-II. By conducting a thorough examination of these models, the present study provides further evidence of the psychometric properties of these instruments, thereby advancing the ongoing debate on the topic. The present study investigated the invariance of the BDI-II and BAI in a representative sample of undergraduate students from Spain, Portugal, and Brazil.

# 2. Materials and Methods

## 2.1. Participants and Procedures

The data were collected in three different settings (University of Extremadura, Spain; University of Coimbra, Portugal; Pontifical Catholic University of Rio de Janeiro [PUC-Rio], Brazil) in 2015. Stratified probability sampling was performed, which makes the total participants representative of the three institutions. More details about the sampling procedures can be found elsewhere (Junior *et al.*, 2020).

The final analytic sample consisted of 1957 undergraduate students. Of these, 62.1% (n = 1210) were from the University of Extremadura (Spain), 21.8% (n = 426) were from the University of Coimbra (Portugal), and 16.1% (n = 293) were from PUC-Rio (Brazil). The mean ages were 21.49 years (SD = 3.02 years) for the Spanish students, 20.43 years (SD = 1.66 years) for the Portuguese students, and 22.83 years (SD = 7.18 years) for the Brazilian students.

A team of trainees and previously trained professionals performed the data collection. To control for eventual bias, the questionnaires were completed between school exam periods. All participants were informed about the study objectives, and their further questions were answered by the research team. Ethical approval was obtained from the local Research Ethics Committees in the respective countries.

## 2.2. Measures

### 2.2.1. Beck Depression Inventory-II

The BDI (Beck *et al.*, 1961) has been extensively used to assess and screen for depressive symptoms in nonclinical and clinical populations (Bardhoshi, Duncan e Erford, 2016). With the release of the DSM-III-R and DSM-IV, the authors revised the BDI, producing the 2nd edition (BDI-II; Beck, A. T., Steer, R. A., & Brown, 1996). Since then, many reviews have shown its good psychometric properties (Beck *et al.*, 1961; Beck, A. T., Steer, R. A., & Brown, 1996; Faro e Pereira, 2020; Huang e Chen, 2015; Junior *et al.*, 2020; Wang e Gorenstein, 2013). For the BDI-II, participants rate different symptoms on a Likert-type scale, ranging from 0 to 3. A total score is calculated by summing the ratings for all 21 items. This instrument has presented evidence of validity from studies in Spain (Magán *et al.*,

2008), Portugal (Campos e Gonçalves, 2011), and Brazil (Cunha, 2001). These studies showcased optimal results, such as the α for the BDI-II equals 0.89 (95% confidence interval [CI] = 0.89–0.90).

### 2.2.2. Beck Anxiety Inventory

The BAI (Beck et al., 1988) (Beck *et al.*, 1988) is a self-report measure that is recommended for different populations and has good psychometric properties (Han-Kyeong Lee *et al.*, 2016). It consists of 21 items that assess the severity of anxiety symptoms on a Likert-type scale, ranging from 0 to 3. Studies have shown that the instrument has evidence of validity in Brazil (Cunha, 2001), Portugal (Quintão, Delgado e Prieto, 2013), and Spain (Magán *et al.*, 2008). Similar to a previous study (Junior *et al.*, 2020), the α for the BAI was 0.90 (95% CI: 0.90–0.91).

## 2.3. Statistical Plan

The data analysis was conducted using a two-fold method. First, CFA was performed by assessing different BDI-II and BAI models (Beck, A. T., Steer, R. A., & Brown, 1996). A $\chi 2$ test (Satorra, 2000) was used to explore fit index differences between the tested models. We also conducted a Multi-Group Factorial Analysis (MGCFA) to check whether invariance was achieved within the selected models in the different countries.

We performed two factorial models for each instrument, checking unidimensional and two-factor structures for the instruments. Because of the ordinal nature of the data, the psychometric analyses were based on the Robust Diagonally Weighted Least Squares as the estimator (Li, 2016).

The threshold values to assess goodness of fit were the following: Comparative Fit Index (CFI) and Tucker–Lewis index (TLI) greater than 0.90 or 0.95 and Root Mean Square Error of Approximation (RMSEA) less than 0.08 or 0.06, with the upper limit of the CI less than 0.10 (Brown, 2015). In accordance with previous literature (Brown, 2015), a $\chi 2$ test was conducted to check significant differences in fit indices between the factorial solutions.

The reliability of the selected models was estimated by the ordinal Cronbach alpha (α) and McDonald omega (ϖ) (Gana e Broc, 2019).

In the second part of the data analysis, an MGCFA was used to test the invariance of the BDI-II and BAI for Brazilian, Spanish, and Portuguese undergraduate students. We used the same aforementioned estimation method and fit indices, with the inclusion of the ΔCFI (less than 0.01) as recommended by Cheung and Rensvold (2002) (Cheung e Rensvold, 2002). The MGCFA assessed three levels of invariance: configural (factor structure), metric (loadings), and scalar (intercept). R Studio 4.02 software was used for all analyses, with the lavaan package (version 0.6.14).

# 3. Results

## 3.1. Preliminary Analyses

### 3.1.1. Descriptive Analyses

Due to their skewness (SI) and kurtosis (KI) indicators surpassing the corresponding thresholds, typically greater than 2 and 7, respectively, for many items (Kline, 2015), BDI-II and BAI scores were non-normally distributed (Table 1). The missing-value inspection did not reveal any inconsistencies (approximately < 1.2%) for all items of both instruments.

PLEASE INSERT TABLE 1 HERE

### 3.1.2. Sample Characteristics

A sample of 1957 participants was assessed in Spain (62.1%), Portugal (21.8%), and Brazil (16.1%). The mean age of the participants was 21.5 years (SD = 3.8 years), but significant differences were found between the three countries ($F$ 2,1926 = 35.19, $p < 0.001$). Significant differences were found in the proportions of males and females between Portuguese and Brazilian students compared with Spanish students ($\chi 21 = 161$, $p < 0.001$). The values were similar to the results of (Junior *et al.*, 2020) (See Table 2).

PLEASE INSERT TABLE 2 HERE

## 3.2. Evidence of Internal Structure

### 3.2.1. BDI-II Results

The unifactorial model did not show satisfactory fit indices. The $\chi 2$ test value was significant ($\chi 2189 = 913.126$, $p < 0.001$). The RMSEA values were adequate (0.045; 90% CI = 0.042–0.048). The CFI (0.787) and TLI (0.764) values were not adequate. Different results were achieved with the two-factor model (Figure 1). The factor loadings values are shown in Table 3. The $\chi 2$ test value was significant ($\chi 2210 = 3613.885$, $p < 0.001$). The RMSEA values were adequate (0.037; 90% CI = 0.034–0.040). The CFI (0.855) and TLI (0.838) values were slightly below the recommended values (Table 4).

PLEASE INSERT FIGURE 1 HERE

PLEASE INSERT TABLE 3 HERE

PLEASE INSERT TABLE 4 HERE

The goodness-of-fit indices are presented in Table 4. The results of the $\chi 2$ test (Satorra, 2000) revealed that the two-factor model better fit the data compared with the unifactorial solution ($\chi 21 = 125.54$, $p < 0.001$), which was previously expected. The reliability of the two-factor model revealed adequate values for both factors of the instrument: "Cognitive" ($\alpha = 0.81$, $\varpi = 0.82$) and "Somatic-Affective" ($\alpha = 0.84$, $\varpi = 0.84$).

### 3.2.2. BAI Results

The BAI results partially followed the same pattern as in the previous analyses. The $\chi 2$ test value was significant ($\chi 2189 = 1026.066$, $p < 0.001$). The RMSEA values

were adequate (0.049; 90% CI = 0.046–0.051). The CFI (0.777) and TLI (0.753) values were outside the current range adopted for cut-offs. Similar results were found for the two-factor model (Figure 1). The $\chi 2$ test value was significant ($\chi 2188 = 971.658$, $p < 0.001$). The RMSEA values were adequate (0.47; 90% CI = 0.044–0.050). The CFI and TLI values were not close to the prescribed values (0.792 and 0.767, respectively). The factor loadings are shown in Table 5.

PLEASE INSERT TABLE 5 HERE

The goodness-of-fit indices are presented in Table 6. The $\chi 2$ test (Satorra, 2000) showed better results for the two-factor model compared with the unifactorial structure ($\chi 21 = 26.854$, $p < 0.001$). The reliability analysis of the two-factor model revealed adequate results for both factors. For the "Somatic Symptoms" factor, the values were $\alpha = 0.84$ and $\varpi = 0.84$. For the "Subjective Affective and Panic Symptoms" factor, the values were $\alpha = 0.82$ and $\varpi = 0.83$.

PLEASE INSERT TABLE 6 HERE

### 3.3. Measurement Invariance

We performed an MGCFA to evaluate the invariance of the two-factor models of the BDI-II and BAI for Brazilian, Spanish, and Portuguese undergraduate students. The goodness-of-fit indices are presented in Table 7. The BDI-II model exhibited good indices at different invariance levels. The $\chi 2$ test was significant at the configural, metric, and scalar levels ($p < 0.001$). The RMSEA was 0.022 (CI = 0.018–0.026). The CFI and TLI values were 0.942 and 0.943, respectively. The $\Delta$CFI (0.012) was slightly above the adopted cut-off point (0.01) [26]. Contrary to the previous results, the two-factor model of the BAI did not show invariance at any level. The $\chi 2$ test was significant at the configural, metric, and scalar levels ($p < 0.001$). The RMSEA was 0.035 (CI = 0.031–0.038). The CFI and TLI values were 0.863 and 0.865, respectively. The $\Delta$CFI value was −0.039 at the scalar level.

PLEASE INSERT TABLE 7 HERE

## 4. Discussion

The main goal of the present study was to investigate the invariance of the BAI and BDI-II. This study was performed with a representative sample of undergraduate students from Brazil, Portugal, and Spain. All of the estimated BDI-II models showed good psychometric properties, such as good reliability values and good standard fit indices. The examined factorial models of the BAI did not demonstrate the anticipated outcomes. After comparing different solutions for each instrument, the two-factor models were sorted. These models exhibited high reliability levels, described by many coefficient estimators. Additionally, the BDI-II showed invariant results at three different levels: configural, metric, and scalar. In contrast to the aforementioned findings, the results of the BAI analysis were not consistent across the three levels of measurement (i.e., configural, metric, and scalar) and thus did not demonstrate invariance properties among the respective samples.

Several confirmatory models (unidimensional and two-factor by country and with the pooled sample) were performed to assess the BDI-II. The results of the factorial models revealed favorable fit indices. In cases in which multiple models are found to be suitable, a challenge arises in terms of determining the most appropriate factorial solution. A thorough examination of the factor loadings across countries and in the total sample revealed that all items exhibited loadings greater than 0.40 on their respective factors, with the exception of the “Loss of sexual interest” and “Feeling of punishment” items for the Brazilian sample (0.33 and 0.36, respectively) and “Suicidal thoughts” item for the Spanish (0.36) and Portuguese (0.34) samples. Altogether, these accumulated results indicated a robust factor structure. An analysis of the loadings for specific items revealed some discrepancies in factor loadings across different countries. Specifically, the “Loss of sexual interest” item in the “Somatic-Affective” factor displayed notable variability in loading, with a correlation of 0.33 for the Brazilian sample and correlations that

ranged from 0.46 to 0.50 for the Spanish and Portuguese samples. Concerns about the lack of salience in this item were raised in the meta-analysis that was conducted by Huang e Chen (2015) for the total sample and some study subgroups. Furthermore, the "Indecision" item exhibited divergent characteristics across countries. In Brazil, the factor loading was 0.50, whereas the respective factor loadings were 0.61 and 0.62 in Spain and Portugal. Additionally, the "Irritability" item displayed a lower correlation in the Brazilian sample (0.49) compared with the Spanish and Portuguese samples (0.60). Furthermore, the "Feeling of punishment" item in the "Cognitive" factor obtained a correlation of 0.36 in the Brazilian sample, 0.50 in the Spanish sample, and 0.47 in the Portuguese sample. Similarly, the "Suicidal thoughts" item displayed disparate behavior among the three countries. In Brazil, the factor loading was 0.55, whereas the corresponding factor loadings were 0.36 and 0.34 in Spain and Portugal, respectively. Altogether, this highlights the need for further examination and analysis to make a definitive selection (Vandekerckhove, Matzke e Wagenmakers, 2014).

Although there were discrepancies in factor loadings of specific items in the BDI-II, their values for the overall sample were prominent, with all items exhibiting a factor loading greater than 0.40. Moreover, the discussion that was proposed by a recent meta-analysis (Huang e Chen, 2015) about inconsistent findings for the factorial model structure of the BDI-II was extensively considered. In this previous study, the results supported the two-correlated-factor model that was proposed by Beck et al. (1996) (Kline, 2015). Thus, a $\chi 2$ test (Satorra, 2000) was performed to investigate differences between our models ($p < 0.001$). Considering such results, the two-factor solution was selected in our study. Reliability analyses were performed for the different factors ("Somatic-Affective" and "Cognitive"), supporting the internal consistency of the instrument.

A thorough examination of factor loadings across countries and in the total sample also revealed that all items of the BAI exhibited loadings greater than 0.40 on their respective factors, with the exception of the "Numbness or tingling" item for the Brazilian sample (0.32), "Face flushed" item for the Spanish sample (0.38), and "Fear of dying" item for the Spanish (0.32), Portuguese (0.32), and overall (0.35) samples. The analysis of loadings for specific items in the "Somatic Symptoms" factor revealed notable discrepancies, with the "Numbness or tingling" item

displaying a correlation of 0.32 for the Brazilian sample, 0.53 for the Spanish sample, and 0.58 for the Portuguese sample.

Additionally, in the "Subjective Affective and Panic Symptoms" factor, the "Unable to relax" item obtained a lower correlation in the Brazilian sample (0.55) compared with the Spanish and Portuguese samples (0.67), and the "Fear of dying" item obtained a correlation of 0.48 in the Brazilian sample and 0.32 in the Spanish and Portuguese samples. Regarding this last result for the Spanish and Portuguese samples, no specific discussion was found for this item in the consulted literature. However, sociocultural factors related to the understanding of the expression "fear of dying" may be associated with that. Added to this, given that Brazil has higher rates of anxiety in its population compared with other countries, more pronounced symptoms of anxiety may manifest more acutely in this population, which may account for these discrepancies in factor loadings (WHO Team, 2022).

Divergent results were achieved with the CFA of the BAI, in which the unifactorial and two-factor solutions did not present good fit indices. The procedure that was adopted to sort the most appropriate model was the same as the aforementioned procedure. Again, the $\chi 2$ test result was significant ($p < 0.001$). Consistent with the findings of a recent meta-analysis with different psychometric properties of the BAI (Han-Kyeong Lee *et al.*, 2016). These authors found no evidence to support the factorial validity of the model that was proposed by Beck and Steer (1993) (Beck, A. T., & Steer, 1993), with CFI values that ranged from 0.69 to 0.90 and RMSEA values that ranged from 0.04 to 0.14. However, the authors included only studies that analyzed English language versions of the BAI.

An MGCFA was conducted to assess the invariance of the two-factor solutions of the BDI-II and BAI for Brazilian, Spanish, and Portuguese undergraduate students. We present results from the lavaan package (version 0.6.14). A parsimonious examination of the characteristics of the models, together with a statistical summary, may facilitate an understanding of the invariant behavior of the BDI-II across configural, metric, and scalar levels. This interpretation is primarily supported by the optimal distribution of residuals as measured by the RMSEA in the configural model, which yielded an interval of 0.031 (0.027–0.035) and CFI value of 0.898, which are at the limit of acceptability. Additionally, the TLI value

was marginally lower than what is currently recommended at 0.886. Furthermore, the CFI and TLI values increased significantly in the metric model (CFI = 0.954, TLI = 0.952), with a decrease in the RMSEA index = 0.20 (0.015–0.025). The behavior of these indices remained consistent, approaching 0.950 in the scalar model (CFI = 0.942, TLI = 0.943, RMSEA = 0.022 [0.018–0.026]).

The present study yielded divergent results for the BAI, in which it did not exhibit invariant behavior across any of the models at the scaled evaluation level. Specifically, at the configural level, the CFI and TLI values were 0.833 and 0.814, respectively, although the RMSEA value was within an acceptable range (0.041 [0.037–0.044]). These findings suggest instability in the factorial structure and other psychometric properties of the BAI among the countries that were included in the sample for the selected model.

With regard to the limitations of the present study, the research was performed among undergraduate students in three countries, thereby biasing the sampling. Additionally, due to the cross-sectional design that was adopted in the original study (Junior *et al.*, 2020), it is not possible to understand the temporal changes of the phenomenon studied, which would be better verified through longitudinal investigations (Widaman, Ferrer e Conger, 2010). Another potential limitation was the lack of measuring bi-factor models of the BDI-II and invariance properties to gender, as proposed in a recent study (Faro e Pereira, 2020). One partial limitation of our analyses was the way we dealt with items of the measures. Because of convergence issues, we treated all items as continuous indicators. However, we used the DWLS estimator, which the current literature also recommends within the categorical analysis framework. Finally, we did not run an EFA to check which latent structure best fits the data collected, which will be carried out in future studies. Altogether, these limitations should be considered when attempting to generalize our findings.

With the COVID-19 pandemic, an increase in the prevalence of depression and anxiety has been detected (Santomauro *et al.*, 2021). Different degrees of anxiety and depression disorders are associated with extensive impairments in cognition that compromise daily living activities (Castaneda *et al.*, 2008). Tools are needed with robust psychometric properties. Our findings demonstrated that the BDI-II is

suitable for clinical use, and its scoring can be considered invariant even between participants in different countries, such as Brazil, Spain, and Portugal.

To our knowledge, this is the first study that evaluated the invariance of the BDI-II and BAI among undergraduate students in Brazil, Portugal, and Spain. As discussed previously, because of the scarcity of studies on the psychometric properties of instruments in developing countries (e.g., Brazil), our findings will contribute to further development in this area. Our study also provided further evidence of the factorial structure of these popular and important instruments. Finally, to advance future research in the field, we strongly recommend investigating the invariance of these instruments with different factorial structures (e.g., bi-factor models) and genders.

## 5. Conclusions

The present study assessed the invariance of the BDI-II and BAI among undergraduate students from three different countries. The analyses revealed good psychometric properties of the BDI-II, whose invariance was fully supported by our analysis, whereas the BAI did not show similar results. The present results add to the evidence of the factor structure of these instruments.

## 6. References

ABRAMS, M. P.; CARLETON, R. N.; ASMUNDSON, G. J. G. An exploration of the psychometric properties of the PASS-20 with a nonclinical sample. **The journal of pain**, v. 8, n. 11, p. 879–886, nov. 2007.

ANUNCIAÇÃO, L. *et al.* An Exploratory Analysis of the Internal Structure of Test Through a Multimethods Exploratory Approach of the ASQ:SE in Brazil. **Journal of Neurosciences in Rural Practice**, v. 13, n. 2, p. 186, 1 abr. 2022.

BAGHERI, Z. *et al.* Assessing the measurement invariance of the 10-item Centre for Epidemiological Studies Depression Scale and Beck Anxiety Inventory questionnaires across people living with HIV/AIDS and healthy people. **BMC Psychology**, v. 9, n. 1, p. 1–11, 1 dez. 2021.

BARDHOSHI, G.; DUNCAN, K.; ERFORD, B. T. Psychometric Meta-Analysis of the English Version of the Beck Anxiety Inventory. **Journal of Counseling and Development**, v. 94, n. 3, p. 356–373, 1 jul. 2016.

BECK, A. T. *et al.* An inventory for measuring depression. **Archives of general psychiatry**, v. 4, n. 6, p. 561–571, 1961.

BECK, A. T. *et al.* An inventory for measuring clinical anxiety: psychometric properties. **Journal of consulting and clinical psychology**, v. 56, n. 6, p. 893–897, 1988.

BECK, A. T., & STEER, R. A. **Beck Anxiety Inventory manual**. San Antonio, TX: Psychological Corporation, 1993.

BECK, A. T., STEER, R. A., & BROWN, G. K. **Manual for the Beck Depression Inventory**. 2nd. ed. San Antonio, TX: [s.n.].

BENUTO, L. T. *et al.* A confirmatory factor analysis of the beck anxiety inventory in Latinx primary care patients. **International Journal of Mental Health**, v. 49, n. 4, p. 361–381, 1 out. 2020.

BROWN, T. **Confirmatory for Analysis for Applied Research**. [s.l.] Guilford Publication, 2015.

CAMPOS, R. C.; GONÇALVES, B. The Portuguese version of the Beck Depression Inventory-II (BDI-II): Preliminary psychometric data with two nonclinical samples. **European Journal of Psychological Assessment**, v. 27, n. 4, p. 258, 2011.

CASTANEDA, A. E. *et al.* A review on cognitive impairments in depressive and anxiety disorders with a focus on young adults. **Journal of Affective Disorders**, v. 106, n. 1–2, p. 1–27, 1 fev. 2008.

CHAPMAN, L. K. *et al.* A confirmatory factor analysis of the Beck Anxiety Inventory in African American and European American young adults. **Journal of Anxiety Disorders**, v. 23, n. 3, p. 387–392, abr. 2009.

CHEUNG, G. W.; RENSVOLD, R. B. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. **https://doi.org/10.1207/S15328007SEM0902_5**, v. 9, n. 2, p. 233–255, 2002.

CUNHA, J. A. **Manual da versão em português das Escalas de Beck**. [s.l.] Casa do Psicólogo, 2001.

DERE, J. *et al.* Cross-cultural examination of measurement invariance of the Beck Depression Inventory-II. **Psychological assessment**, v. 27, n. 1, p. 68–81, 1 mar. 2015.

DUNLOP, B. W. *et al.* Somatic symptoms in treatment-naïve Hispanic and non-Hispanic patients with major depression. **Depression and Anxiety**, v. 37, n. 2, p. 156–165, 1 fev. 2020.

FARO, A.; PEREIRA, C. R. Factor structure and gender invariance of the Beck Depression Inventory–second edition (BDI-II) in a community-dwelling sample of adults. **Health Psychology and Behavioral Medicine**, v. 8, n. 1, p. 16–31, 1 jan. 2020.

GANA, K.; BROC, G. **Structural Equation Modeling with lavaan**. [s.l: s.n.].

GANJI, K. K. *et al.* COVID-19 and stress: An evaluation using Beck's depression and anxiety inventory among college students and faculty members of Jouf University. **Work (Reading, Mass.)**, v. 72, n. 2, p. 399–407, 2022.

GBD. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. **The Lancet Psychiatry**, v. 9, n. 2, p. 137–150, 1 fev. 2022.

HAN-KYEONG LEE *et al.* Psychometric Properties of the Beck Anxiety Inventory in the Community-dwelling Sample of Korean Adults. **Korean Journal of Clinical Psychology**, v. 35, n. 4, p. 822–830, 2016.

HUANG, C.; CHEN, J. H. Meta-Analysis of the Factor Structures of the Beck Depression Inventory–II. **Assessment**, v. 22, n. 4, p. 459–472, 2015.

JUNIOR, A. A. *et al.* Depression and Anxiety Symptoms in a Representative Sample of Undergraduate Students in Spain, Portugal, and Brazil. **Psicologia: Teoria e Pesquisa**, v. 36, p. 1–7, 25 nov. 2020.

KLINE, R. B. **Principles and Practice of Structural Equation Modeling, Fourth Edition - Rex B. Kline - Google Books**. Fifth ed. [s.l.] The Guilford Press, 2015.

LI, C. H. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. **Behavior research methods**, v. 48, n. 3, p. 936–949, 1 set. 2016.

LIANG, Y.; WANG, L.; ZHU, J. Factor structure and psychometric properties of Chinese version of Beck Anxiety Inventory in Chinese doctors. **Journal of Health Psychology**, v. 23, n. 5, p. 657–666, 1 abr. 2018.

MAGÁN, INÉS *et al.* Psychometric Properties of a Spanish Version of the Beck Anxiety Inventory (BAI) in General Population. **The Spanish Journal of Psychology**, v. 11, n. 2, p. 626–640, 2008.

OSMAN, A. *et al.* Factor structure, reliability, and validity of the Beck Anxiety Inventory in adolescent psychiatric inpatients. **Journal of clinical psychology**, v. 58, n. 4, p. 443–456, 2002.

QUINTÃO, S.; DELGADO, A. R.; PRIETO, G. Validity study of the Beck Anxiety Inventory (Portuguese version) by the Rasch Rating Scale model. **Psicologia: Reflexão e Crítica**, v. 26, n. 2, p. 305–310, 2013.

SANTOMAURO, D. F. *et al.* Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. **The Lancet**, v. 398, n. 10312, p. 1700–1712, 6 nov. 2021.

SATORRA, A. Scaled and Adjusted Restricted Tests in Multi-Sample Analysis of Moment Structures. p. 233–247, 2000.

SILVA, M. A. DA; WENDT, G. W.; ARGIMON, I. I. DE L. Inventário de depressão de beck II: análises pela teoria do traço latente. **Revista Avaliação Psicológica**, v. 17, n. 03, p. 339–350, 15 ago. 2018.

STEER, R. A.; CLARK, D. A. Psychometric characteristics of the beck depression inventory-II with college students. **Measurement and Evaluation in Counseling and Development**, v. 30, n. 3, p. 128–136, 1997.

VANDEKERCKHOVE, J.; MATZKE, D.; WAGENMAKERS, E.-J. Model comparison and the principle of parsimony. **Oxford Handbook of**, p. 1–29, 2014.

WANG, Y. P.; GORENSTEIN, C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. **Revista brasileira de psiquiatria (Sao Paulo, Brazil : 1999)**, v. 35, n. 4, p. 416–431, 2013.

WHO TEAM. World Mental Health Report: Transforming Mental Health For All - Executive Summary. p. 1–28, 2022.

WIDAMAN, K. F.; FERRER, E.; CONGER, R. D. Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. **Child development perspectives**, v. 4, n. 1, p. 10, 4 abr. 2010.

Table 1 - Descriptive statistics of the BDI-II and BAI

| Item (BAI) | Mean | SD | Skewness | Kurtosis | Item (BDI-II) | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| 1. Numbness or tingling | 0.41 | 0.63 | 1.39 | 1.17 | 1. Sadness | 0.25 | 0.53 | 2.36 | 6.02 |
| 2. Feeling hot | 0.69 | 0.75 | 0.74 | −0.29 | 2. Pessimism | 0.54 | 0.69 | 1.26 | 1.67 |
| 3. Wobbliness in legs | 0.34 | 0.63 | 1.76 | 2.30 | 3. Feeling of failure | 0.30 | 0.57 | 2.02 | 4.15 |
| 4. Unable to relax | 0.89 | 0.89 | 0.66 | −0.51 | 4. Feeling of guilt | 0.37 | 0.59 | 1.50 | 2.08 |
| 5. Fear of worst happening | 0.69 | 0.90 | 1.06 | 0.00 | 5. Feeling of guilt | 0.49 | 0.61 | 1.13 | 1.65 |
| 6. Dizzy or lightheaded | 0.28 | 0.61 | 2.23 | 4.29 | 6. Feeling of punishment | 0.16 | 0.51 | 3.87 | 16.61 |
| 7. Heart pounding/racing | 0.34 | 0.68 | 1.99 | 3.29 | 7. Disconformity with oneself | 0.39 | 0.70 | 1.93 | 3.33 |
| 8. Unsteady | 0.36 | 0.65 | 1.86 | 3.04 | 8. Self-criticism | 0.79 | 0.72 | 0.71 | 0.37 |
| 9. Terrified or afraid | 0.20 | 0.53 | 2.88 | 8.19 | 9. Suicidal thoughts | 0.07 | 0.32 | 5.68 | 40.13 |
| 10. Nervous | 1.04 | 0.88 | 0.50 | −0.49 | 10. Crying | 0.34 | 0.69 | 2.05 | 3.52 |
| 11. Feeling of choking | 0.22 | 0.59 | 2.85 | 7.81 | 11. Agitation | 0.54 | 0.69 | 1.32 | 1.99 |
| 12. Hands trembling | 0.30 | 0.63 | 2.19 | 4.36 | 12. Loss of interest | 0.43 | 0.62 | 1.46 | 2.37 |
| 13. Shaky/unsteady | 0.17 | 0.47 | 3.14 | 10.56 | 13. Indecision | 0.46 | 0.79 | 1.84 | 2.86 |
| 14. Fear of losing control | 0.24 | 0.59 | 2.65 | 6.84 | 14. Devaluation | 0.31 | 0.67 | 2.07 | 3.16 |
| 15. Difficulty breathing | 0.22 | 0.57 | 2.88 | 8.30 | 15. Loss of energy | 0.56 | 0.69 | 1.03 | 0.64 |
| 16. Fear of dying | 0.18 | 0.56 | 3.42 | 11.57 | 16. Changes in sleeping habits | 0.85 | 0.78 | 0.76 | 0.37 |
| 17. Scared | 0.39 | 0.69 | 1.76 | 2.51 | 17. Irritability | 0.43 | 0.64 | 1.51 | 2.33 |
| 18. Indigestion | 0.55 | 0.77 | 1.27 | 0.85 | 18. Changes in appetite | 0.53 | 0.71 | 1.31 | 1.52 |
| 19. Faint/lightheaded | 0.14 | 0.44 | 3.41 | 12.29 | 19. Difficulties in concentration | 0.76 | 0.80 | 0.65 | −0.64 |
| 20. Face flushed | 0.50 | 0.71 | 1.30 | 1.06 | 20. Tiredness or fatigue | 0.53 | 0.69 | 1.23 | 1.30 |
| 21. Hot/cold sweats | 0.37 | 0.66 | 1.77 | 2.62 | 21. Loss of sexual interest | 0.14 | 0.45 | 3.93 | 17.19 |

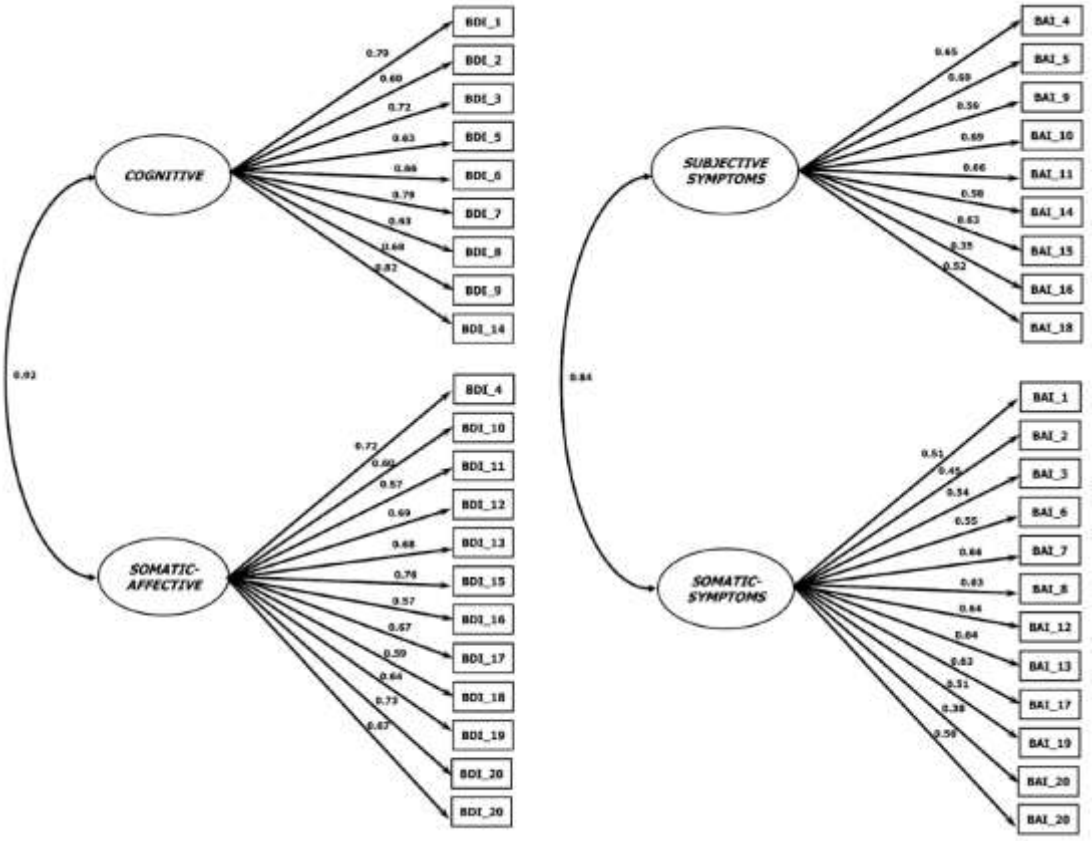
*Note.* SD, standard deviation.

Table 2 - Descriptive results of demographic variables

| | Participants | | Sex | | | Age | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Country** | ***n*** | **Total %** | **Male (%)** | **Female (%)** | ***p*** | **Total (Years [SD]** | **Male (Years [SD])** | **Female (Years [SD])** | ***p*** |
| Brazil | 315 | 16.1 | 149 (47) | 166 (53) | 0.37 | 22.8 (7.2) | 22.1 (4.3) | 23.4 (8.8) | 0.11 |
| Portugal | 426 | 21.8 | 203 (48) | 223 (52) | 0.36 | 20.4 (1.6) | 20.5 (1.6) | 20.4 (1.7) | 0.39 |
| Spain | 1210 | 62.1 | 384 (32) | 825 (68) | <0.001 | 21.5 (3.0) | 21.4 (3.5) | 21.5 (2.8) | 0.67 |
| Total | 1957 | 100.0 | 736 (38) | 1214 (62) | | 21.5 (3.8) | 21.3 (3.3) | 21.5 (4.1) | |

*Note. n* = number of participants; % = absolute frequency.

Figure 1 - Factorial solutions of the BDI-II and BAI




*Note.* The right model refers to the BDI-II and the left model refers to the BAI. “Subjective Symptoms” refers to “Subjective Affective and Panic Symptoms”.

Table 3 - Standardized regression weights (factor loadings) of BDI-II items

| BDI-II Item | Factor Loadings | | | |
|---|---|---|---|---|
| | **Brazilian** | **Spanish** | **Portuguese** | **Total** |
| **Factor 1: Somatic-Affective** | | | | |
| 4. Loss of pleasure | 0.52 | 0.62 | 0.63 | 0.72 |
| 10. Crying | 0.56 | 0.48 | 0.42 | 0.60 |
| 11. Agitation | 0.48 | 0.52 | 0.47 | 0.57 |
| 12. Loss of interest | 0.56 | 0.58 | 0.63 | 0.69 |
| 13. Indecision | 0.50 | 0.61 | 0.62 | 0.68 |
| 15. Loss of energy | 0.64 | 0.67 | 0.65 | 0.76 |
| 16. Changes in sleeping habits | 0.53 | 0.48 | 0.48 | 0.57 |
| 17. Irritability | 0.49 | 0.60 | 0.60 | 0.67 |
| 18. Changes in appetite | 0.52 | 0.52 | 0.49 | 0.59 |
| 19. Difficulties in concentration | 0.51 | 0.59 | 0.60 | 0.64 |
| 20. Tiredness or fatigue | 0.71 | 0.60 | 0.63 | 0.73 |
| 21. Loss of sexual interest | 0.33 | 0.50 | 0.46 | 0.67 |
| **Factor 2: Cognitive** | | | | |
| 1. Sadness | 0.65 | 0.64 | 0.69 | 0.79 |
| 2. Pessimism | 0.61 | 0.51 | 0.53 | 0.60 |
| 3. Feeling of failure | 0.61 | 0.58 | 0.63 | 0.72 |
| 5. Feeling of guilt | 0.54 | 0.57 | 0.53 | 0.63 |
| 6. Feeling of punishment | 0.36 | 0.50 | 0.47 | 0.66 |
| 7. Disconformity with oneself | 0.60 | 0.69 | 0.72 | 0.79 |
| 8. Self-criticism | 0.58 | 0.57 | 0.59 | 0.63 |
| 9. Suicidal thoughts | 0.55 | 0.36 | 0.34 | 0.68 |
| 14. Devaluation | 0.64 | 0.66 | 0.65 | 0.82 |

Table 4 - Confirmatory factor analysis of BDI-II for the total sample and by country

| Factor Structure | Sample | $\chi^2$ | *df* | CFI | TLI | RMSEA (90% CI) |
|---|---|---|---|---|---|---|
| BDI-II Two-Factors (Beck et al., 1996) | Brazilian | 258.002 * | 188 | 0.922 | 0.912 | 0.035 (0.024–0.045) |
| | Portuguese | 257.632 * | 188 | 0.887 | 0.873 | 0.030 (0.020–0.039) |
| | Spanish | 471.875 * | 188 | 0.850 | 0.832 | 0.036 (0.032–0.040) |
| | Total | 682.023 * | 188 | 0.855 | 0.838 | 0.037 (0.034–0.040) |

*Note. df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Standard Root Mean Square Error of Approximation; CI = confidence interval.
* $p < 0.001$

Table 5 - Standardized regression weights (factor loadings) of BAI items

| BAI Item | Factor Loadings | | | |
|---|---|---|---|---|
| | **Brazilian** | **Spanish** | **Portuguese** | **Total** |
| Factor 1: Somatic Symptoms | | | | |
| 1. Numbness or tingling | 0.32 | 0.53 | 0.58 | 0.51 |
| 2. Feeling hot | 0.46 | 0.45 | 0.42 | 0.45 |
| 3. Wobbliness in legs | 0.54 | 0.54 | 0.53 | 0.54 |
| 6. Dizzy or lightheaded | 0.65 | 0.50 | 0.62 | 0.55 |
| 7. Heart pounding/racing | 0.73 | 0.63 | 0.63 | 0.64 |
| 8. Unsteady | 0.54 | 0.64 | 0.68 | 0.63 |
| 12. Hands trembling | 0.59 | 0.65 | 0.65 | 0.64 |
| 13. Shaky/unsteady | 0.61 | 0.64 | 0.62 | 0.64 |
| 17. Scared | 0.72 | 0.62 | 0.61 | 0.63 |
| 19. Faint/lightheaded | 0.45 | 0.49 | 0.56 | 0.51 |
| 20. Face flushed | 0.45 | 0.38 | 0.45 | 0.38 |
| 21. Hot/cold sweats | 0.48 | 0.51 | 0.49 | 0.50 |
| Factor 2: Subjective Affective and Panic Symptoms | | | | |
| 4. Unable to relax | 0.55 | 0.67 | 0.67 | 0.65 |
| 5. Fear of worst happening | 0.73 | 0.69 | 0.69 | 0.69 |
| 9. Terrified or afraid | 0.62 | 0.58 | 0.58 | 0.59 |
| 10. Nervous | 0.70 | 0.69 | 0.70 | 0.69 |
| 11. Feeling of choking | 0.55 | 0.68 | 0.67 | 0.66 |
| 14. Fear of losing control | 0.75 | 0.55 | 0.54 | 0.58 |
| 15. Difficulty breathing | 0.45 | 0.58 | 0.53 | 0.55 |
| 16. Fear of dying | 0.48 | 0.32 | 0.32 | 0.35 |
| 18. Indigestion | 0.48 | 0.52 | 0.56 | 0.52 |

Table 6 - Confirmatory factor analysis of BAI for the total sample and by country

| Factor Structure | Sample | $\chi^2$ | *df* | CFI | TLI | RMSEA (90% CI) |
|---|---|---|---|---|---|---|
| BAI Two-Factors (Beck et al., 1993) | Brazilian | 287.705 * | 188 | 0.807 | 0.784 | 0.043 (0.033–0.052) |
| | Portuguese | 290.533 * | 188 | 0.862 | 0.846 | 0.036 (0.028–0.044) |
| | Spanish | 653.464 * | 188 | 0.804 | 0.781 | 0.046 (0.042–0.050) |
| | Total | 971.658 * | 188 | 0.792 | 0.767 | 0.047 (0.044–0.050) |

*Note. df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Standard Root Mean Square Error of Approximation; CI = confidence interval.
* $p < 0.001$

Table 7 - Multi-group CFA of BDI-II and BAI for Brazilian, Spanish, and Portuguese undergraduate students

| Instrument/Model | $\chi^2$ | *df* | CFI | ΔCFI | TLI | RMSEA (90% CI) |
|---|---|---|---|---|---|---|
| BDI-II Two-Factors | | | | | | |
| Configural | 911.549 * | 564 | 0.898 | - | 0.886 | 0.031 (0.027–0.035) |
| Metric | 757.522 * | 602 | 0.954 | 0.056 | 0.952 | 0.020 (0.015–0.025) |
| Scalar | 835.641 * | 640 | 0.942 | −0.012 | 0.943 | 0.022 (0.018–0.026) |
| BAI Two-Factors | | | | | | |
| Configural | 1151.617 * | 564 | 0.833 | - | 0.814 | 0.041 (0.037–0.044) |
| Metric | 945.792 * | 602 | 0.902 | 0.069 | 0.898 | 0.040 (0.035–0.044) |
| Scalar | 1221.724 * | 640 | 0.863 | −0.039 | 0.865 | 0.035 (0.031–0.038) |

*Note. df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Standard Root Mean Square Error of Approximation; CI = confidence interval.
* $p < 0.001$

## Article 2

**DO NASCIMENTO, Rodrigo Leão Ferreira; PINHEIRO, Ana Paula Azzam Gadelha; ANUNCIAÇÃO, Luis; LANDEIRA-FERNANDEZ, J.; ALVES, Maracy Domingues.** Psychometric Evidence of the Penibility Assessment Inventory (IAP) among Hospital Professionals. This manuscript was written in Portuguese and translated (with the aid of an LLM platform, see link https://chatgpt.com/share/67572c8b-c86c-800a-b70a-876bb2124150) for the current thesis. Submitted to the journal *Assessment Psychology* in December 2024.

## Abstract

Professionals working in hospital settings often report high levels of penibility. This type of suffering is associated with the work environment, typically related to burnout symptoms, leading to various negative consequences for both professionals and organizations. A gap is observed in the Brazilian context regarding instruments to assess levels of penibility. Therefore, this research aimed to develop and provide validity evidence for the Penibility Assessment Inventory (IAP). A total of 246 hospital workers participated in this study. Our results support both the unifactorial structure and the three-dimensional structure, investigated using Confirmatory Factor Analysis techniques and Item Response Theory. It is concluded that the IAP presents good preliminary psychometric evidence, with potential to assess penibility among hospital workers in the Brazilian context.



## Keywords

penibility; burnout syndrome; evidence of validity.

# 1. Introduction

The relationship between human beings and work is subject to systematic investigation by various fields of knowledge (e.g., Sociology, Law, Economics, etc.). Regarding its interface with human health, work can either elicit health or contribute to illness. This proposition is clearly supported by the current concept of mental health adopted by the World Health Organization (WHO), which defines it as "a state of well-being in which an individual realizes their own abilities, can cope with the normal stresses of life, can work productively, and is able to contribute to their community" (OMS, 2022). Given that mental health is an integral part of the broader concept of health, according to the same institution, it is the role of disciplines related to workers' health (e.g., Ergonomics, Medicine, Psychology, etc.) to investigate such phenomena.

Within this context, professional practice in the hospital work environment, especially in emergency services, has historically been associated with negative mental health outcomes for a significant portion of these workers (Alanazy e Alruwaili, 2023), resulting in high rates of burnout syndrome both during the COVID-19 pandemic (Chor *et al.*, 2021), and prior to it (Moreira, Souza e Yamaguchi, 2018). This syndrome, along with depressive and/or anxiety disorders, are among the psychopathologies associated with prolonged exposure to chronic stress (Beck e Bredemeier, 2016; Maslach e Jackson, 1981; Roche, Kostadinov e Fischer, 2017), and they are classified as occupational diseases according to Ministerial Order GM/MS Nº 1.999, 27th November, 2023 (BRASIL, 2023).

The body of evidence indicates that suffering associated with the work context exists and generates harm to workers and society at large. This form of suffering (hereinafter referred to as “penibility”) was introduced into Brazilian legislation in the 1960s, alongside the concepts of health and risk premiums. However, due to a "legal vacuum", despite being provided for in the 1988 Federal Constitution of Brazil through Article 7, which defines the “rights of urban and rural workers, in addition to other rights aimed at improving their social condition,” it lacks appropriate regulation (Petrus, 2017).

Despite the importance of the subject, there are few studies that address it in the Brazilian context. The debate on penibility has been more developed in countries

such as Portugal and France, resulting in both the regulation of labor rights and a better understanding and clarification of the concept (Petrus, 2017; Pina *et al.*, 2015). In Brazil, this topic was studied among urban bus drivers in São Paulo by researcher Leny Sato in the 1990s (Alves, 2019). Among her findings, Sato (1993) highlighted the definition of penible work as work characterized by discomfort, effort, and physical and mental suffering, over which workers exert no control. More recently, the subject has been investigated among train conductors in the railway system of Minas Gerais (Petrus, 2017) and workers in the automotive industry in São Paulo (Pina *et al.*, 2015). In general, given the complexity of the subject, the authors propose conceptual definitions and highlight the challenges in operationalizing the construct.

One of the main challenges in this debate is the development of a theoretical framework (see Figure 1) that encompasses the inherent characteristics of the construct of penibility and is capable of being measured. To address this demand, a proposal was developed by Alves (2019) based on the works of Christophe Dejours on Work Psychodynamics and Serge Moscovici, through the Theory of Social Representations, especially its societal approach formulated by Willem Doise. According to Alves, the construct of penibility can be investigated through a representational perspective linked to three dimensions: familiarity, power, and subjective limit.

PLEASE INSERT FIGURE 1 HERE

According to Alves (2019), familiarity can be understood as a gradual process of becoming accustomed to the work, enabled through the construction of specific knowledge by the workers themselves. Power refers to the worker's ability to influence and change the prescriptions that define task-related norms realized by them. The subjective limit can be understood as the worker’s awareness of their own limits, i.e., "how much," "when," and "what" can be endured at work. This last dimension is subject to dual determination, as it depends on both the work context and the individual characteristics of each worker, with no rigid or immutable boundary.

Based on the dimensions outlined above, Alves (2019) emphasizes the need to operationalize them through levels of representation using a societal approach. At the intra-individual level, the focus is on how individuals organize their experiences with the environment. The interpersonal level addresses interindividual and situational processes, seeking explanatory principles typical of social dynamics in interaction systems. The intergroup level considers the different positions that individuals occupy in social relationships and analyzes how these positions modulate processes at the first and second levels. Finally, at the societal level, the emphasis is on belief systems, representations, evaluations, and social norms, adopting the assumption that cultural and ideological productions, characteristic of a society or specific groups, give meaning to individual behaviors and create social differentiations, based on general principles.

The challenges in developing theoretical frameworks that are articulated with psychological science have been the subject of recent debate in the literature (Eronen e Bringmann, 2021). In their article, the authors discuss the relevance of psychological explanations, through constructs and concepts of this nature, to address issues at this level (a perspective known as "holistic pragmatism," see Eronen, 2021). To this end, several points are raised as instrumental. Two of them are highlighted here: (i) the robust empirical limitation of psychological phenomena that serve as the basis for theory development; and (ii) advances in methodological studies. According to the author, there is a profusion of psychological phenomena being investigated without a solid empirical foundation. These "phenomena" would serve as the basis for the construction of psychological theories, which in turn would fuel sterile research hypotheses. Regarding the need for validity studies, the authors highlight the need for more research on construct validity evidence, for example. They emphasize that, in general, there is a significant gap in validity studies, while data on reliability measures (e.g., Cronbach's alpha) are more commonly found.

With regard to the construct of penibility, as far as the authors of this study have knowledge, there are no instruments with psychometric validity evidence specifically designed to measure it. At the international level, a study conducted by Baurin (2018) was identified, which used secondary public data from the *European Working Conditions Survey 2015*, a survey that evaluates Self-Rated Health Status (SRHS) and working conditions, in addition to life expectancy (LE) data, stratified

by occupations, from the *National Longitudinal Mortality Survey (1980s–1990s)* in the United States. This study aimed to investigate the concept of penibility.

Given the impacts of the topic on workers' rights and health (e.g., social security), this research aims to present preliminary psychometric evidence for the Penibility Assessment Inventory (IAP) in the context of emergency services in a public health unit in the city of Rio de Janeiro.

## 2. Method

### 2.1 Participants

The present study employed a non-probabilistic intentional sampling process, interviewing a total of 300 out of 400 workers assigned to a public hospital emergency unit in the city of Rio de Janeiro. These professionals were categorized according to the four Major Groups (GG) of occupations defined by the Brazilian Classification of Occupations (CBO) from the Brazilian Ministry of Labor and Employment. In collaboration with the hospital's management team, it was decided to establish analytical codes based on the level of responsibility in the institution's activities, using a scale from 1 to 8 to ensure an appropriate sample distribution (see Table 1). Out of the 300 professionals interviewed, 54 participants were excluded from the final sample due to incomplete data, resulting in a final sample of 246 workers (61.50% of the target population).

PLEASE INSERT TABLE 1 HERE

### 2.2 Instruments

The Penibility Assessment Inventory (IAP) was developed to evaluate the social representations of psychological suffering among hospital emergency workers in a public unit in Rio de Janeiro. In its initial version, it was applied as a hetero-assessment instrument, using a structured interview format with twelve open-ended questions. Two items were adapted in their structure to account for the occupation title and workplace context (see Figure 2).

PLEASE INSERT FIGURE 2 HERE

## 2.3 Procedures

For this study, five preceptors working at the Miguel Couto Municipal Hospital (HMMC) and ten psychology students affiliated with the Brazilian Ministry of Health's Pet-Saúde Program were selected. All members of the research team participated in training sessions to prepare for data collection, which took place before the COVID-19 pandemic. Initially, each participant was invited to answer a structured interview composed of questions about how they perceive their own work in the hospital emergency setting. The interviews were conducted within the emergency department premises of the HMMC, covering all three work shifts.

During the interviews, which were conducted by pairs of properly trained researchers, each item was read aloud to the worker, and their responses were audio-recorded for subsequent transcription. Based on the transcribed data, each member of the researcher pairs independently assigned a score to each interview response using content analysis (Bardin, 2016). In cases of disagreement, a third evaluator intervened to harmonize the scores, ensuring a consensus was reached. The scoring indicators were as follows: 0 – “no negative aspect”, 1 – “mild negative with positive aspect”, 2 – “strong negative with positive aspect”, 3 – “mild negative without positive aspect”, 4 – “strong negative without positive aspect”, and 5 – “unbearable”.

Each interview was conducted on a voluntary basis, with participants providing informed consent as per the Informed Consent Form (TCLE). The research pairs began by presenting the study's objectives to the workers. The TCLE, which had been approved by the Research Ethics Committee of the Municipal Health and Civil Defense Secretariat of the City of Rio de Janeiro, was then provided to each participant for their review and signature.

## 2.4 Data Analysis

The data analysis process was conducted in three stages. Initially, descriptive statistics were calculated to characterize the sample using sociodemographic variables and item-level analysis (both individual items and total scores). Student's t-tests were also employed to identify differences based on gender, level of

responsibility in public service, and work shift, considering the sum of scores for the dimensions (Familiarity, Power, and Subjective Limit) as well as the overall instrument score. The "level of responsibility in public service" variable was created by collapsing occupations with more “complex” responsibilities, ranging from 1 to 4 (e.g., physicians, nursing technicians, etc.), into one group, while the remaining occupations formed the “basic” group. The work shift variable grouped workers into a “day shift” category, while the other shifts were combined into a “night shift” group.

Following these steps, confirmatory factor analysis (CFA) was conducted to verify the instrument's internal structure, testing both a three-dimensional model and a unidimensional structure. To assess model fit, the following indices were calculated: chi-square/degrees of freedom ratio ($\chi^2$/df), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). According to Brown (2015), an acceptable fit is indicated when the $\chi^2$/df ratio is less than 3, CFI and TLI values are greater than 0.95, and RMSEA values are below 0.08, with the upper bound of the 90% confidence interval not exceeding 0.10. To determine which model showed the best fit for the data, a chi-square difference test ($\chi^2$) was performed following the procedure proposed by Satorra & Bentler (2001). Finally, ordinal Cronbach's alpha ($\alpha$) and McDonald's omega ($\varpi$) coefficients were calculated to estimate the reliability of the investigated factor solutions.

The final stage of the analysis involved using Item Response Theory (IRT) to examine item discrimination and difficulty thresholds, as well as the model's fit characteristics. Given the nature of the instrument, which contains polytomous items, the Graded Response Model (GRM) proposed by Samejima (1997) was used. Model fit was assessed using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Sample-Size Adjusted BIC (SABIC), in addition to indicators derived from the M2 statistic (Paek *et al.*, 2020). The analyses presented in this study were conducted using R Studio, employing the following R packages: *mirt*, *lavaan*, *semTools*, *emmeans*, and *afex*.

## 3. Results

### 3.1 Sample characteristics

Table 2 presents information on the sociodemographic characteristics of the final sample. The majority of participants are male (60.1%) and aged between 21 and 30 years (31.7%). Approximately 54% of the sample resides with a spouse and/or children. Regarding financial responsibility, around 46.3% of participants report being partially responsible for household finances. Most participants work as on-duty shift workers (63.8%), with 59.3% of them working during the daytime. Regarding tenure, both in emergency care and at the HMMC, the majority of participants (32.9%) have been active for a period ranging from 2 to 10 years.

PLEASE INSERT TABLE 2 HERE

The descriptive statistics for the items and total scores of the instrument are presented in Table 3. The item with the highest mean was *"What is society's perception of situations involving your field of work?"* (M = 2.78, SD = 1.57), while the item with the lowest mean was *"How is your work when other groups also participate in it?"* (M = 0.72, SD = 1.2). Regarding the total scores for each dimension, Subjective Limit had the highest mean (M = 6.58, SD = 3.89), followed by Power (M = 6.07, SD = 3.39) and Familiarity (M = 5.20, SD = 3.27). The overall total score of the instrument had a mean of 17.85 (SD = 8.35).

PLEASE INSERT TABLE 3 HERE

No significant differences were found by gender for the dimensions of Familiarity ($p = 0.476$), Power ($p = 0.975$), Subjective Limit ($p = 0.753$), or for the Penibility total score ($p = 0.905$). When comparing workers by their level of responsibility in public service, significant differences were found in all dimensions and the total score. In addition, significant differences were observed in the dimensions and the

overall total, except for Familiarity, when comparing workers who work exclusively during the day shift with those who work exclusively or partially during the night shift (see Table 4).

PLEASE INSERT TABLE 4 HERE

## 3.2 Internal structure

The unidimensional model showed partially satisfactory fit indices. The chi-square test was significant ($\chi^2(54) = 85.99$, $p < 0.001$), but the chi-square/degrees of freedom ratio ($\chi^2/df = 1.59$) was within the acceptable range. The RMSEA was adequate (0.049; 90% CI [0.028 – 0.068]), while the CFI (0.922) and TLI (0.904) were below the recommended threshold of 0.95. The three-dimensional model yielded similar results. The chi-square test was significant ($\chi^2(51) = 80.98$, $p = 0.005$), and the chi-square/degrees of freedom ratio ($\chi^2/df = 1.59$) remained acceptable. The RMSEA was again adequate (0.049; 90% CI [0.027 – 0.068]), the CFI (0.927) was closer to the recommended value of 0.95, but the TLI (0.905) was still below the expected threshold. A chi-square difference test was conducted to compare the fit of the two models. The results indicated that the fit indices for the three-dimensional structure were not significantly better than those of the unidimensional solution ($\chi^2(3) = 5.42$, $p = 0.14$). Table 5 presents a summary of these results.

PLEASE INSERT TABLE 5 HERE

The reliability of the Penibility Assessment Inventory (IAP) was evaluated using the ordinal Cronbach's alpha (α) and McDonald's omega (ϖ) for both the unidimensional structure and the three-dimensional structure. For the unidimensional model, the reliability indices were considered adequate, with α (0.72) and ϖ (0.68), indicating acceptable internal consistency. In contrast, the three-dimensional model presented lower-than-expected reliability indices for each

dimension: Familiarity (α = 0.47; ϖ = 0.45); Power (α = 0.29; ϖ = 0.23); and Subjective Limit (α = 0.53; ϖ = 0.48). Table 6 provides the standardized factor loadings for each item in both the unifactorial and three-factor models.

PLEASE INSERT TABLE 6 HERE

### 3.3 Item Response Theory

The unidimensional model showed good convergence in fitting the data after 21 iterations. The log-likelihood value was -4122.399, and the model fit indices were as follows (AIC = 8388.797, BIC = 8641.181, SABIC = 8412.944). Further evaluation of the model's fit indicated acceptable indices (M2 = 7.651, RMSEA = [0.033: 0 - 0.093] (90% CI), TLI = 0.746, CFI = 0.915, df = 6, p-value = 0.264). The parameter estimates, including discrimination coefficients, difficulty thresholds, and the average difficulty parameter for the items, can be reviewed in Table 7.

PLEASE INSERT TABLE 7 HERE

## 4. Discussion

The main objective of the present study was to investigate and provide preliminary evidence of the validity of the Penibility Assessment Inventory (IAP) in its hetero-assessment version. The data used in the final sample comprised 246 participants, distributed across different occupations within an emergency hospital unit in the municipality of Rio de Janeiro. Several analyses were conducted to describe and explore the characteristics of the sample, items, and the internal structure of the instrument. In general, the IAP showed satisfactory fit indices for most of its theoretical dimensions and the unidimensional solution. Variables mentioned in the literature, such as responsibility level in care and work shift (Petrus, 2017), showed significant differences regarding the total penibility score and nearly all separate

dimensions. The instrument's items displayed good ranges of difficulty and discrimination, as assessed using an Item Response Theory (IRT) model.

The internal structure evidence of the IAP was investigated through CFA, given the presence of a robust theoretical hypothesis (Brown, 2015). Inspection of the fit indices for both the three-dimensional model and the unidimensional solution was considered satisfactory, with no significant differences found between the models. However, the reliability coefficients diverged depending on the model, being adequate for the unidimensional solution but below expectations for the separate dimensions. A possible explanation for this result may be related to the type and number of occupations assessed (n = 23). When comparing the alpha (α) value in the Familiarity dimension, grouping by the responsibility level 'complex' (α = 0.41) and 'basic' (α = 0.46), with the value obtained in the general sample (α = 0.47), a difference was observed that cannot be explained by the sample size in each group, as the first group has more participants (n = 155) than the second (n = 91).

The standardized factor loadings in both models ranged from low to moderate, with the highest being the item "What does society think about professionals in your field of work?" in the Familiarity dimension (λ = 0.583) and (λ = 0.592) in the unidimensional model. The second item with the highest standardized regression coefficient was "What is society's view of situations involving your field of work?" in the Subjective Limit dimension (λ = 0.538) and (λ = 0.543) in the unidimensional model. Overall, this dimension showed the highest values of λ, followed by Familiarity and Power, respectively. The lowest standardized factor loading was from the item "How do you perceive your autonomy in performing your work activities?" in the Power dimension (λ = 0.209), and (λ = 0.263) in the unidimensional model. The Subjective Limit dimension includes items covering factors such as 'work hours', which is associated with penibility, both in terms of work shift type and the number of hours worked (Petrus, 2017; Pina *et al.*, 2015); and addressing 'situations experienced with colleagues and/or clients'. Due to the emergency hospital environment, the item content might elicit episodic memories, which are subject to emotional biases, producing a loading effect common among study participants (Silva *et al.*, 2021; Watkins, Martin e Stern, 2000).

The analysis of the discrimination parameters revealed that the items had good discriminatory capacity, with an average of 0.82 (SD = 0.26). When evaluating the average difficulty parameter of the items, it was observed that some items had higher theta values. In general, the set of evidence points to the robustness of the evaluated models. Although the unidimensional model showed more satisfactory indicators than the three-dimensional model, it is important to note that the heterogeneity of occupations may be associated with this result. Given the scarcity of instruments of this nature, further investigation into this model is needed, considering the properties identified in this study.

Limitations of this study include the small size of each occupational category within the total sample. Although representative of the location chosen for this research, these results cannot be generalized to other categories. Another limitation is the lack of a measure of inter-rater agreement in assigning the degree of penibility to the items of the instrument. Future studies may employ invariance techniques, considering larger, stratified samples by specific occupations, to investigate the instrument's different properties at various levels (e.g., scaling); and deepen the studies considering relationships with other variables. With this study, we hope to foster the development of new psychological instruments that can investigate penibility among workers in various occupations, thus aiding in the prevention of mental health issues.

## 5. References

ALANAZY, A. R. M.; ALRUWAILI, A. The Global Prevalence and Associated Factors of Burnout among Emergency Department Healthcare Workers and the Impact of the COVID-19 Pandemic: A Systematic Review and Meta-Analysis. **Healthcare (Switzerland)**, v. 11, n. 15, 1 ago. 2023.

ALVES, M. D. **Sofrimento Psíquico do Trabalho: Construção de um Instrumento para o Diagnóstico de Penosidade**. Rio de Janeiro, Brazil: Pontifícia Universidade Católica do Rio de Janeiro, 5 abr. 2019.

BARDIN, L. **Análise de Conteúdo**. 70. ed. SãoPaulo: [s.n.].

BAURIN, A. **Job penibility: measurements and policy discussions**. [s.l: s.n.].

BECK, A. T.; BREDEMEIER, K. A Unified Model of Depression: Integrating Clinical, Cognitive, Biological, and Evolutionary Perspectives. **https://doi.org/10.1177/2167702616628523**, v. 4, n. 4, p. 596–619, 29 mar. 2016.

BRASIL. **PORTARIA GM/MS Nº 1.999, DE 27 DE NOVEMBRO DE 2023 - PORTARIA GM/MS Nº 1.999, DE 27 DE NOVEMBRO DE 2023 - DOU - Imprensa Nacional**, 2023. Disponível em: <https://www.in.gov.br/web/dou/-/portaria-gm/ms-n-1.999-de-27-de-novembro-de-2023-526629116>. Acesso em: 13 out. 2024

BROWN, T. **Confirmatory for Analysis for Applied Research**. [s.l.] Guilford Publication, 2015.

CHOR, W. P. D. *et al.* Burnout amongst emergency healthcare workers during the COVID-19 pandemic: A multi-center study. **The American Journal of Emergency Medicine**, v. 46, p. 700, 1 ago. 2021.

ERONEN, M. I. The levels problem in psychopathology. **Psychological Medicine**, v. 51, n. 6, p. 927–933, 2021.

ERONEN, M. I.; BRINGMANN, L. F. The Theory Crisis in Psychology: How to Move Forward. **Perspectives on Psychological Science**, v. 16, n. 4, p. 779–788, 2021.

MASLACH, C.; JACKSON, S. E. The measurement of experienced burnout. **Journal of Organizational Behavior**, v. 2, n. 2, p. 99–113, 1981.

MOREIRA, H. DE A.; SOUZA, K. N. DE; YAMAGUCHI, M. U. Síndrome de *Burnout* em médicos: uma revisão sistemática. **Revista Brasileira de Saúde Ocupacional**, v. 43, n. 0, p. 3, 12 mar. 2018.

PAEK *et al.* **Using R For Item Response Theory Model Applications**. New York: Routledge, 2020.

PETRUS, A. M. F. Da atividade de trabalho nos trilhos ao debate político e epistemológico sobre penosidade. 22 fev. 2017.

PINA, J. A. *et al.* Intensificação do trabalho e saúde dos trabalhadores: um estudo na Mercedes Benz do Brasil, São Bernardo do Campo, São Paulo1. **Saúde e Sociedade**, v. 24, n. 3, p. 826–840, 1 jul. 2015.

ROCHE, A.; KOSTADINOV, V.; FISCHER, J. **Stress and Addiction Introduction and Background Defining Stress**. [s.l: s.n.].

SAMEJIMA, F. Graded Response Model. **Handbook of Modern Item Response Theory**, p. 85–100, 1997.

SATO, L. A representação social do trabalho penoso. *Em*: **O conhecimento no cotidiano: as representações sociais na perspectiva da psicologia social**. [s.l.] Editora Brasiliense, 1993. v. 2p. 188–211.

SATORRA, A.; BENTLER, P. M. A scaled difference chi-square test statistic for moment structure analysis. **Psychometrika 2001 66:4**, v. 66, n. 4, p. 507–514, 2001.

SILVA, R. DE A. DA *et al.* Autobiographical Memory and Episodic Specificity Across Different Affective States in Bipolar Disorder. **Frontiers in Psychiatry**, v. 12, p. 641221, 7 maio 2021.

WATKINS, P. C.; MARTIN, C. K.; STERN, L. D. Unconscious Memory Bias in Depression: Perceptual and Conceptual Processes. **Journal of Abnormal Psychology**, v. 109, n. 2, p. 282–289, 2000.

WORLD HEALTH ORGANIZATION. **Mental health**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>. Acesso em: 13 out. 2024.

Figure 1 - Theoretical framework of the Penibility Assessment Inventory (IAP)

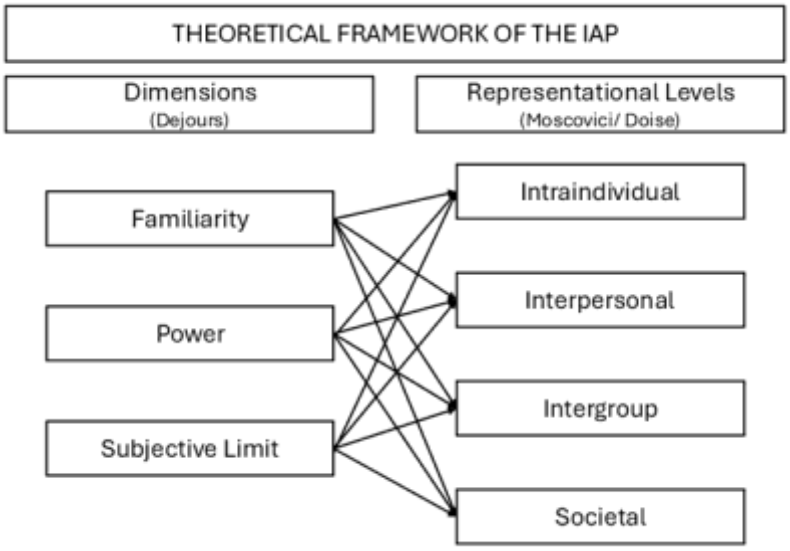


Table 1: Occupational characteristics based on the Brazilian Occupational Classification (CBO)

| Major Group (GG) | Occupation | Responsibility level | n | % |
|---|---|---|---|---|
| Sciences Professionals | Doctor | 1 | 114 | 46.34 |
| | Nurse<br>Social Worker<br>Dentist<br>Speech Therapist | 2 | | |
| | Nutritionist<br>Psychologist | 3 | | |
| Mid-Level Technicians | Nursing Technician | 4 | 51 | 20.73 |
| | Surgical Instrument Technician<br>Casting Technician<br>Radiology Technician<br>Secretary | 5 | | |
| Administrative Workers | Administrative Assistant<br>Receptionist | 6 | 25 | 10.16 |
| Service Workers | General Services Assistant<br>Cook<br>Janitor<br>Attendant<br>Security Guard<br>Watchman | 7 | 35 | 14.23 |
| Students | Academic Intern | 8 | 21 | 8.53 |
| Total | | | 246 | 100.00 |

*Note*. n = Number of observations; % = Relative frequency

Figure 2 - Interview Protocol of the Penibility Assessment Inventory (IAP)

| Dimension | Level | Item |
|---|---|---|
| Familiarity | Intraindividual | What does it mean to you to be a [occupation name] working at [workplace]? |
| | Interpersonal | What is your daily routine like with colleagues and/or clients? |
| | Intergroup | How do you perceive your work when compared to others in different groups of people? |
| | Societal | What does society think about professionals in your field of expertise? |
| Power | Intraindividual | How do you perceive your autonomy in performing your work tasks? |
| | Interpersonal | Describe a situation where you had to intervene in a situation with colleagues and/or clients. |
| | Intergroup | How is your work when other groups are also involved? |
| | Societal | How is the value placed on the [occupation name] professional working at [workplace]? |
| Subjective Limit | Intraindividual | Comment on your work schedule. |
| | Interpersonal | Describe a situation that occurred with you and your colleagues and/or clients. |
| | Intergroup | How is your communication with family and friends when you are on the work shift? |
| | Societal | What is society's view on situations involving your field of expertise? |

Table 2 - Sociodemographic characteristics of the sample (n = 246)

| Variable | n | % |
|---|---|---|
| Gender | | |
| Feminine | 98 | 39.8 |
| Masculine | 148 | 60.1 |
| Age | | |
| 21-30 | 78 | 31.7 |
| 31-40 | 50 | 20.3 |
| 41-50 | 54 | 21.9 |
| 51 or more | 64 | 26.0 |
| Family composition | | |
| Single | 34 | 13.8 |
| With spouse | 36 | 14.6 |
| With spouse and/ or with children | 135 | 54.9 |
| With relatives and/ or friends | 41 | 16.6 |
| Financial Responsible | | |
| Yes | 79 | 32.1 |
| No | 53 | 21.5 |
| Partial | 114 | 46.3 |
| Work scale | | |
| Weekdays | 82 | 33.3 |
| On duty | 157 | 63.8 |
| Dual-registered | 7 | 2.8 |
| Work shift | | |
| Day | 146 | 59.3 |
| Night | 10 | 4.1 |
| Both | 90 | 36.6 |
| Years of Occupation | | |
| Up to 1 year | 81 | 32.9 |
| 02 - 10 years | 81 | 32.9 |
| 11 - 20 years | 34 | 13.8 |
| 21 - 30 years | 36 | 14.6 |
| 31 years or more | 14 | 5.7 |
| Years of Emergency | | |
| Up to 1 year | 72 | 29.2 |
| 02 - 10 years | 79 | 32.1 |
| 11 - 20 years | 38 | 15.4 |

| | | |
|---|---|---|
| 21 - 30 years | 38 | 15.4 |
| 31 years or more | 19 | 7.7 |
| | | |
| Years of Emergency at HMMC | | |
| Up to 1 year | 43 | 17.5 |
| 02 - 10 years | 93 | 37.8 |
| 11 - 20 years | 36 | 14.6 |
| 21 - 30 years | 43 | 17.5 |
| 31 years or more | 31 | 12.6 |
| Total | 246 | 100.0 |

*Note*. n = Number of participants; % = Relative frequency

Table 3 - Descriptive statistics of items and total scores per dimension of the IAP

| Item/ Total | μ | SD | Med | Min | Max | S | K |
|---|---|---|---|---|---|---|---|
| What does it mean to you to be a [occupation name] working at [workplace]? | 0.81 | 1.27 | 0 | 0 | 5 | 1.56 | 1.62 |
| What is your daily routine like with colleagues and/or clients? | 0.76 | 1.14 | 0 | 0 | 5 | 1.64 | 1.97 |
| How do you perceive your work when compared to others in different groups of people? | 1.06 | 1.43 | 0 | 0 | 5 | 1.18 | 0.16 |
| What does society think about professionals in your field of expertise? | 2.57 | 1.53 | 3 | 0 | 5 | -0.24 | -1.22 |
| How do you perceive your autonomy in performing your work tasks? | 0.94 | 1.4 | 0 | 0 | 5 | 1.34 | 0.45 |
| Describe a situation where you had to intervene in a situation with colleagues and/or clients. | 1.98 | 1.69 | 2 | 0 | 5 | 0.19 | -1.39 |
| How is your work when other groups are also involved? | 0.72 | 1.2 | 0 | 0 | 5 | 1.56 | 1.4 |
| How is the value placed on the [occupation name] professional working at [workplace]? | 2.43 | 1.71 | 2 | 0 | 5 | 0.02 | -1.35 |
| Comment on your work schedule. | 1.2 | 1.55 | 0 | 0 | 5 | 1.06 | -0.11 |
| Describe a situation that occurred with you and your colleagues and/or clients. | 1.61 | 1.7 | 1 | 0 | 5 | 0.62 | -0.99 |
| How is your communication with family and friends when you are on the work shift? | 0.98 | 1.36 | 0 | 0 | 5 | 1.2 | 0.14 |
| What is society's view on situations involving your field of expertise? | 2.78 | 1.57 | 3 | 0 | 5 | -0.38 | -1.14 |
| Total Familiarity | 5.2 | 3.27 | 5 | 0 | 16 | 0.73 | 0.3 |
| Total Power | 6.07 | 3.39 | 6 | 0 | 17 | 0.42 | -0.06 |
| Total Subjective Limit | 6.58 | 3.89 | 6 | 0 | 17 | 0.5 | -0.28 |
| Total Penibility | 17.85 | 8.35 | 17 | 1 | 41 | 0.34 | -0.35 |

*Note*. μ = Mean; SD = Standard Deviation; Med = Median; Min = Minimum; Max = Maximum; S = Skewness; K = Kurtosis

Table 4 - Results of the Student's t tests of IAP total scores separated by Work shift and Responsibility

| Score | Work shift | | *p* | d | Responsibility | | *p* | d |
|---|---|---|---|---|---|---|---|---|
| | Night (n = 100) | Morning (n = 146) | | | Complex (n = 155) | Basic (n = 91) | | |
| Total Familiarity | 4.94 (3.13) | 5.58 (3.4) | 0.134 | - | 4.08 (2.84) | 5.86 (3.33) | < 0. 001 | 0.561 |
| Total Power | 5.67 (3.15) | 6.65 (3.6) | 0.027 | 0.289 | 5.12 (3.28) | 6.63 (3.33) | < 0. 001 | 0.456 |
| Total Subjective Limit | 6.16 (3.69) | 7.19 (4.09) | 0.040 | 0.267 | 5.05 (3.77) | 7.47 (3.68) | < 0. 001 | 0.650 |
| Total Penibility | 16.78 (7.88) | 19.42 (8.78) | 0.015 | 0.319 | 14.26 (8.16) | 19.96 (7.73) | < 0. 001 | 0.722 |

*Note*. The values refers to the means and standard deviations; d = Cohen's coefficient

Table 5 - Fit Indices of the Penibility Assessment Inventory (IAP)

| Model | $\chi^2$ | df | *p*-valor | $\chi^2$/ df | CFI | TLI | RMSEA | CI (90%) |
|---|---|---|---|---|---|---|---|---|
| Unifactorial | 85.9 9 | 54 | <0.001 | 1.59 | 0.922 | 0.904 | 0.049 | 0.028 – 0.068 |
| Three factors | 80.9 7 | 51 | 0.005 | 1.59 | 0.927 | 0.905 | 0.049 | 0.027 – 0.068 |

*Note*. $\chi^2$ = Chi-squared test; df = Degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; CI (90%) = Confidence Interval

Table 6 - Standardized factor loadings of unifactorial and three-factor models

| Item | Factor loadings | | | |
|---|---|---|---|---|
| | F | P | LS | Unifactorial |
| What does it mean to you to be a [occupation name] working at [workplace]? | 0.409 | | | 0.415 |
| What is your daily routine like with colleagues and/or clients? | 0.249 | | | 0.253 |
| How do you perceive your work when compared to others in different groups of people? | 0.528 | | | 0.534 |
| What does society think about professionals in your field of expertise? | 0.583 | | | 0.592 |
| How do you perceive your autonomy in performing your work tasks? | | 0.209 | | 0.263 |
| Describe a situation where you had to intervene in a situation with colleagues and/or clients. | | 0.31 | | 0.382 |
| How is your work when other groups are also involved? | | 0.36 | | 0.442 |
| How is the value placed on the [occupation name] professional working at [workplace]? | | 0.313 | | 0.394 |
| Comment on your work schedule. | | | 0.524 | 0.529 |
| Describe a situation that occurred with you and your colleagues and/or clients. | | | 0.486 | 0.492 |
| How is your communication with family and friends when you are on the work shift? | | | 0.352 | 0.354 |
| What is society's view on situations involving your field of expertise? | | | 0.538 | 0.543 |

*Note*. F = Familiarity; P = Power; LS = Subjective Limit

Table 7 - Discrimination and difficulty parameters of the IAP

| Item | *a* | *b*1 | *b*2 | *b*3 | *b*4 | *b*5 | *b* |
|---|---|---|---|---|---|---|---|
| What does it mean to you to be a [occupation name] working at [workplace]? | 0.779 | 0.691 | 1.684 | 2.853 | 3.936 | 5.345 | 2.902 |
| What is your daily routine like with colleagues and/or clients? | 0.443 | 0.672 | 3.551 | 5.179 | 6.519 | 12.651 | 5.714 |
| How do you perceive your work when compared to others in different groups of people? | 1.069 | 0.107 | 1.066 | 1.628 | 2.344 | 4.12 | 1.853 |
| What does society think about professionals in your field of expertise? | 1.332 | -1.967 | -0.859 | -0.217 | 0.473 | 2.528 | -0.008 |
| How do you perceive your autonomy in performing your work tasks? | 0.453 | 0.713 | 2.876 | 3.456 | 4.994 | 9.924 | 4.393 |
| Describe a situation where you had to intervene in a situation with colleagues and/or clients. | 0.686 | -1.338 | -0.235 | 0.454 | 1.793 | 4.262 | 0.987 |
| How is your work when other groups are also involved? | 0.82 | 1.053 | 1.769 | 2.776 | 4.215 | 6.239 | 3.210 |
| How is the value placed on the [occupation name] professional working at [workplace]? | 0.721 | -2.35 | -0.938 | 0.329 | 0.825 | 2.867 | 0.146 |
| Comment on your work schedule. | 0.945 | 0.058 | 0.81 | 1.613 | 2.322 | 3.533 | 1.667 |
| Describe a situation that occurred with you and your colleagues and/or clients. | 0.877 | -0.479 | 0.227 | 1.194 | 1.776 | 3.303 | 1.204 |
| How is your communication with family and friends when you are on the work shift? | 0.625 | 0.316 | 1.878 | 2.552 | 4.076 | 8.003 | 3.365 |
| What is society's view on situations involving your field of expertise? | 1.117 | -2.323 | -1.106 | -0.427 | 0.236 | 2.305 | -0.263 |

*Note*. *a* = discrimination values; *b*1 a *b*5 = difficulty thresholds; *b* = average of the difficulty thresholds

## Article 3

**DO NASCIMENTO, Rodrigo Leão Ferreira; MURPHY, Kimberly; ANUNCIAÇÃO, Luis; LANDEIRA-FERNANDEZ, J.; SQUIRES, Jane.** Exploring the Psychometric Properties of the Environmental Screening Questionnaire Research Edition (ESQ-RE) using a Network Approach. Submitted to the *Journal of Child and Family Studies* in June 2024.

## Abstract

The assessment of environmental risk factors associated with healthy child development is a critical area of research associated with improving developmental outcomes. Despite its importance, only few instruments are available to address environmental risk factors for children in their preschool years. This study aims to explore the psychometric structure of the ESQ-RE using a network analysis approach. A sample of 22,391 children aged from 1 to 83 months from the US was used to conduct a network analysis. This method was estimated using regularization procedures adequate to Ising model. The node strength, closeness, and betweenness were computed, and a community detection analysis was implemented. The network model presented 275 (97.4%) positive significantly correlations among nodes. The ESQ-RE items were clustered in six different communities; ESQ-RE item related to access of support programs (WIC, SNAP, Medicaid) was the most important of the network. The results point to robust preliminary properties of the ESQ-RE on measuring contextual risk factors associated with child development.

## Keywords

Psychometrics; Child development; Network Analysis

# 1. Introduction

The adverse childhood experiences (ACEs) seminal study posited that severe adverse childhood conditions were associated with compromising aspects in an adult's life (Lazar *et al.*, 1982). These findings were consistently reaffirmed over the two decades after the original study (Squires e Bricker, 2020). The literature in the field has indicated that ACEs are traumatic events that happen during childhood, including but not limited to experiencing violence, abuse, or neglect (Bhutta *et al.*, 2023; Lazar *et al.*, 1982; Squires e Bricker, 2020). Subsequently, an increased research focus has turned to the development of measures able to assess the impact of ACEs on child development (Squires e Bricker, 2020).

A recent review conferred an additional scope on the extent of the detrimental effects of ACEs on a subject's lifetime (Bhutta *et al.*, 2023). The authors suggest that mild to severe adverse exposure levels occurring even in the prenatal period of life are associated with compromised health outcomes throughout a lifespan. They argue that in addition to the most well-known factors established in ACEs studies (e.g. poverty, sexual abuse), it is important to consider the emerging events of armed conflicts, climate change, environmental degradation, and pandemic crises. In line with these risk factors, a study that focused on the impact of the COVID-19 pandemic on maternal mental health, parental practices, and early childhood development found evidence suggesting that the latter was affected by families' financial burden and loss of access to health services (Penna *et al.*, 2023). These contextual factors are related to stress-related risks for typical child development outcomes.

Due to its importance in subsequent life stages, preventing or mitigating the consequences of stressful events in child neurodevelopment is a global need (World Health Organization, 2023), especially in the early childhood years when a child's brain is highly sensitive to this exposure (Bhutta *et al.*, 2023; Khawli, El *et al.*, 2018; Lupien *et al.*, 2001). On an individual level, early stress exposure is associated with alterations in different brain networks and areas (e.g. amygdala centered resting-state functional connectivity, ventrolateral Prefrontal Cortex, etc.), neurocognitive functions (e.g. memory, executive functions), and capabilities (e.g. reappraisal) (Khawli, El *et al.*, 2018; Lupien *et al.*, 2001). A meta-analysis study

found evidence of the predisposition of these subjects towards the occurrence of future mental health problems (e.g. depression) (Kraaijenvanger *et al.*, 2020). The ubiquitous presence of stress in such studies has led researchers to investigate its associations with child development alterations using different measures (e.g. cortisol levels, maltreatment, socioeconomic status) (Bhutta *et al.*, 2023). According to these authors there is an overlap between people experiencing stressful environmental factors and those living under ACEs circumstances. Thus, not only biodevelopmental markers must be measured but also additional contextual risk factors.

For optimal developmental outcomes, the investigation of social determinants of health (SDOH) is vital. The World Health Organization (2013) defined SDOH as "the circumstances in which people are born, grow up, live, work and age, and the systems put in place to deal with illness. These circumstances are in turn shaped by a wider set of forces: economics, social policies, and politics." SDOH assessment and monitoring are part of a global effort driven by the WHO towards the mitigation of the effects derived from inequalities within different factors, including 'early life and education' (World Health Organization, 2023). The development of SDOH environmental screening measures for the early childhood population should: (1) focus on efforts implemented by WHO; (2) aid in identifying broadly contextual factors associated with the above-mentioned detrimental effects on children's development and wellbeing; (3) and finally, contribute to the development of knowledge about early childhood development and risk.

Despite its undoubted relevance, the measurement of the SDOH in early childhood has not received adequate focus (Sokol *et al.*, 2019). These authors found evidence of robust psychometric properties for eleven measures, but only four of them addressed the early childhood area. In the face of this scarcity of measures targeting SDOH in early childhood, the Environment Screening Questionnaire-Research Edition (Squires e Bricker, 2020, 2007) was developed to identify environmental risk factors that affect children's subsequent lives. The ESQ-RE addresses environmental risk factors in six areas: 1) Education and Employment; 2) Housing; 3) Child and Family Health; 4) Economics and Finances; 5) Family Life; and 6) Community. When used with developmental screening measures, these tools offer

a comprehensive range of information, enabling the early identification of children who may require further in-depth assessment and support.

A previous study explored the ESQ-RE initial psychometric properties including its utility (Moxley-South *et al.*, 2015). A total of 324 caregivers with young children participated completing print forms (N = 72) and online (N = 252) forms. Among the main results, preliminary evidence of concurrent validity was obtained comparing the ESQ-RE and the Parenting Stress Index-Short Form (PSI-SF;Abdin, 1995). Results indicated that the online caregivers with higher ESQ-RE scores (i.e., more risk factors) tended to have slightly higher parenting stress. Other significant correlations were found between the ESQ-RE score and the Ages & Stages Questionnaire: Social-Emotional (ASQ:SE; Squires et al., 2002, 2015) scores for the online caregiver sample at 6 months, and 48 months. In addition, 82% of agency professionals indicated they planned to use this measure in future occasions and reported that the ESQ-RE was useful for identifying family's strengths and needs, and to monitor their status.

An emergent field of research known as network psychometrics approach has been adopted to add to current knowledge of the ESQ-RE. This field of research on psychometrics posits that nodes of a network reinforce each other causing its activation and stability (Isvoranu, 2022). Network analysis has been applied in recent studies to different areas such as psychopathology (Borsboom e Cramer, 2013; Cai *et al.*, 2024) and cognition (Maas, van der *et al.*, 2017) among others (Borsboom *et al.*, 2021)

That said, this study aims to explore the psychometric structure of the ESQ-RE using a network analysis approach. This investigation will provide insights not only into structure but also into the conditions and factors that might affect a child's development.

## 2. Methods

### 2.1 Participants and Procedures

This study was part of a project that assessed early childhood development using parent-completed assessments. The data collection started in 2019, online (N =

30,902). This data collection procedure is often used in epidemiological settings due to its large reach (Tyrer e Heyman, 2016). For the larger study, respondents completed the ASQ-3, ASQ:SE-2, PSI-SF, and ESQ-RE. More details about the project are available elsewhere (https://agesandstagesresearch.com). The target of this study was on the ESQ-RE completed by a subsample of parents/caregivers (N = 22,391).

## 2.2 Sample characteristics

Parents or caregivers completed the ESQ-RE for 22,391 children. The child's age ranged from 1 to 83 months (approximately 7 years old). Males made up more than half (55.6%) of our sample. The family income range varied across categories; the majority of the respondent's education level was 4-year college or above (44.2%). Finally, a total of 14,116 (61%) of the respondents were White. Table 1 shows the demographic characteristics of the sample.

INSERT TABLE 1 HERE

## 2.3 Ethical issues

All participants were informed of the goals of this study and gave informed consent. The approval for the project entitled "Improving Early Identification: Renorming the ASQ: IRB approval was obtained before the start of the study.

## 2.4 Measure

### 2.4.1 Environmental Screening Questionnaire - Research Edition (ESQ-RE)

The ESQ-RE was developed by Squires & Bricker (2007; 2015) to assess areas known to impact the welfare of young children. It was developed to help in the identification of specific information in order to assist in targeting family needs. The ESQ-RE assesses a total of 6 areas: 1) Education and Employment; 2) Housing; 3) Child and Family Health; 4) Economics and Financials; 5) Family Life; and 6) Community. Each area has five related questions, in which caregivers use a "yes-no" option. They also mark if the question represents a concern. Regarding the

score, a `yes` response receives 10 points, while a `no` response receives 0 points. If the concerned box is checked for an item, an additional 5 points are added to that item. Thus, all areas range within 0-75 points and higher scores indicate more risk factors in a child's life that may affect caregivers' ability to meet their children's needs. Areas of greatest risk and concern to parents can be targeted for intervention. Examples from the ESQ-RE appear in supplementary table 1.

## 2.5 Statistical Plan

The ESQ-RE items fall in six areas, which were totaled separately for this analysis. Descriptive analyses of the ESQ-RE were conducted by checking area distributions using range, mean, and standard deviation. In addition to the previous analyses, a one-way repeated measure ANOVA was computed to check for significant differences within the groups followed by *post-hoc* tests.

We estimated a network model to explore the internal structure of the ESQ-RE, including its central measures and communities. Due to binary nature of all items, we followed the procedures recommended by Isvoranu and colleagues (2022) for estimating networks with binary non-normally distributed data. The first step was selecting the function `binarize` from the R package `bootnet`, which was used to recode the variables to 0 or 1. This step was necessary to allow the R package `IsingFit` to function adequately. Secondly, the network model was estimated using the eLasso procedure which combines l1-regularized logistic regression with model selection based on the Extended Bayesian Information Criterion (EBIC) on the Ising model (Borkulo, Van *et al.*, 2014). This regularization procedure sets two types of parameters: the tuning parameter lambda (λ) and the hyperparameter gamma (γ). The tuning parameter lambda (λ) reduces spurious edges in the network amplifying the number of true edges. The λ value choosing is in some instances still under debate but the default value (λ = 0.50) has been consistently suggested due to its good results in simulation studies (Epskamp e Fried, 2018), therefore we opted to maintain it. On the other hand, the hyperparameter gamma (γ) controls how much the EBIC prefers simpler models. The Ising models usually adopt the value (γ = 0.25) due to the type of data (Epskamp e Fried, 2018).

The degree of importance of each node in the network is investigated with centrality measures, and we investigated the node strength, closeness, and betweenness of the

network model. The node strength includes the sum of absolute edge weights connected to each node, which provides a direct measure of the strength of associations among the node with other nodes in the network; while the closeness is computed by taking the inverse of the sum of all distances from one node to all other nodes, which informs how much the node is distant from other nodes of the network. Finally, betweenness is computed by investigating how many shortest paths go through a node of interest, which informs how much the node provides access to other nodes of the network (Isvoranu, 2022).

Lastly, the R Package `EGAnet` was used to detect the number of communities in the network. This analysis is especially relevant to exploring new structures in complex networks (Golino e Epskamp, 2017). We opted for using the Spinglass algorithm to reveal how many clusters would be found in the network model. All analyses were performed using R Studio packages `bootnet`, `qgraph`, `IsingFit`, and `EGAnet`.

## 3. Results

The format of the distribution of the ESQ-RE areas was asymmetric, indicating that lower results (i.e. fewer risk factors) were more frequently present in the data. Table 2 describes the descriptive statistics of the ESQ-RE, including its range, mean, and standard deviation.

INSERT TABLE 2 HERE

The mean results of the six ESQ-RE areas were significantly different ($F(5, 111950) = 2538.6$, $p < 0.001$, $\eta^2 = .063$). A post-hoc test demonstrated that the only non-significant comparison was between “Education and Employment” and “Child and Family Health” ($p = 0.584$). Overall, the highest mean was obtained in the “Economics and Finances” area (8.07 ± 10.27), higher than Education and

Employment (7.45 ± 8.70), Child and Family Health (7.33 ± 9.07), Housing (1.79 ± 5.31), Family Life (4.51 ± 7.70), and Community (5.23 ± 8.40).

## 3.1 Network Analyses

Figure 1 shows the network analysis of the ESQ-RE items using the Ising model. The weighted matrix was composed of 463 association pairs. From which, 275 (59.4%) significant correlations were different from zero. Among the 275 significant associations achieved, only 12 (2.6%) were significantly negative, and 224 (97.4%) were positive. Supplementary table 2 highlights items with the strongest positive associations found in our results.

INSERT FIGURE 1 HERE

### 3.1.1 Centrality analysis

The inspection of the centrality measures appears in Figure 2. These measures provide information about the strength of the direct association of one node with others (node strength), an indirect measure related to the distance of one node with others (closeness), and about how much one node intermediate the path to other nodes (betweenness). This analysis revealed that the node ‘Do you have regular transportation?’ presented the highest degree of closeness and betweenness compared to the other nodes, and the second highest strength degree. It was surpassed only by the node ‘Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?’ which showed a strength degree slightly superior. The nodes ‘Do you have a spouse/partner who lives with you most of the time?’ and ‘Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?’ achieved the second and third highest levels of closeness compared to the rest of the network. Thus, these three nodes are possibly the most important in the network. In addition, ‘Do you have regular transportation?’, ‘Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?’, and ‘Does anyone in your home have problems with depression, anger,

or anxiety?' showed the highest betweenness levels, which means that it is the shortest path among pairs of nodes.

INSERT FIGURE 2 HERE

### 3.1.2 Community structure of the ESQ-RE network

The results of the community detection analysis performed using the Spinglass algorithm revealed six clusters of items, see Figure 3. The yellow cluster is composed of nine nodes from all areas except the areas 'Housing' and 'Education and Employment'. Most of its nodes were from the 'Community' area. The blue cluster is formed by five nodes three from the 'Child of Family Health' area, and the rest of the 'Education and Employment' area. The green cluster groups all nodes from the 'Housing' area, except the nodes 'Have you or your child/children witnessed violence in your home or neighborhood?' and 'Is your housing in below-average condition?' that pertains to the orange cluster. In addition to these nodes, this cluster is composed of two nodes from the Family Life area and the node 'Does anyone in your home have alcohol or drug problems?'. The purple cluster counts with three nodes in the 'Economics and Finances' area. Finally, the red cluster is composed of nodes from the 'Education and Employment' area in addition to the nodes 'Do you have a spouse/partner who lives with you most of the time?' and 'Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?'.

INSERT FIGURE 3 HERE

## 4. Discussion

This study aimed to investigate the preliminary psychometric properties of the ESQ-RE. To achieve this goal, a network analysis was performed with a sample composed of 22,391 children, whose ages ranged from 1 to 83 months

(approximately 7 years old). This distribution is in line with the statistics extracted from the U.S. Census Bureau population data related to the years 2019 to 2023 (U.S. Census Bureau, 2023). The ESQ-RE areas scores were asymmetric, with the Economics and Finances area achieving the greatest mean (M = 8.07, SD = 10.27), while the Housing area presented the lowest mean (M = 1.79, SD = 5.35). The network analysis found 235 significant correlations which were majority positive (N = 224), with few negative (N = 11). The inspection of the central measures showed that the item Economics and Finances 3 ('Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?') presented the greatest strength degree. This item was followed by the item Community 5 ('Do you have regular transportation?') which exhibited the highest degree of closeness and betweenness. An additional community detection analysis grouped all items into six different clusters.

When exploring the mean score values for each ESQ-RE area, significant differences were found among them ($p < 0.001$, $\eta^2 = .063$). A post-hoc test showed a non-significant comparison between "Education and Employment" and "Child and Family Health" ($p = 0.584$). The greatest result was obtained in the "Economics and Finances" area (8.07 ± 10.27), which surpassed Education and Employment (7.45 ± 8.70), Child and Family Health (7.33 ± 9.07), Community (5.23 ± 8.40), Family Life (4.51 ± 7.70), and Housing (1.79 ± 5.31). Within the "Economics and Finances" area, the item 'Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?' had the greatest mean (3.75 ± 4.84) and was the second most reported problem (37%) among all ESQ-RE items. Following this item, 'Do you have credit problems?' (2.15 ± 4.11), and 'Does your income cover your monthly expenses?' (1.24 ± 3.29) were consecutively the second and third items more endorsed in this area. Altogether, these values could be partially influenced by different patterns of consumers' credit card usage compared to before and during the pandemic crisis in the United States of America, as suggested in the report named 'Credit Cards` (U.S. Government Accountability Office, 2023). This report describes how the pandemic assistance likely helped to reduce balances in the sample studied. However, higher interest rates, lower credit limits, and longer-carried balances were more prevalent in zip codes with a majority of Black or African American or Hispanic or Latino residents than White zip codes.

The most prevalent parent concern among all the items was found in the “Child and Family Health” area concerning mental health, ‘Does anyone in your home have problems with depression, anger, or anxiety?’ (4.07 ± 4.91, 41% of the participants). Different sources suggest that mental health problems are increasing worldwide, especially after the pandemic outbreak (Mucci *et al.*, 2023; Wolf e Schmitz, 2023). In addition to this, the second most endorsed item in this area was ‘Do you have a child with a learning or behavior problem?’ (1.42 ± 3.49). The relationship between families' untreated mental health concerns and children`s learning or behavior problems is well-established in the field (Lazar *et al.*, 1982). Within the “Housing” area the item most endorsed was ‘Do you need to live with friends or family not by choice?’ (0.65 ± 2.47), while ‘Do you consider yourself homeless?’ (0.08 ± 0.91) obtained the lowest mean. In addition to slight increases in income levels due to pandemic assistance policies, these results were expected considering the sampling procedures adopted which required majority access to an online website.

The ESQ-RE areas (e.g. housing, community) can be seen as dynamically integrated, which reinforces the importance of the network approach. Thus, for the estimation of the ESQ-RE network structure we used the eLasso procedure which was suggested as useful for explorative model search (Borsboom, 2022). Among the 275 significant correlations achieved, 97.4% (N = 263) were significantly positive, while only 2.4% (N = 12) were significantly negative.

Among the positive correlations, the strongest associations found were between the nodes `Do you have frequent spouse/partner conflicts?`, `Are you in a relationship in which you have been physically hurt, felt threatened, or been controlled by someone else?`, and `Have you or your child/children witnessed violence in your home or neighborhood?`. Altogether, these results seem to suggest the presence of increased risk related to familial conflicts in our sample. According to the study review conducted during the pandemic period (Penna *et al.*, 2023), even with additional barriers to assessing intimate partner violence (IPV), the results pointed to increased rates of this and child abuse. One of the factors suggested by the authors to explain this finding was related to the need of victims to cohabitate with aggressors at that time. Other sets of risk factors with positive correlations were job concerns driven by language barriers; instability regarding accommodation and housing; lack of social support and access to child care; unavailability of material

resources (e.g. phone, transportation); lack of access to health insurance; and finally, economic difficulties related to basic needs (e.g. monthly expenses, food).

According to (Isvoranu, 2022), centrality measures are usually computed to estimate the position and role of a node in a network. The nodes ‘Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?’ and ‘Do you have regular transportation?’ showed the highest strength degree among all nodes. This measure reveals the importance of those nodes to the network, which points to the content addressed by these items. The use of support programs to provide basic needs and difficulties to obtain regular transportation are issues really central to any subject’s life. The nodes ‘Do you have regular transportation?’, ‘Do you have a spouse/partner who lives with you most of the time?’ and ‘Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?’ achieved the first, second and third highest levels of closeness compared to the rest of the network. The high closeness indicates that node reach all other nodes relatively fast. Again, transportation and access to support programs show up as items more prone to affect the other nodes. In addition to them, the marriage status appears influencing indirectly other nodes of the network. The marriage status has been considered a protective factor to different life outcomes, including life expectation (Tatangelo *et al.*, 2017). Finally, the nodes `Do you have regular transportation?`, `Do you currently use support programs such as WIC, food stamps (SNAP), or Medicaid?` and ‘Does anyone in your home have problems with depression, anger, or anxiety?’ showed the highest betweenness levels, which means that they tend to connect clusters of nodes together. In addition to the previous items mentioned, mental health concerns influence all other items in an indirect way.

The results of the community detection analysis revealed the presence of 6 clusters. The main goal of this analysis is to verify highly connected clusters that exhibit greater connectivity within than between clusters (Golino e Epskamp, 2017). As expected, all nodes grouped forming different clusters (or communities). The inspection of each cluster showed how the different nodes of each ESQ-RE area were associated to another ones.

In general, it is possible to see how different aspects of each node are directly or indirectly influencing the structure of the ESQ-RE network. Despite the added evidence on the ESQ-RE psychometric properties, this study presented some limitations. First, the collection procedure required access to the internet. Thus, people facing severe housing issues or living in poverty might not have participated. Second, our data covered a period of time pre and during COVID-19 pandemic, which could, on one hand, influence the endorsement of some items. But as noted earlier, the likelihood of events like this are predicted in the future (Bhutta *et al.*, 2023).

This study is of sum importance to the field. In face of the scarcity of instruments to cover environmental aspects that affect child development, the search of sources of evidence for new measures able to address this entangled network of contextual risk factors is important to the community. Future studies must focus on the reliability, concurrent validity, and utility aspects of the ESQ-RE.

## 5. Conclusion

The ESQ-RE has presented satisfactory preliminary psychometric properties using a network analysis. This analysis indicated the relevant information about different roles displayed by items. The cumulative evidence suggest that a 6-cluster structure is well fitted for the data and our results are partially related to the ESQ-RE theoretical areas: education and employment, housing, child and family health, economics and financials, family life, and community.

# 6. References

ABDIN, R. R. **Parenting Stress Index, Third Edition: Professional Manual**. Third Edition ed. Odessa, FL.: Psychological Assessment Resources, Inc., 1995.

BHUTTA, Z. A. *et al.* Adverse childhood experiences and lifelong health. **Nature Medicine 2023 29:7**, v. 29, n. 7, p. 1639–1648, 18 jul. 2023.

BORKULO, C. D. VAN *et al.* A new method for constructing networks from binary data. **Scientific Reports**, v. 4, 1 ago. 2014.

BORSBOOM, D. *et al.* Network analysis of multivariate data in psychological science. **Nature Reviews Methods Primers 2021 1:1**, v. 1, n. 1, p. 1–18, 19 ago. 2021.

___. Possible Futures for Network Psychometrics. **Psychometrika**, v. 87, n. 1, p. 253–265, 1 mar. 2022.

BORSBOOM, D.; CRAMER, A. O. J. Network analysis: an integrative approach to the structure of psychopathology. **Annual review of clinical psychology**, v. 9, p. 91–121, mar. 2013.

CAI, H. *et al.* A network model of depressive and anxiety symptoms: a statistical evaluation. **Molecular Psychiatry 2024**, p. 1–15, 18 jan. 2024.

EPSKAMP, S.; FRIED, E. I. A tutorial on regularized partial correlation networks. **Psychological methods**, v. 23, n. 4, p. 617–634, 1 dez. 2018.

GOLINO, H. F.; EPSKAMP, S. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. **PloS one**, v. 12, n. 6, 1 jun. 2017.

ISVORANU, A.-M. **Network Psychometrics with R.** . [s.l.] Taylor & Francis., 2022.

KHAWLI, E. EL *et al.* Early-Life stress modulates neural networks associated with habitual use of reappraisal. **Behavioural Brain Research**, v. 337, p. 210–217, 30 jan. 2018.

KRAAIJENVANGER, E. J. *et al.* Impact of early life adversities on human brain functioning: A coordinate-based meta-analysis. **Neuroscience and biobehavioral reviews**, v. 113, p. 62–76, 1 jun. 2020.

LAZAR, I. *et al.* Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies Effects of Early Education: A Report from the Consortium

for Longitudinal Studies (1982), pp. i+iii+v-vii+ix-xiv+1-151. **Source: Monographs of the Society for Research in Child Development**, v. 47, n. 2, 1982.

LUPIEN, S. J. *et al.* Can poverty get under your skin? Basal cortisol levels and cognitive function in children from low and high socioeconomic status. **Development and Psychopathology**, v. 13, n. 3, p. 653–676, jun. 2001.

MAAS, H. L. J. VAN DER *et al.* Network Models for Cognitive Development and Intelligence. **Journal of Intelligence 2017, Vol. 5, Page 16**, v. 5, n. 2, p. 16, 20 abr. 2017.

MOXLEY-SOUTH, K. *et al.* The Environmental Screening questionnaires: A brief family risk and resilience Screening. **Journal of Human Services**, v. 35, n. 1, p. 62–72, 2015.

MUCCI, F. *et al.* Navigating the "Mental Health Crisis" in Adolescents in the Aftermath of Covid-19 Pandemic: Experience and Insights from Frontline Psychiatric Service. **Clinical Neuropsychiatry**, v. 20, n. 4, p. 309, 1 ago. 2023.

PENNA, A. L. *et al.* Impact of the COVID-19 pandemic on maternal mental health, early childhood development, and parental practices: a global scoping review. **BMC Public Health 2023 23:1**, v. 23, n. 1, p. 1–26, 24 fev. 2023.

SOKOL, R. *et al.* Screening Children for Social Determinants of Health: A Systematic Review. **Pediatrics**, v. 144, n. 4, 2019.

SQUIRES, J. *et al.* **Ages & Stages Questionnaires: Social-Emotional: A parent-completed, child-monitoring system for social-emotional behaviors**. 1. ed. Baltimore: Paul H. Brookes, 2002.

SQUIRES, J.; BRICKER, D. Screening for Social Determinants of Health: Environmental Screening Questionnaire. **International Forum of Special Education and Child Development**, v. 2, n. 11, p. 1–11, 12 nov. 2020.

SQUIRES, J.; BRICKER, D.; TWOMBLY, E. **Ages and Stages Questionnaires: Social- Emotional: A parent-completed child-monitoring system - 2nd edition.** 2 Edition ed. Baltimore: Brookes Publishing, 2015.

SQUIRES, JANE.; BRICKER, D. D. An activity-based approach to developing young children's social emotional competence. p. 276, 2007.

TATANGELO, G. *et al.* Gender, marital status and longevity. **Maturitas**, v. 100, p. 64–69, 1 jun. 2017.

TYRER, S.; HEYMAN, B. Sampling in epidemiological research: Issues, hazards and pitfalls. **BJPsych Bulletin**, v. 40, n. 2, p. 57–60, 2016.

U.S. CENSUS BUREAU. **Vintage 2022 Population Estimates**.

U.S. GOVERNMENT ACCOUNTABILITY OFFICE. **CREDIT CARDS Pandemic Assistance Likely Helped Reduce Balances, and Credit Terms Varied among Demographic Groups Report to Congressional Committees United States Government Accountability Office**. [s.l: s.n.]. . Acesso em: 27 maio. 2024.

WOLF, K.; SCHMITZ, J. Scoping review: longitudinal effects of the COVID-19 pandemic on child and adolescent mental health. **European Child & Adolescent Psychiatry 2023 33:5**, v. 33, n. 5, p. 1257–1312, 21 abr. 2023.

WORLD HEALTH ORGANIZATION. **Social determinants of health: Key concepts**. Disponível em: <https://www.who.int/news-room/questions-and-answers/item/social-determinants-of-health-key-concepts>. Acesso em: 27 maio. 2024.

___. **Integrating the social determinants of health into health workforce education and training**. [s.l: s.n.]. . Acesso em: 27 maio. 2024.

Table 1 - Demographic characteristics of the sample

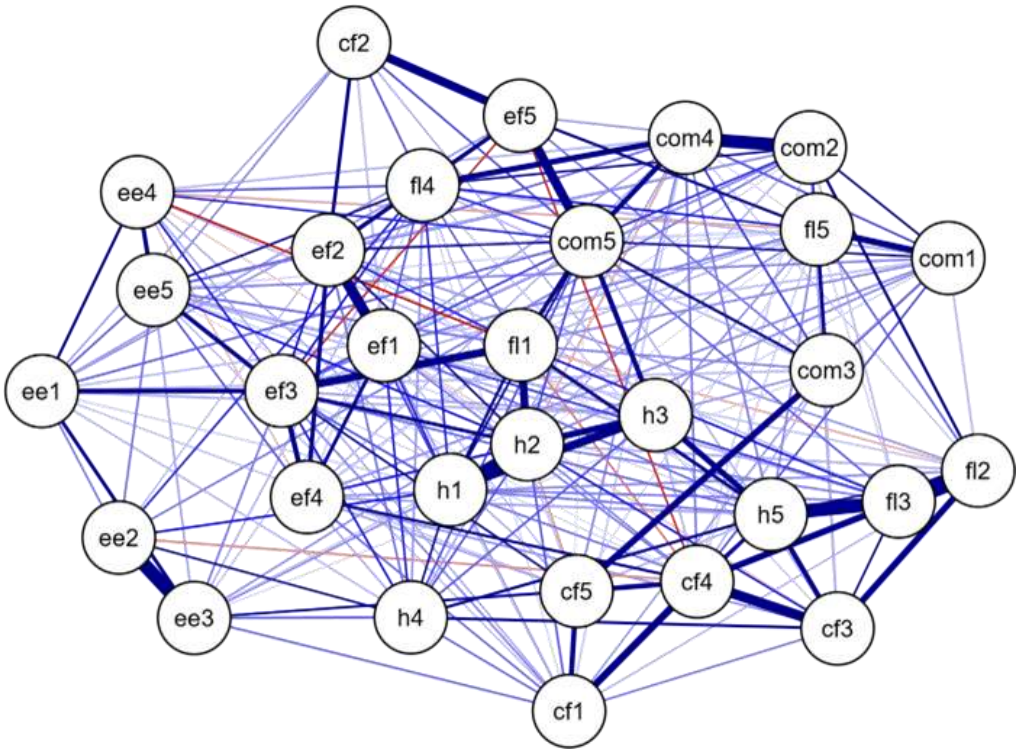
| | *N* | *%* |
|---|---|---|
| Gender | | |
| 1 Male | 12,446 | 55.6 |
| 2 Female | 9,936 | 44.4 |
| 3 Other | 9 | 0.04 |
| | | |
| Income (USD) | | |
| 1 $0-16,000 | 2,315 | 11.0 |
| 2 $16,501 – $24,500 | 1,967 | 9.3 |
| 3 $24,501 - $44,500 | 3,647 | 17.3 |
| 4 $44,501- $60,000 | 6,653 | 31.6 |
| 5 Over $60,000 | 6,487 | 30.7 |
| 6 Don't Know | 0 | 0 |
| | | |
| Mother's education | | |
| 1 less than high school | 627 | 2.8 |
| 2 high school | 8,144 | 36.4 |
| 3 AA degree | 3,404 | 15.2 |
| 4 4 4-year college or above | 9,893 | 44.2 |
| 5 Don't know | 323 | 1.4 |
| | | |
| Race/ethnicity | | |
| 1 Asian | 579 | 2.8 |
| 2 White | 14116 | 61.0 |
| 3 American Indian or Alaska Native | 182 | 0.8 |
| 4 Hawaiian or Pac Islander | 41 | 0.2 |
| 5 Black / African American | 2,529 | 11.5 |
| 6 Hispanic / Latino/a | 2,181 | 11.8 |
| 8 Some Other | 189 | 0.8 |
| 9 Don't know | 92 | 0.6 |
| 10 2 or more races | 2,482 | 10.5 |
| | | |
| Total cases | 22,391 | 100.0 |

Table 2 - Statistical results of all areas of the ESQ-RE

| Area | Results |
|---|---|
| Education and Employment | |
| - Mean (SD) | 7.45 (8.70) |
| - Range | 0.00 - 75.00 |
| Housing | |
| - Mean (SD) | 1.79 (5.31) |
| - Range | 0.00 - 75.00 |
| Child and Family Health | |
| - Mean (SD) | 7.33 (9.07) |
| - Range | 0.00 - 65.00 |
| Economics and Finances | |
| - Mean (SD) | 8.07 (10.27) |
| - Range | 0.00 - 65.00 |
| Family Life | |
| - Mean (SD) | 4.51 (7.70) |
| - Range | 0.00 - 70.00 |
| Community | |
| - Mean (SD) | 5.23 (8.40) |
| - Range | 0.00 - 75.00 |

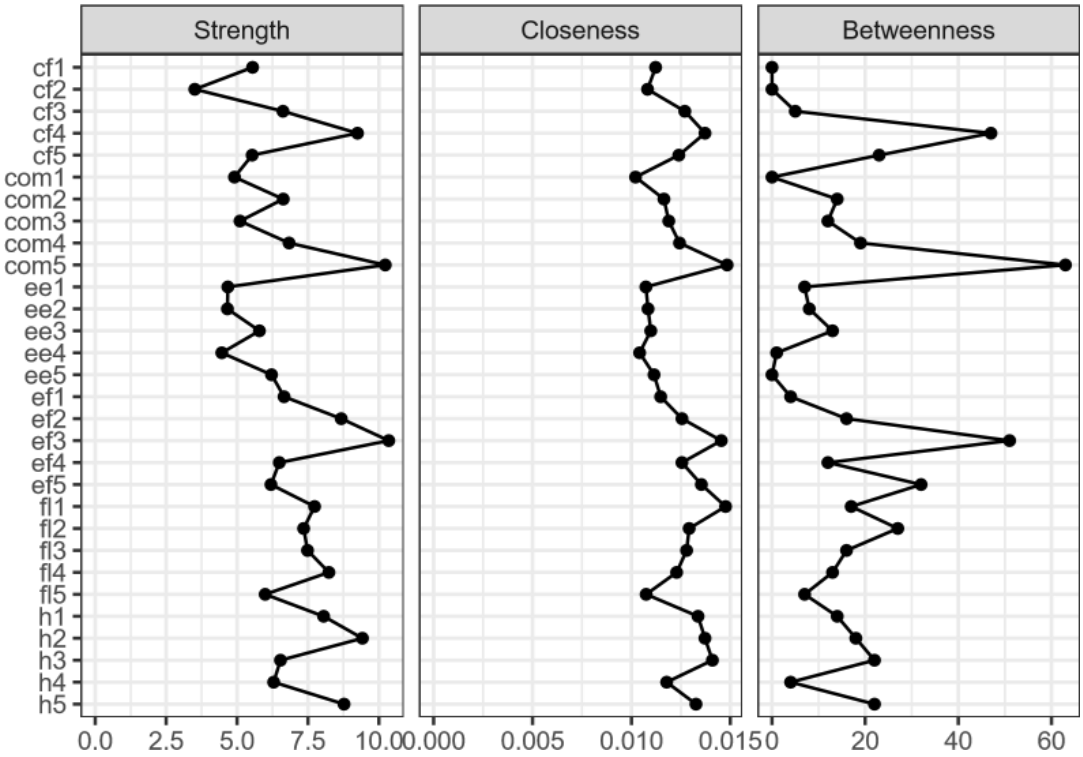
*Note*. SD = Standard Deviation

Figure 1 - Estimated network model for the ESQ-RE items.



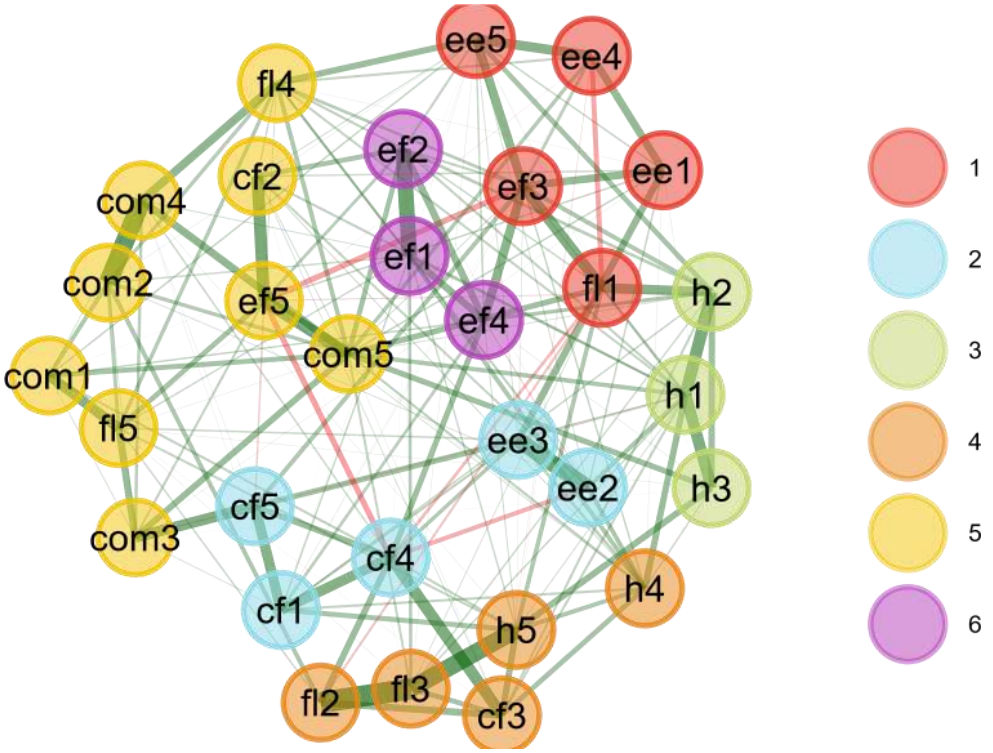

*Note*. The dark blue lines represent positive correlations while the red ones represent negative correlations. The edge thickness represents the strength of the association between the nodes.

Figure 2 - Centrality measures of the ESQ-RE items




*Note.* The figure shows the standardized values for the strength, closeness, and betweenness of the ESQ-RE nodes

Figure 3 - Community structure of the ESQ-RE containing 6 clusters




*Note*. Nodes starting with ‘h’ refers to housing, ‘ee’ to education and employment, ‘ef’ to economics and finances, ‘cf’ to child and family health, ‘fl’ to family life, and ‘com’ to community

**Supplementary Table 1.**

Examples of items of the ESQ-RE

| |
|---|
| A. Education and Employment |
| Are you a high school or GED graduate? |
| Are you employed at the level you would like to be? |
| B. Housing |
| Do you consider yourself homeless? (Examples include living in a shelter or car, or camping because you don't have a home or apartment.) |
| Have you or your child/children witnessed violence in your home or neighborhood? |
| C. Child and Family Health |
| Do you or does anyone in your home have major health problems? (Major means the problem is chronic and affects everyday life.) |
| Do you have a child with a learning or behavior problem? |
| D. Economics and Finances |
| Do you worry about having enough food for your family? |
| Do you have access to a phone when you need to make calls? |
| E. Family Life |
| Do you have a spouse/partner who lives with you most of the time? |
| Are you able to read, play, or sing with your child/children several times per week? |
| F. Community |
| Does your family join in community activities? (Examples include going to the library, playing sports, going to church, or attending other community events.) |
| Do you have regular transportation? (Examples include access to a car, bus, train, or subway.) |

**Supplementary Table 2.**

Nodes strongly associated in the ESQ-RE Network

| |
|---|
| `Do you have frequent spouse/partner conflicts?` |
| `Are you in a relationship in which you have been physically hurt, felt threatened, or been controlled by someone else?` |
| `Have you or your child/children witnessed violence in your home or neighborhood?` |
| `Do language problems get in the way of your finding or keeping a job?` |
| `Do you have problems with reading or writing? |
| `Do you consider yourself homeless?` |
| `Do you need to live with friends or family not by choice?` |
| `Have you moved three or more times in the past year?` |
| ; `Do you have people to talk to about your problems?` |
| `Does your child/do your children get along well with other children?` |
| `Do you have child care that meets your family's needs?` |
| `Do you and your family members have health insurance or access to regular medical and dental care?`, |
| `Do you have access to a phone when you need to make calls?` |
| `Do you have regular transportation?` |
| `Do you worry about having enough food for your family?` |
| `Does your income cover your monthly expenses?`. |

## Article 4

**DO NASCIMENTO, Rodrigo Leão Ferreira; MENDES, Elias Rego; ANUNCIAÇÃO, Luis; LANDEIRA-FERNANDEZ, J.** Large Amounts of Missing Data & IRT: A Brief Overview of Current Challenges and Opportunities. *In preparation*.

## Abstract

Item Response Theory (IRT) models are usually adopted in large-scale assessments in order to provide information about item functioning. However, these tools usually present large amounts of missing data which creates bias to the model estimates obtained with IRT analysis. On missing data field research, Little and Rubin's theory is widely adopted. In general lines, it posits that three mechanisms cause distinct missing data types: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Depending on various reasons, MNAR produces nonignorable missing values. Several approaches have been implemented to handle nonignorable missing values in educational settings with the purpose of estimating reliable item parameters. Overall, these approaches can be separated into deterministic and model-based ones. Despite this topic has been broadly debated in educational settings, it has been less investigated in the human developmental field, particularly across childhood development phases. Thus, this study aimed to provide a brief overview of current challenges and opportunities of implementing IRT within large amounts of missing data in this area of study.



## Keywords

IRT; missing data; childhood development.

# 1. Introduction

Early childhood is undoubtedly one of the most important periods in one's lifetime. Due to its importance in a person's life further outcomes such as health, economics and professional areas, it has been targeting worldwide policies which aimed to foster nurture environments for growth of children in different countries (Currie e Rossin-Slater, 2015; Irwin, Siddiqi e Hertzman, 2007). To this end, several strategies are usually adopted including monitoring and assessing early childhood development (World Health Organization, 2023). The latter is inherently dependent on measurement, particularly Psychometrics.

Several tools have been developed in order to track childhood developmental delays (Irwin, Siddiqi e Hertzman, 2007). Depending on which framework the tool was developed (e.g. Factor Analysis, Item Response Theory), it lies in different assumptions (e.g. local independence) (Brown, 2015). This constitutes the empirical part of any psychometrical tool development process accounting for one of its major steps, validity evidence gathering. However, turning back on previous steps, important decisions must be made regarding tool design, administration, scoring etc. Longstanding evidence converges with the occurrence of detrimental effects of tool's length on test taker's motivation, data quality, and response rates (Galesic e Bosnjak, 2009; Herzog e Bachman, 1981). Altogether, the evidence raises a relevant (but, underrated) debate about the trade-off between tool design and respondent's experience.

Recently, a few solutions have been suggested to mitigate the attentional burden caused by tool length on respondent's experience. Commonly, these solutions are aimed at constructing short tools or even shortening the existing ones. A recent study conducted by Schultze e Lorenz ([s.d.]) provided a broad overview of automated item selection procedures associated with this end. For the purpose of guiding researchers on the implementation of these procedures, the authors highlight three 'critical decisions' to be made prior to the use of an algorithm. Wise e Sidarus ([s.d.]) described a method for shortening existing measures referred to as Factor Score Item Reduction with Lasso Estimator (FACSIMILE), which showed reasonable accuracy with 'as few as three items'.

In addition to this burden, the respondent's experience is influenced by a set of different issues associated with the tool design in each context such as educational (Frey, Hartig e Rupp, 2009). In this study, the authors investigated the construction of booklet designs that are typically used in large-scale assessments of student achievement, such as the Programme for International Student Assessment (PISA), and the National Assessment of Educational Progress (NAEP) in the United States, among others. The authors pointed out how the process of 'finding a solution for allocating a large number of questions to a smaller number of booklets' encompasses context-specific constraints. For instance, the 'testing time allotted for the administration of the items' was described as one of the key constraints for booklet designs that imposes challenges on further statistical analyses.

In summary, for whatever reason listed above, the development of new tools aligned with these needs must address challenges imposed by tool design on further psychometrical analyses. In particular, the handling of missing data on IRT models and item/person parameter estimates has been extensively discussed in the field (Rose, Davier, von e Xu, 2010; Waterbury, 2019). According to Frey, Hartig e Rupp (2009), the missing data resulting from insufficient time for certain students compared to others on completing large-scale assessments can potentially distort parameter estimates which create bias in person ability estimates. Besides this, considering the high number of omitted responses found in these exams given that 'the underlying reasons for omissions are not completely understood' (Rose, Davier, von e Xu, 2010), different approaches have been adopted to treat data. Before looking in depth into this discussion, it is important to summarize the different types of missing data.

## 2. Missing data

Fortunately, one doesn't have to miss out on good definitions of missing data. Recently, it was defined as 'when an observation has no value assigned to it' (Mirzaei *et al.*, 2022). In spite of its simplicity, the 'problem of missing data', as introduced by Rubin (1976), requires different answers to one big question: which steps must be followed to handle it? To shed light on this field, Rubin (1976) investigated the conditions that cause missing data. In this seminal study, the author provided a set of notations that formalized three conditions which includes the

definition of the: (i) data ‘missing at random’; (ii) data ‘observed at random’; (iii) and, finally, the verification of differences between $\theta$ and $\phi$, given a random variable where the parameter of the data is $\theta$, and the parameter of the conditional distribution of the missing-data process is $\phi$. In summary, this study settled a solid ground for the investigation of missingness. Almost 30 years later, Rubin’s theory was updated, expanding the study of missingness. A more detailed view on this problem was discussed by Little e Rubin (2002), which have defined three distinct missing data mechanisms: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) and missing not at random (MNAR).

Each one of the different mechanisms was formalized as shown in Table 1. For a better understanding of the table’s notations, it is necessary to present some definitions. Given a certain data matrix composed of rows and columns, rows represent observations and columns refer to variables, Little e Rubin (2002) defined the complete data as $Y = (y_i)$, while $M = (M_{ij})$ as the missing-data indicator matrix. The conditional distribution of $M$ given $Y$ characterizes a missing-data mechanism, let’s say $f(M \mid Y, \phi)$, where $\phi$ refers to unknown parameters. Put simply, MCAR occurs if missingness does not depend on the values of the data, independent of being or not observed. It would happen if someone unwittingly skipped an item. Once the missingness depends exclusively on the components of $Y_{obs}$ of $Y$ that are observed it is named MAR. For instance, when someone young ($Y_{obs}$, age) skipped an item about income. Finally, MNAR refers to the missingness dependent on $y_i$ values that are missing. A good example is related to income as well. In cases when people’s annual incomes are too low or too high, omitting this information is related directly to the variable ($y_i$, income) one is trying to measure.

INSERT TABLE 1 HERE

Despite its widespread use in the field, some aspects of the above-mentioned theory have been criticized, as debated by Enders (2023). In this study, the author refers to the criticisms made by Manski. In addition to that, Enders (2023), highlights a change on Little and Rubin’s terminology made recently in their textbook, which rephrased the mechanism ‘not missing at random’ (NMAR) to ‘missing not at random’ (MNAR). Aside from these issues, Little and Rubin’s framework remains

significantly relevant in this field. In particular, MNAR mechanisms have received attention among psychometricians due to their applications in educational settings regarding nonignorable missing data.

According to Sikov (2018), given a data matrix that contains missing data caused by a MNAR mechanism, it is called ‘nonignorable’ when one assumes that the missingness is dependent of the unobserved variables ($Y_{mis}$) after conditioning on certain covariates ($Y_{obs}$). As highlighted by Rose, Davier e von e Xu (2010), to neglect the status of ‘nonignorable missing data’ on IRT analyses has the potential to distort person and item parameter estimates.

## 3. Missing data and IRT

Estimation approaches of IRT modelling within nonignorable missing values were investigated in different ways. Some examples are the: two-stage method, multiple imputation methods, and treating missingness as an extra dimension in a latent model (for a review see Enciso, 2016). A study conducted by Rose, Davier e von e Xu (2010) aimed at verifying the effects of different treatments for nonignorable missing data in educational large-scale survey assessments, in which several criticisms were made to a ‘common practice in operational data analyses’ (p. 3), named ‘deterministic replacement’. This practice consists of recoding missing values derived from not-reached items as incorrect responses. According to the authors, it neglects the stochastic relation between the latent proficiency variable and the manifested item values.

As argued by Bock e Gibbons (2021), IRT models assume a stochastic mechanism that is modeled by an unobservable random variable and a threshold. Rose, Davier, von e Xu (2010), compared deterministic-based models with model-based ones. The latter had some innovations that can be summarized as follow: (1) with the adoption of a latent regression approach, the risk of increasing the dimensionality of the IRT scaling model was diminished; (2) the presence of a covariate based on the amount of missingness can be defined even with few omissions to support IRT-based scaling of response propensity indicator variables; (3) the latent regression approach would not increase analyses-related costs. In fact, it was found evidence of decreased bias for three distinct model-based models compared with four

different deterministic-based models. As reviewed by Gorgun e Bulut (2021), alternative model-based approaches were developed based innovations such as on joint modeling of ability and response times. In a recent doctorate thesis, Enciso (2016), present a summarized version of the missingness treatment in IRT modelling context. The author distinguishes two main approaches: (1) generation of imputed values with IRT models; (2) adjustments of IRT models to incorporate the missingness as latent variable, as a manifest variable using an indicator, or as grouping factor using missingness levels. One limitation of this classification, for our purpose, is regarded as the missing data mechanism that is not restricted to MNAR.

## 4. Current and future challenges in childhood developmental field

All above-mentioned initiatives were proposed in educational setting. However, almost no studies have been conducted in developmental settings to evaluate those methods, especially considering attrition (Enders, 2013). Given the recent trend of shortening existing tools to decrease attentional burden on respondents it is important to anticipate some challenges regarding IRT applications in this context. A possible challenge is related to planned missing data designs. According to Xu e Logan (2024), the latter 'allow researchers to deliberately incorporate missingness into their data collection plans by randomly assigning participants to receive or not receive specific measures or items' (p. 1233). In developmental field, missing data below the floor would be designated as MNAR or MAR? Is it possible to incorporate imputation methods to decrease bias on IRT parameters estimations? These methodological questions arise naturally once one synthetizes evidence of these different fields.

Altogether, we aimed to provide a brief overview of current challenges and opportunities of implementing IRT within large amounts of missing data in this area of study. In summary, considering the sophisticated advances achieved in educational settings in addition to the scarcity of studies in developmental field, current and future challenges were addressed in order to raise awareness on this debate.

## 5. References

BOCK, R. DARRELL.; GIBBONS, R. D. . **Item response theory**. [s.l.] John Wiley & Sons, Inc., 2021.

BROWN, T. **Confirmatory for Analysis for Applied Research**. [s.l.] Guilford Publication, 2015.

CURRIE, J.; ROSSIN-SLATER, M. Early-Life Origins of Life-Cycle Well-Being: Research and Policy Implications. **Journal of Policy Analysis and Management**, v. 34, n. 1, p. 208–242, 1 jan. 2015.

ENCISO, S. M. S. **The Effects of Missing Data Treatment on Person Ability Estimates Using IRT Models**. [s.l.] University of Nebraska, 2016.

ENDERS, C. K. Dealing With Missing Data in Developmental Research. **Child Development Perspectives**, v. 7, n. 1, p. 27–31, 1 mar. 2013.

___. Missing Data: An Update on the State of the Art. **Psychological Methods**, 2023.

FREY, A.; HARTIG, J.; RUPP, A. A. An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. **Educational Measurement: Issues and Practice**, v. 28, n. 3, p. 39–53, set. 2009.

GALESIC, M.; BOSNJAK, M. Effects of questionnaire length on participation and indicators of response quality in a web survey. **Public Opinion Quarterly**, v. 73, n. 2, p. 349–360, maio 2009.

GORGUN, G.; BULUT, O. A Polytomous Scoring Approach to Handle Not-Reached Items in Low-Stakes Assessments. **Educational and Psychological Measurement**, v. 81, n. 5, p. 847–871, 1 out. 2021.

HERZOG, A. R.; BACHMAN, J. G. **Effects of Questionnaire Length on Response Quality Downloaded fromPublic Opinion Quarterly**. [s.l: s.n.]. Disponível em: <http://poq.oxfordjournals.org/>.

IRWIN, L. G.; SIDDIQI, A.; HERTZMAN, C. **Early Childhood Development : A Powerful Equalizer Early Child Development : A Powerful Equalizer Final Report**. [s.l: s.n.]. Disponível em: <www.earlylearning.ubc.ca/WHO>.

LITTLE, R. J.; RUBIN, D. B. **Statistical Analysis with Missing Data**. 2nd. ed. [s.l: s.n.].

MIRZAEI, A. *et al.* Missing data in surveys: Key concepts, approaches, and applications. **Research in Social and Administrative Pharmacy**, v. 18, n. 2, p. 2308–2316, 1 fev. 2022.

ROSE, N.; DAVIER, M. VON; XU, X. MODELING NONIGNORABLE MISSING DATA WITH ITEM RESPONSE THEORY (IRT). **ETS Research Report Series**, v. 2010, n. 1, p. i–53, 1 jun. 2010.

RUBIN, D. B. **Inference and missing dataBiometrika**. [s.l: s.n.]. Disponível em: <http://biomet.oxfordjournals.org/>.

SCHULTZE, M.; LORENZ, T. **A TUTORIAL ON AUTOMATED ITEM SELECTION 1 I choo-choo-choose you: A tutorial on automated item selection in scale construction**. [s.l: s.n.]. Disponível em: <https://orcid.org/0000-0003-1925-2403>.

SIKOV, A. A Brief Review of Approaches to Non-ignorable Non-response. **International Statistical Review**, v. 86, n. 3, p. 415–441, 1 dez. 2018.

WATERBURY, G. **Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation**. [s.l: s.n.]. Disponível em: <https://www.researchgate.net/publication/333520847>.

WISE, T.; SIDARUS, N. **Reducing the Burden of Psychological Questionnaire Measures Through Selective Item Re-Weighting**. [s.l: s.n.].

WORLD HEALTH ORGANIZATION. **Integrating the social determinants of health into health workforce education and training**. [s.l: s.n.]. . Acesso em: 27 maio. 2024.

XU, M.; LOGAN, J. A. R. Two-Method Measurement Planned Missing Data With Purposefully Selected Samples. **Educational and Psychological Measurement**, 1 dez. 2024.

Table 1 - Notations of missing data mechanisms proposed by Little and Rubin (2002)

| Type of mechanism | Notation |
|---|---|
| MCAR | $f(M \mid Y, \phi) = f(M \mid \phi)$ for all $Y, \phi$ |
| MAR | $f(M \mid Y, \phi) = f(M \mid Y\text{obs}, \phi)$ for all $Y\text{mis}, \phi$ |
| MNAR | $f(Y, M \mid \theta, \phi) = f(Y \mid \theta)\, f(M \mid Y, \phi) = \prod f(y\text{i} \mid \theta) \prod f(M\text{i} \mid y\text{i}, \phi)$ |

# IV. GENERAL DISCUSSION

The present thesis is aimed at exploring distinct aspects of psychometrical applications as its nuanced interplays. For this end, three empirical studies were conducted with the purpose of gathering valid evidence for different constructs across human development. In addition to that, a brief overview of current challenges in the field was provided in order to foster debate in the field. To achieve these goals, it was needed to introduce, following a historical approach, different psychometrical traditions such as Network Analysis and Item Response Theory, as presenting its limitations and current challenges. Underlying this review, it emphasized the relevance of the substantive counterpart of statistical modelling associated to these traditions. From the author's point of view, substantive, statistical, and theoretical aspects of Psychometrics would be equally considered in psychometrical applications.

The first study investigated the internal structure of the BDI-II and BAI and their level of invariance. From a representative sample of undergraduate students from Spain (n = 1216), Portugal (n = 426), and Brazil (n = 315), Confirmatory Factor Analysis and Multigroup Confirmatory Factor Analysis were performed. Albeit no evidence of invariance properties was achieved to BAI, a two-factor structure of BDI-II reached invariant properties at three levels. This study adds evidence to the field of multicultural mental health. Given the current challenges faced by many countries across the world regarding mental health concerns at populational levels (WHO Team, 2022), the evidence achieved in this study has the potential to aid on the establishment of public policies to young adults in Brazil, Spain, and Portugal.

The second article presented here aimed at collecting valid evidence of the internal structure of the IAP in the Brazilian hospital context. This tool was developed to assess an underrated, although important construct, named 'Penibility'. Given the scarcity of studies about the theme, this article is innovative because it evaluated IAP's internal structure with robust psychometrical techniques (CFA and IRT). In addition to that, it provided an operational definition of Penibility following previous work conducted by Alves (2019). Findings obtained in this study partially converged in the direction of the proposed framework which presumes three factors. In this seminal study, unidimensional structure was more consistent across

the different fit indices implemented. However, due to sample size or even plurality of occupations assessed, especially bearing their levels of complexity, the authors reinforced the relevance of exploring it in future studies.

The third study implemented a Network Analysis to verify the preliminary psychometrical properties of the ESQ-RE. This instrument was developed to measure contextual risk factors involved in early childhood development. One of the most strengths regarding this tool is related its association with the SDOH (World Health Organization, 2023). The results obtained in this study showed that its network structure was stable and the number of six communities theoretically stated was verified with the use of community detection analysis. However, a few nodes were out of the expected communities. The application of Network Model to this tool was substantively motivated by the assumption that SDOH are not latent factors. Thus, the current article showed that it is necessary to keep in mind that constructs may be psychometrically assessed putting all these levels in the decision table.

Finally, the fourth manuscript provided a brief overview of the association between IRT modeling and large amounts of missing data in the developmental field. This study aimed at fostering the urgent need for rethinking test taker's experience, especially considering attentional burden. Given the trend on shortening tools, through the development of new or even modifying the existing ones, tool design must anticipate further problems regarding IRT modelling in the presence of missing data. To this end, this study explored this theme not in its traditional setting (educational) but in developmental field. Again, scarce literature was found, being Enders (2013), a relevant voice in this area of research.

The current thesis presents some limitations. First, none of the empirical studies conducted were longitudinal. Cross-sectional findings difficult the generalizability of findings obtained. However, some constructs, for instance, depression and anxiety, are widely investigated, being the results achieved here in line with previous findings (Bardhoshi, Duncan e Erford, 2016). Second, no study investigated network models and factor models at the same time. This would be interesting to add evidence to psychometrical literature as methodological comparative studies (Christensen, Golino e Silvia, 2020) aid advancing this debate.

Despite lacking empirical evidence, this issue was theoretically debated across the different sections of this thesis. The present study leveraged methods and models across diverse contexts. To this end, multicultural huge samples were submitted to statistical procedures to describe and infer patterns of association and differences. The thesis investigated constructs associated with direct social impact with the potential to influence public policies worldwide. Bearing in mind the contextual and environmental risks associated with wars and pandemics in human health, it's noteworthy that future studies must advance the debate to mental health and human development at populational levels.

## V. REFERENCES


AERA; APA; NCME. **Standards in educational and psychological testing**. [s.l.] AERA, 2014.

ALVES, M. D. **Sofrimento Psíquico do Trabalho: Construção de um Instrumento para o Diagnóstico de Penosidade**. Rio de Janeiro, Brazil: Pontifícia Universidade Católica do Rio de Janeiro, 5 abr. 2019.

BARDHOSHI, G.; DUNCAN, K.; ERFORD, B. T. Psychometric Meta-Analysis of the English Version of the Beck Anxiety Inventory. **Journal of Counseling and Development**, v. 94, n. 3, p. 356–373, 1 jul. 2016.

BOCK, R. DARRELL.; GIBBONS, R. D. . **Item response theory**. [s.l.] John Wiley & Sons, Inc., 2021.

BOECK, P. DE *et al.* Questioning Psychological Constructs: Current Issues and Proposed Changes. **Psychological Inquiry**, v. 34, n. 4, p. 239–257, 2023.

BORSBOOM, D. *et al.* Chapter 1. Network perspectives. *Em*: ISVORANU, A. M. *et al.* (Eds.). . **Network psychometrics with R: guide for behavioral and social scientists.** 1. ed. New York: Routledge, Taylor & Frances Group, 2022. v. 1.

BORSBOOM, D.; CRAMER, A. O. J. Network analysis: an integrative approach to the structure of psychopathology. **Annual review of clinical psychology**, v. 9, p. 91–121, mar. 2013.

BROWN, T. **Confirmatory for Analysis for Applied Research**. [s.l.] Guilford Publication, 2015.

CHRISTENSEN, A. P.; GOLINO, H.; SILVIA, P. J. A Psychometric Network Perspective on the Validity and Validation of Personality Trait Questionnaires. **https://doi.org/10.1002/per.2265**, v. 34, n. 6, p. 1095–1108, 1 dez. 2020.

ENCISO, S. M. S. **The Effects of Missing Data Treatment on Person Ability Estimates Using IRT Models**. [s.l.] University of Nebraska, 2016.

ENDERS, C. K. Dealing With Missing Data in Developmental Research. **Child Development Perspectives**, v. 7, n. 1, p. 27–31, 1 mar. 2013.

ERONEN, M. I.; BRINGMANN, L. F. The Theory Crisis in Psychology: How to Move Forward. **Perspectives on Psychological Science**, v. 16, n. 4, p. 779–788, 2021.

GOODWIN, C. JAMES. **A history of modern psychology**. 6th. ed. [s.l.] Wiley, 2022.

KORBMACHER, M. *et al.* The replication crisis has led to positive structural, procedural, and community changes. **Communications Psychology 2023 1:1**, v. 1, n. 1, p. 1–13, 25 jul. 2023.

MACHADO, W. DE L.; VISSOCI, J.; EPSKAMP, S. Análise de rede Aplicada À Psicometria e À Avaliação Psicológica. *Em*: HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. (Eds.). . **Psicometria**. Porto Alegre: Artmed, 2015. p. 1–192.

NOSEK, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. **Annual Review of Psychology**, v. 73, p. 719–748, 2022.

PASQUALI, L. **Psicometria: Teoria dos testes na psicologia e na educação**. 5ª ed. [s.l.] Vozes, 2013.

RINDSKOPF, D. Reliability: Measurement. *Em*: **International Encyclopedia of the Social & Behavioral Sciences: Second Edition**. [s.l.] Elsevier Inc., 2015. p. 248–252.

ROSE, N.; DAVIER, M. VON; XU, X. MODELING NONIGNORABLE MISSING DATA WITH ITEM RESPONSE THEORY (IRT). **ETS Research Report Series**, v. 2010, n. 1, p. i–53, 1 jun. 2010.

RUST, J.; GOLOMBOK, S. **Modern Psychometrics**. 3rd. ed. [s.l.] Routledge, 2008.

SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. DE. Avanços na psicometria: da Teoria Clássica dos Testes à Teoria de Resposta ao Item. **Psicologia: Reflexão e Crítica**, v. 26, n. 2, p. 241–250, 2013.

WATERBURY, G. **Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation**. [s.l: s.n.]. Disponível em: <https://www.researchgate.net/publication/333520847>.

WHO TEAM. World Mental Health Report: Transforming Mental Health For All - Executive Summary. p. 1–28, 2022.

WORLD HEALTH ORGANIZATION. **Integrating the social determinants of health into health workforce education and training**. [s.l: s.n.]. . Acesso em: 27 maio. 2024.