UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA
DEPARTAMENTO DE PSIQUIATRIA

ANTONIO REIS DE SÁ JUNIOR

**Aplicação da teoria de resposta ao item na análise de diferença de gênero de
sintomas depressivos na população universitária brasileira**

Tese apresentada à Faculdade de Medicina
da Universidade de São Paulo para
obtenção do título de Doutor em Ciências

Programa de Psiquiatria

Orientador: Prof. Dr. Wang Yuan Pang

São Paulo

2019

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Sá Júnior, Antônio Reis de
Aplicação da teoria de resposta ao item na
análise de diferença de gênero de sintomas
depressivos na população universitária brasileira /
Antônio Reis de Sá Júnior. -- São Paulo, 2019.
Tese(doutorado)--Faculdade de Medicina da
Universidade de São Paulo.
Programa de Psiquiatria.
Orientador: Wang Yuan Pang.

Descritores: 1.Depressão 2.Gênero 3.Viés
4.Psicometria 5.Inventário de depressão de Beck
6.Teoria de resposta ao item

USP/FM/DBD-029/19

Responsável: Erinalva da Conceição Batista, CRB-8 6755

DEDICATÓRIA

Dedico este trabalho ao meu amado filho,
que me estimula todos os dias a querer ser
uma pessoa melhor, me dando as forças
necessárias para transpor os obstáculos em
meu caminho e me ensinando o valor das
lutas e das conquistas que a vida nos oferece.

AGRADECIMENTOS

Nossa! Foram muitos os que me ajudaram a concluir este trabalho.

Meus sinceros agradecimentos...

... a Deus por permitir que eu realizasse esta viagem até aqui;

... à minha família, por ter me acompanhado por todo o caminho,

por compreender a escassez da minha presença,

mas sempre incentivando e apoiando todos os meus passos;

... ao meu orientador, um professor compreensivo nas horas certas

e um amigo nas horas incertas, sempre atento e dedicado ao meu

crescimento acadêmico, com sua minuciosa e esmerada orientação;

... à Professora Doutora Clarice Gorenstein, por aceitar colaborar

com valiosas sugestões, comentários e críticas que tanto contribuíram

para o aperfeiçoamento desta pesquisa;

... ao Programa de Pós-graduação em Psiquiatria: em especial às

secretárias do Programa, Eliza Fukushima e Isabel Ataíde,

por todo o suporte durante o curso;

... aos colegas da FMUSP, pelo companheirismo nos momentos de

descontração;

... a todos os amigos, por todo o incentivo que me deram, de todas as formas,

de perto ou de longe.

**O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento
de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001**

NORMALIZAÇÃO ADOTADA

Esta tese está de acordo com as seguintes normas, em vigor no momento desta publicação:

Referências: adaptado de *International Committee of Medical Journals Editors* (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. *Guia de apresentação de dissertações, teses e monografias*. Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A. L. Freddi, Maria F. Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3a ed. São Paulo: Divisão de Biblioteca e Documentação; 2011.

Abreviaturas dos títulos dos periódicos de acordo com *List of Journals Indexed in Index Medicus*.

SUMÁRIO

RESUMO

ABSTRACT

LISTAS

1. INTRODUÇÃO.....	1
1.1 Depressão: características clínicas e epidemiológicas.....	1
1.2 Prevalência comparativa entre Homem Mulher	2
1.3 Inventário de Depressão de Beck-II	5
1.4 Avaliação Psicométrica	7
1.4.1 Conceituação e Modelos.....	7
1.4.2 Teoria de Resposta ao Item como ferramenta de análise	8
1.4.3 Funcionamento Diferencial do Item	12
1.4.4 Métodos de análise do Funcionamento Diferencial do Item	16
1.5 Importância do presente estudo	18
2. HIPÓTESES DO ESTUDO.....	20
3. JUSTIFICATIVA	22
4. OBJETIVOS	24
4.1 Objetivos Gerais	24
4.2 Objetivos Específicos	24
5. MATERIAL E MÉTODO	25
5.1 Delineamento do Estudo I Levantamento Nacional Sobre o uso de álcool, tabaco e outras drogas entre universitários das 27 capitais brasileiras.....	25
5.2 Caracterização da População do Estudo	27
5.3 Características do instrumento Inventário de Depressão de Beck, versão II.....	28
5.4 Análise Estatística.....	29
5.5 Aspectos Éticos.....	33
6. RESULTADOS	34
6.1 Confiabilidade e análise descritiva	34
6.2 Parâmetros dos Itens pela Teoria da Resposta ao Item.....	38
7. DISCUSSÃO	51
8. CONCLUSÃO.....	58
9. ANEXOS	60
Anexo A.....	61
Anexo B.....	64

Anexo C.....	65
Anexo D.....	66
Anexo E	74
10. REFERÊNCIAS	103

11. APÊNDICES

Apêndice A

Apêndice B

Apêndice C

Apêndice D

RESUMO

Sá Júnior AR. *Aplicação da teoria de resposta ao item na análise de diferença de gênero de sintomas depressivos na população universitária brasileira* [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2019.

INTRODUÇÃO: Embora seja amplamente difundida a ideia de que não existem diferenças significativas entre homens e mulheres em termos dos sintomas que experimentam durante os episódios depressivos, pesquisas recentes sugerem que as diferenças sutis no perfil de sintomas podem existir e apontar para fundamentais diferenças de gênero na fisiopatologia dos estados depressivos. As mulheres são duas vezes mais propensas que os homens a sofrer de transtorno depressivo maior durante a sua vida e uma série de estudos têm indicado que as mulheres deprimidas tendem a apresentar mais sintomas depressivos "atípicos" e mais ansiedade e sintomas de somatização quando comparadas com os homens. Alguns dos problemas com as abordagens tradicionais para medir estas diferenças se devem às deficiências da teoria estatística subjacente dessas abordagens, como é o caso da teoria clássica dos testes (TCT). Nos últimos anos, o uso da teoria de resposta ao item (TRI) começou a substituir TCT na construção e avaliação de testes e alguns autores têm seguido com o uso de métodos baseados na TRI para estudar significativas variações clínicas individuais em vez dos métodos tradicionais baseados em escores totais. Um modelo para a TRI muitas vezes utilizado com a pontuação de itens politômicos é o modelo de resposta gradual (MRG). O MRG é adequado para a análise de escalas psicométricas com as categorias de resposta ordenadas e crescentes. O funcionamento diferencial dos itens da escala do Inventário de Depressão de Beck segunda versão

(BDI-II) foi investigado, por gênero, comparando estes subgrupos de alunos.

MÉTODOS: Os 21 itens do BDI-II foram aplicados transversalmente em uma amostra representativa de 12.677 estudantes universitários brasileiros. A confiabilidade foi avaliada com base no coeficiente alfa de Cronbach. Os parâmetros gravidade (b_i) e discriminação (a) de cada item do BDI-II foram calculados através do MRG. A influência do gênero foi testada para o funcionamento diferencial do item (DIF) dentro da abordagem baseada na TRI. **RESULTADOS:** O BDI-II apresentou boa confiabilidade ($\alpha = 0,91$). As mulheres apresentaram significativamente maior probabilidade de depressão (ponto de corte > 13) do que os homens. Em geral, os participantes endossaram mais facilmente itens cognitivo-somáticos do que itens afetivos da escala. “Sentimento de culpa”, “ideias suicidas” e “perda de libido” foram os itens que mais provavelmente indicaram maior gravidade da depressão ($b_3 \geq 3,60$). No entanto, todos os itens da BDI-II mostraram de moderada a alta discriminação ($a \geq 1,32$) para o estado depressivo. Enquanto um item foi sinalizado como apresentando DIF, “chorar”, com maior probabilidade das mulheres endossarem esse sintoma, o peso global desse item no escore total foi insignificante. **CONCLUSÕES:** Embora o gênero dos entrevistados possa apresentar influência no padrão de resposta dos sintomas depressivos, as medidas dos sintomas autorreferidos não inflacionaram os escores de gravidade. Esses achados fornecem suporte adicional à validade do uso do BDI-II para avaliar a depressão em contextos acadêmicos e destacam o valor de se considerar sintomas comuns de depressão relacionados ao sexo.

Descritores: depressão; gênero; viés; psicometria; inventário de depressão de Beck; teoria de resposta ao item.

ABSTRACT

Sá Júnior AR. Using item response theory to analyze gender difference in depressive symptoms in Brazilian college students [thesis]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

INTRODUCTION: Although it is widely held that there are no significant differences between men and women in terms of the symptoms that they experience during depressive episodes, recent research suggests that subtle differences in symptom profile may exist and may point to fundamental gender differences in the pathophysiology of depressive states. Women are up to twice as likely as men to suffer from major depressive disorder during their lifetime and a number of studies have found that depressed females tend to exhibit more ‘atypical’ depressive symptoms and more anxiety and somatization symptoms compared to men. Some of the problems with the traditional approaches to measure change are due to the shortcomings of the underlying statistical theory of these approaches, which is classical test theory (CTT). In recent years, the use of item response theory (IRT) has started to replace CTT in test construction and test evaluation and some authors have called for the use of IRT based methods to study individual clinical significant change instead of traditional methods based on total scores. An often-used IRT model for polytomous item scores is the graded response model (GRM). The GRM is suitable for analyzing psychometric scales with ordered and growing response categories. Potential differential item functioning of the scale items of the Beck Depression Inventory–II (BDI-II) is investigated, by gender, to compare across sub-groups of

students. **METHODS:** The 21-item BDI-II was cross-sectionally administered to a representative sample of 12,677 Brazilian college students. Reliability was evaluated based on Cronbach's alpha coefficient. Severity (b_i) and discrimination (a) parameters of each BDI-II items were calculated through the GRM. The influence of gender and age were tested for differential item functioning (DIF) within the IRT-based approach. **RESULTS:** The BDI-II presented good reliability ($\alpha = 0.91$). Women significantly presented a higher likelihood of depression (cut-off > 13) than men. In general, participants endorsed more easily cognitive-somatic items than affective items of the scale. "Guilty feelings", "suicidal thoughts", and "loss of interest in sex" were the items that most likely indicated depression severity ($b_3 \geq 3.60$). However, all BDI-II items showed moderate-to-high discrimination ($a \geq 1.32$) for depressive state. While one item was flagged for DIF, "crying", for gender, with women more likely endorsed this symptom than men, the global weight of this item on the total score was negligible. **CONCLUSIONS:** Although respondents' gender might present influence on response pattern of depressive symptoms, the measures of self-reported symptoms have not inflated severity scores. These findings provide further support to the validity of using BDI-II for assessing depression in academic contexts and highlight the value of considering gender-related common symptoms of depression.

Descriptors: depression; gender identity; bias; psychometrics; Beck depression inventory; item response theory.

LISTAS

Lista de siglas e abreviaturas

OMS	- Organização Mundial da Saúde
NCS	- National Comorbidity Survey
ECA	- Epidemiologic Catchment Area
BDI	- Inventário de Depressão de Beck
BDI-IA	- Inventário de Depressão de Beck primeira versão atualizada
BDI-II	- Inventário de Depressão de Beck segunda versão
DSM-IV	- Manual Diagnóstico e Estatístico de Transtornos Mentais,
quarta edição	
TCT	- Teoria clássica dos testes
TRI	- Teoria de resposta ao item
Θ	- Traço latente (letra grega teta)
CCI	- Curva Característica do Item
CII	- Curva de informação do item
a	- Discriminação
b	- Dificuldade
GRM	- Modelo de resposta gradual
DIF	- Funcionamento Diferencial do Item, ou como denominado em
inglês, Differential Item Functioning	
GR	- Grupo referência
GF	- Grupo focal
SENAD	- Secretaria Nacional de Políticas sobre Drogas
GREAFMUSP	- Grupo Interdisciplinar de Estudos de Álcool e Drogas da
Faculdade de Medicina da Universidade de São Paulo	

IES

- Instituições de Ensino Superior

TCLE

- Termo de consentimento livre e esclarecido

Lista de figuras

Figura 1. Curva Característica do Item de um item hipotético sob o modelo logístico de dois parâmetros (curva sólida)

Figura 2. Curva de Informação do Item de quatro itens sob o modelo logístico de dois parâmetros

Figura 3. Representação gráfica de um item com Funcionamento Diferencial do Item uniforme

Figura 4. Representação gráfica de um item com Funcionamento Diferencial do Item não-uniforme

Figura 5: Curva Característica do Item para os 21 itens

Figura 6: Curva de informação dos 21 itens

Figura 7: O primeiro gráfico representa a pontuação que representa o construto avaliado (pontuação da teoria da resposta ao item) da amostra de estudantes universitários sobre o Funcionamento Diferencial dos Itens, para o item 10 (Choro). O segundo gráfico mostra a Curva Característica do Item para o item 10. As curvas mostram a probabilidade de endossar um item específico

Figura 8: Impacto geral do Funcionamento Diferencial do Item (DIF) para gênero nas Curvas Características do Teste (TCC). O primeiro gráfico exibe o TCC de todos os itens e o segundo gráfico exibe o TCC de itens específicos do DIF

Lista de tabelas

Tabela 1: Características psicométricas do Inventário de Depressão de Beck-II para uma amostra de estudantes universitários ($N = 12,677$)

Tabela 2: Proporção da resposta do BDI-II (%) para as alternativas de categorias de escolha nos itens selecionados

Tabela 3: Parâmetros estimados do modelo de resposta gradual (GRM) dos itens do BDI-II ($N = 12,677$).

Tabela 4: Critérios Utilizados para Detectar o Funcionamento Diferencial dos Itens.

Tabela suplementar: Proporção da resposta do BDI-II (%) para as alternativas de categorias de escolha nos 21 itens

1. INTRODUÇÃO

1.1 Depressão: características clínicas e epidemiológicas

A depressão é uma das principais causas de morbidade em todo o mundo, sendo que a Organização Mundial da Saúde (OMS) estima que será a segunda causa de incapacidade mundial em 2020, logo depois das doenças cardíacas (1). Considerada uma doença comum, grave e complexa que afeta um número estimado de 121 milhões de pessoas em todo o mundo, sendo uma condição relativamente comum (2), de curso crônico (3) e recorrente (4–6). Está frequentemente associada com incapacitação funcional (7) e comprometimento da saúde física (8–10). Os pacientes deprimidos apresentam limitação da sua atividade e bem-estar (11) além de uma maior utilização de serviços de saúde (12).

Estudos de prevalência em diferentes países ocidentais mostram que a depressão é um transtorno frequente. A *National Comorbidity Survey Replication* (NCS-R), um estudo de base populacional norte-americana, publicou resultados que mostram uma prevalência de 6,6% no último ano e 16,2% a longo da vida (5). No Brasil, a prevalência de depressão na população geral ao longo da vida é de aproximadamente 17% (13). Em um estudo internacional, realizado em 18 países, a prevalência encontrada foi de 11,1%, e entre os países de renda média, a maior prevalência encontrada foi no Brasil, com percentual de 18,4% (14).

Existem poucos estudos populacionais longitudinais em larga escala sobre o curso da depressão, mas os estudos clínicos disponíveis mostram que uma proporção substancial de pessoas que procuram tratamento para depressão maior apresenta um curso crônico e recorrente (15,16). Estudos longitudinais de longo prazo também mostram que algumas pessoas com diagnóstico de depressão deixam de relatar sua história de depressão em estudos transversais (17,18). O maior desafio na estimativa da prevalência ao longo da vida é a possibilidade de viés de memória, onde dois mecanismos poderiam levar a esse viés: esquecimento e ressignificação. Algumas pessoas podem simplesmente esquecer sintomas específicos que eles experimentaram no passado. Outras podem reconfigurar cognitivamente suas experiências de modo que não mais as recontem de uma maneira que apoie a atribuição de um diagnóstico (18). Além disso, alguns sintomas podem gerar constrangimento em seu relato, tais como as ideias suicidas, sendo que entre 60 e 70% dos pacientes, na fase aguda da depressão, experenciam este sintoma (17).

A causalidade em depressão parece ser multifatorial, incluindo interações entre genes e eventos estressantes, ou entre traumas nos primeiros anos de vida e estresse na vida adulta. O tempo de início e remissão da depressão varia muito, indicando diferentes trajetórias de sintomas durante longos períodos de tempo, com resultados diferentes devido a possíveis causas diferentes. Eventos precoces da vida e do desenvolvimento parecem ser importantes fatores de risco para a depressão, incluindo a exposição ao estresse agudo e crônico nos primeiros anos de vida (19).

1.2 Prevalência comparativa entre Homem Mulher

Um dos achados mais consistentes entre os estudos epidemiológicos em depressão é a preponderância da prevalência nas mulheres em relação aos homens de sintomas depressivos e dos transtornos depressivos unipolares (20). No mundo ocidental, a depressão é aproximadamente duas vezes mais prevalente em mulheres do que em homens (21). As razões para as diferenças de gênero em transtornos depressivos unipolares ainda não são adequadamente compreendidas (22). Existem várias revisões da literatura (21,23,24) com o objetivo de analisar as principais explicações para a maior prevalência de depressão em mulheres. Diferentes modelos explicativos têm sido sugeridos, como genéticos, hormonais, psicológicos e socioculturais, mas falta um consenso a respeito (21). Um número crescente de pesquisadores reconhece que a depressão deve ser entendida a partir de uma perspectiva multifatorial, ou seja, o resultado de uma interação variável entre os fatores biológicos, psicológicos e sociais (23,25). O Conselho Sueco de Avaliação de Tecnologias em Saúde, em uma análise das pesquisas sobre a depressão, chegou à conclusão de que os fatores socioculturais (tais como experiências adversas na infância), em combinação com processos de socialização e fatores psicológicos (tais como a vulnerabilidade a eventos adversos de vida e habilidades de enfrentamento) foram os fatores explicativos mais importantes para as diferenças de gênero na depressão (26).

Pesquisas que envolvem os determinantes sociais da preponderância de depressão no sexo feminino se concentram em eventos de vida estressantes, como problemas com a ruptura conjugal e dificuldades associadas ao emprego. Alguns modelos mostram que as disparidades na depressão em relação ao gênero são devidas a uma maior exposição das mulheres a questões sociais ligadas a esses eventos (27). Uma constatação recorrente é que as mulheres são mais influenciadas pelo apoio

emocional e por relacionamentos pessoais em que a intimidade emocional, confiança e solidariedade estão presentes (28). As mulheres parecem ter "o custo de cuidar" (27), o que significa que elas estão mais expostas e mais sensíveis a eventos de redes sociais, levando ao aumento dos níveis de depressão quando o estresse interpessoal ou a perda interpessoal são experimentados. Para os homens, há uma grande quantidade de evidências que sugerem que eles são mais propensos à depressão devido a eventos estressantes relacionados com o trabalho (27). No entanto, estudos empíricos frequentemente relatam resultados conflitantes, sugerindo que a associação do índice de risco de depressão com eventos de vida negativos e tensões crônicas requer uma análise mais aprofundada (20).

As manifestações clínicas da depressão têm se mostrado diferentes entre os gêneros, com as mulheres mais frequentemente relatando alterações de apetite e do sono, fadiga, ansiedade e relatos compatíveis com hipocondria somática (29–31). As mulheres são mais propensas que os homens a tentar o suicídio, mas os homens são mais efetivos na tentativa (32). Assim, alguns autores têm considerado a necessidade de avaliarmos diferentes itens, ou mesmo diferentes escalas e algoritmos de classificação entre homens e mulheres, acreditando que estas medidas podem influenciar as taxas de depressão. Contudo, variáveis determinantes na explicação das diferenças de gênero em transtornos depressivos ainda precisam ser estabelecidas e sua combinação em modelos etiológicos integrados continua exigindo novas evidências para a confirmação das teorias disponíveis.

Uma possibilidade para explicar a diferença de gênero encontrada nos levantamentos epidemiológicos pode estar relacionada aos critérios diagnósticos ou instrumentos utilizados para avaliar a depressão. Supõe-se que os sintomas avaliados para o diagnóstico da depressão sejam igualmente eficazes na identificação do

transtorno em ambos os gêneros, embora seja bem reconhecido que mulheres e homens experimentam e expressam seus sintomas de uma maneira diferente (29–32). Há uma crescente sugestão de que os critérios utilizados, atualmente, nos instrumentos diagnósticos para avaliar a psicopatologia da depressão, possam ser mais representativos de um padrão do transtorno segundo o gênero feminino, sendo menos sensíveis à depressão em homens (33,34). Para uma maior precisão diagnóstica, é importante que os critérios e instrumentos utilizados para avaliar a psicopatologia da depressão reflitam adequadamente os sintomas e as experiências do transtorno em mulheres e homens.

1.3 Inventário de Depressão de Beck-II

Não há um padrão-ouro para o diagnóstico da depressão (35). Uma das alternativas para torná-lo menos subjetivo e tentar quantificá-lo é a aplicação de questionários, ferramentas capazes de fornecer informação para apoio diagnóstico, evolução ou intensidade do transtorno, mensurando-o. Para a detecção de casos em estudos populacionais, a realização de inquéritos contendo duas fases consome recursos importantes (36). Devido a isso, diversos instrumentos são aplicados em pesquisas comunitárias, com razoável acurácia (37). Essas ferramentas resultam em pontuações que podem sugerir a presença de transtorno depressivo maior e a gravidade do caso. Até o momento, não se sabe qual instrumento é superior ao outro.

O Inventário de Depressão de Beck (BDI) foi desenvolvido em 1961 e passou por várias revisões (38). Sua revisão começou em 1971 com o objetivo de melhorar a redação dos itens e o instrumento final revisado foi publicado em 1979 (BDI-IA) (39). Um manual técnico para o BDI-IA foi publicado em 1987 e revisto em 1993. O

BDI-IA, o que é comumente referido na literatura como simplesmente "BDI", é semelhante ao original, exceto no item prazo onde estende-se "durante a última semana, incluindo hoje", e alguns itens foram reformulados para evitar declarações duplas negativas. O BDI-II, publicado em 1996, contém uma revisão substancial do original e revisto BDI-IA, e omite itens relativos à imagem corporal e perda de peso, a hipocondria, e dificuldade em trabalhar para que a avaliação de sintomas correspondesse aos critérios para depressão do Manual Diagnóstico e Estatístico de Transtornos Mentais, quarta edição (DSM-IV). O prazo BDI-II estende-se por duas semanas para corresponder com os critérios do DSM-IV para transtorno depressivo maior.

O BDI original e suas versões posteriores têm sido amplamente aceitos e utilizados em psicologia e psiquiatria para avaliar a intensidade da depressão em populações psiquiátricas e normais. Estudos têm sido realizados em uma variedade de configurações usando populações clínicas como as mais diversas patologias. A confiabilidade e validade da versão em português brasileiro do BDI-II foram avaliadas por Gomes-Oliveira e cols. em uma amostra não clínica de adultos e mostrou ser uma ferramenta fidedigna e válida para mensurar sintomatologia depressiva na população brasileira (40).

1.4 Avaliação Psicométrica

1.4.1 Conceituação e Modelos

A psicometria tem duas vertentes: a teoria clássica dos testes (TCT) e a teoria de resposta ao item (TRI). O objetivo da psicometria é tentar explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas ou avaliações, tipicamente chamadas de itens. A TCT se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total (T). Por exemplo, o T em uma avaliação de 30 itens de nível de gravidade dos sintomas seria a soma dos itens assinalados como presente. Se for dado 1 para um item presente e 0 para um ausente, e o sujeito indicou 20 itens como presente e 10 como ausente, seu escore T seria de 20. A TCT, então, se pergunta “o que significa este 20 para o sujeito?”. A TRI, por outro lado, não está interessada no escore total em uma avaliação; ela se interessa especificamente por cada um dos 30 itens e quer saber qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser indicado como presente ou ausente (em avaliações de nível de sintomatologia por exemplo). No final, então, temos ou avaliações válidas (TCT) ou itens válidos (TRI), itens com os quais se poderão construir tantas avaliações válidas quantos se quiser ou o número de itens permitir.

1.4.2 Teoria de Resposta ao Item como ferramenta de análise

A TRI tem sido considerada por muitos especialistas como um marco para a psicometria (41). A TRI é na verdade um conjunto de modelos matemáticos que considera o item como unidade básica de análise (e não o escore total como na TCT), sendo adequada para situações em que o construto de interesse é representado por um (ou mais) atributo quantitativo não observável (denominado traço latente) avaliado através de variáveis observáveis categorizadas (instrumentos e escalas compostos por itens fechados com duas ou mais categorias de resposta, ordenadas ou não).

O traço latente é uma característica não observável do sujeito, que determina sua forma de responder à avaliação que é apresentada, sendo que o traço latente é qualificado na TRI com a letra grega *teta* – θ . Um modelo de traço latente especifica a relação entre a gravidade dos sintomas observável de um examinando em uma avaliação e o traço latente não observável ou intensidade do sintoma, que se supõe subjacente à avaliação no teste. Portanto, especificamente na avaliação dos sintomas psiquiátricos, os modelos da TRI têm por base a ideia de que os constructos psicológicos, como os sintomas da depressão estão latentes, ou seja, eles não são diretamente observáveis. O conhecimento sobre itens latentes pode ser obtido através da observação das respostas das pessoas a um conjunto de itens que avaliam o sintoma (42,43). Na TRI, a relação entre a estimativa da variável latente de uma pessoa (θ) e a probabilidade de que uma pessoa ofereça uma resposta a um determinado item pode ser descrita através de um modelo específico. Na prática, quando modelamos a pontuação dos itens de um indivíduo em uma escala através de um modelo TRI, as diferenças de escores da escala são expressas na métrica θ em vez da pontuação total da escala métrica.

A TRI assume o postulado de que a relação entre a avaliação dos sintomas e os traços latentes pode ser descrita por uma função matemática monotônica crescente, cujo gráfico é chamado de Curva Característica do Item – CCI (44). O modelo matemático padrão da CCI é a forma cumulativa da função logística. Portanto, a CCI representa a função que define a relação entre a probabilidade de resposta positiva a um determinado item (por exemplo, ter o sintoma) e o traço latente.

Figura 1. Curva Característica do Item de um item hipotético sob o modelo logístico de dois parâmetros (curva sólida).

Definindo o índice i para denotar item, o índice j para denotar indivíduo e considerando a variável resposta X igual a 1, se o indivíduo apresenta o respectivo sintoma, e 0, caso contrário, a equação que expressa o modelo logístico de dois parâmetros (ML2) é dada por:

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

O quanto um item contribui para avaliar o traço latente em questão é representado pela curva de informação do item (CII). Em geral, itens com maior parâmetro de discriminação (a) contribuem mais para aumentar a precisão das estimativas do traço

latente (θ) do que itens com menor valor de a . Além disso, essa precisão também aumenta quanto mais próximo o valor do parâmetro b estiver de θ (Figura 2).

Figura 2. Curva de Informação do Item de quatro itens sob o modelo logístico de dois parâmetros.

Um modelo TRI frequentemente usado para as notas do item politômicos é o modelo de resposta gradual (GRM) (45). O GRM é adequado para a análise de escalas clínicas com as categorias de resposta ordenadas, tais como escalas do tipo Likert, e vários pesquisadores usaram esse modelo para analisar escalas clínicas e avaliar a mudança (46,47).

1.4.3 Funcionamento Diferencial do Item

O termo Funcionamento Diferencial do Item, ou como denominado em inglês, *Differential Item Functioning* (DIF), tem como objetivo detectar se um item funciona igual ou diferentemente para grupos de distintas características, cujos sujeitos componentes tenham a mesma magnitude na variável medida. Essa expressão é a usada correntemente na literatura psicométrica e dizer que um item apresenta DIF significa que ele mostra diferentes propriedades estatísticas em razão de diferentes grupos.

A literatura tem feito distinção dos estudos de DIF quanto aos objetivos pretendidos (48). Portanto, se pode falar de investigações para a detecção do DIF e outras para verificar o impacto do DIF. No primeiro grupo estão as investigações que utilizam algum método tradicional para a identificação do DIF. Neste caso, os estudos objetivam somente a detecção do DIF, quer dizer, determinar a possível diferença entre as CCI dos itens de acordo com os grupos comparados. No segundo grupo estão as investigações para a busca e identificação das causas do DIF. Neste contexto, o objetivo do investigador que utiliza as distintas técnicas para a detecção do DIF, é tentar saber quais são as causas (psicológicas, educativas, culturais, sociais, atitudinais, etc.) que ocasionam o funcionamento diferencial de um item (49).

No contexto da TRI um item não tem DIF quando a CCI é idêntica para os grupos comparados em um mesmo nível ou magnitude da variável latente medida através do item (50). Geralmente os estudos para a determinação do DIF utilizam dois grupos, denominados de referência (GR) e focal (GF). Os distintos métodos para a detecção do DIF foram desenvolvidos baseados em dois tipos de DIF. O primeiro tipo é denominado DIF uniforme ou consistente, e se observa quando as CCI do item

estudado com respeito aos grupos de referência e focal são diferentes, porém não se cruzam. Em outras palavras, um DIF é considerado uniforme ou consistente quando não se detecta interação entre o atributo medido e o fato de pertencer a um determinado grupo. Este tipo de DIF está representado na Figura 3.

Figura 3. Representação gráfica de um item com Funcionamento Diferencial do Item uniforme.

Na Figura 3, observamos que a CCI das mulheres (grupo focal) está situada mais à esquerda que a dos homens (grupo de referência), indicando que o item é mais fácil para o grupo focal. Tal diferença no parâmetro b (dificuldade) supõe que o item possui DIF.

O segundo tipo de DIF é denominado DIF não uniforme ou inconsistente, e se observa quando as CCIs do item estudado, com respeito aos grupos de referência e focal, são diferentes e, ademais, se cruzam em algum ponto do intervalo atitudinal. Em outras palavras, quando há uma vantagem relativa para um desses grupos investigados, cujo valor é variável ao longo de todo o intervalo atitudinal. Este caso ocorre quando os parâmetros a , b ou c têm valores distintos nas duas CCIs, isto é, quando as CCIs não são paralelas. Este tipo de DIF está representado na Figura 4.

Figura 4. Representação gráfica de um item com Funcionamento Diferencial do Item não-uniforme.

A avaliação não é grupo-dependente, ou seja, as estimativas das dificuldades dos itens são independentes do grupo respondente do instrumento. Esta característica torna a TRI uma ferramenta muito útil na detecção de comportamento diferencial do item, conhecido como DIF, isto é, um item com comportamento diferente entre grupos de indivíduos (51).

Vários programas podem ser utilizados para ajustar modelos da TRI. Entre eles, podem ser citados: Bilog e Bilog-MG, Xcalibre, Multilog, Parscale, Noharm e Mplus.

1.4.4 Métodos de análise do Funcionamento Diferencial do Item

A TRI possibilita um caminho apropriado para a análise do DIF (50,52). A lógica subjacente à detecção do DIF se apoia em: estimar os parâmetros métricos dos itens para os grupos de comparação; colocá-los em uma mesma escala; representá-los através de suas CCIs; comparar os valores dos seus parâmetros entre os grupos escolhidos. Tendo a TRI como referência, um item terá DIF se sua CCI não é a mesma para os grupos comparados, cujos sujeitos tenham a mesma magnitude na variável latente θ (53). Por isso, o procedimento mais adequado para avaliar o DIF é a estimação da CCI para ambos os grupos e sua posterior comparação e se existir diferença significativa é possível afirmar que o item tem DIF (54). O desafio nesta investigação está em como medir com precisão a discrepância entre as CCIs de diferentes subgrupos. Existem vários diferentes procedimentos para o estudo do DIF.

As avaliações da equivalência de mensuração do item em nível de traço latente passaram a enfocar o DIF, definido como diferentes probabilidades de indicar ou endossar itens que se referem ao construto de grupos e com isso controlar a característica subjacente medida pelo teste (55). Existem algumas estratégias para detecção de DIF, incluindo a formulação de modelos de resposta a itens explicativos (56), a formulação de Múltiplos Indicadores de Múltiplas Causas (Multiple Indicators Multiple Causes - MIMIC) (57) e o método SIBTEST (58). O presente estudo tem como proposta usar o método de regressão logística ordinal, que permite detectar vários tipos de DIF (59).

O método de regressão logística ordinal é aprimorado com o uso das estimativas dos traços latentes obtidas através da TRI como critério de correspondência, no lugar do escore total, e ainda faz uso de um processo iterativo

que considera o DIF nas estimativas dos traços latentes em cada etapa, descartando a identificação errônea ou a não identificação em virtude da presença de DIF (59). A combinação destes métodos possibilita identificar várias fontes de DIF e ainda levar em consideração o impacto do DIF na mensuração do traço latente.

Neste método são criados três modelos de regressão logística ordinal para cada item. Em todos os modelos se adiciona um novo termo explicativo. O modelo 1, portanto, supõe que não há a presença de DIF. O modelo 2 supõe a presença de DIF uniforme, e o modelo 3 a presença de DIF uniforme e não uniforme.

Quando a inserção de um novo termo explicativo melhora significativamente o modelo, é sinal de que o item analisado apresenta DIF. Podemos comparar as diferenças entre os modelos a partir de testes da razão de verossimilhança, o que possibilita a identificação da presença, ou não, de DIF (60).

A diferença entre os modelos 1 e 2 é comparada com uma distribuição χ^2 com 1 grau de liberdade. Caso a estatística para o item seja significativa, o mesmo é identificado como tendo DIF uniforme. Fazendo-se a mesma comparação, considerando os modelos 2 e 3, pode-se identificar a presença do DIF não uniforme. Comparar diretamente os modelos 1 e 3, considerando uma distribuição χ^2 com 2 graus de liberdade, é uma forma de atestar a presença de DIF como um todo, sem definir se este é uniforme ou não uniforme.

Como o poder estatístico é associado ao tamanho da amostra, uma diferença muito pequena para efeitos práticos pode ser considerada estatisticamente significativa, dada uma amostra grande o suficiente. Assim, outros critérios foram propostos, tentando levar em consideração a magnitude desta diferença. Entre elas, foi proposto o uso de estatísticas pseudo- R^2 como uma medida de magnitude e classificação do DIF (61). São comparadas as estatísticas pseudo- R^2 entre os modelos,

e Zumbo (1999) sugere que uma diferença entre elas menor que 0,13 é insignificante, entre 0,13 e 0,26 é considerada moderada e maior do que 0,26 é considerada grande (61). Outros autores, entretanto, consideram que esses níveis são muito altos e, portanto, podem levar a não identificação de DIF (62,63). Sendo que isso vale, principalmente, para amostras relativamente grandes, onde o pseudo- R^2 , com mudança $\geq 0,02$, tem sido consistentemente utilizado como um critério para detecção de DIF (59). Com isso, se a diferença entre os modelos 1 e 3 mostra uma diferença $R^2 > 0,02$, então o item é considerado como mostrando DIF (59). Se a diferença entre os modelos 2 e 3 mostras uma diferença $R^2 > 0,02$, então o item é considerado como não-uniforme. DIF.

A literatura especializada propõe o uso da diferença entre o parâmetro β_1 dos modelos 1 e 2 para a detecção de DIF uniforme (64). Caso $\Delta\beta_1$ seja maior do que 10%, o item é identificado como tendo DIF. Entretanto, outros estudos usaram pontos de corte diferentes, como 5% ou até mesmo 1%, a fim de detectar efeitos diferenciais menores (64).

Em resumo, as investigações para a detecção e estudo do DIF estão baseadas em uma mesma argumentação: a existência do DIF é um fator que influencia a validade da interpretação, realizada a partir da pontuação obtida pelo sujeito num item ou teste. Na interpretação da pontuação reside toda a credibilidade e reputação da investigação.

1.5 Importância do presente estudo

Dentre os dados coletados no I Levantamento Nacional sobre o Uso de Álcool, Tabaco e Outras Drogas entre Universitários das 27 Capitais Brasileiras, encontra-se

o BDI-II, o qual foi respondido por 12.721 universitários. Esta população foi utilizada como referência amostral para avaliarmos os itens que compõem o BDI-II, objetivando analisar diferenças de gênero no preenchimento deste inventário.

Um modelo para a TRI muitas vezes utilizado com a pontuação de itens politômicos é o modelo de resposta gradual (MRG). O MRG é adequado para a análise de escalas psicométricas com as categorias de resposta ordenadas e crescentes (45). O MRG foi usado em nosso estudo para identificar os itens com comportamentos diferentes entre homens e mulheres, e com isso facilitar a compreensão das diferenças clínicas dos sintomas depressivos nesta população e ainda apoiar a formulação de novas hipóteses que forneçam suporte para futuras pesquisas na área.

É importante lembrar ainda que a presente proposta de estudo representa uma oportunidade de analisar, de uma forma inovadora, os sintomas depressivos em nossa amostra total e dividida por gênero. São poucas as pesquisas que usaram a ferramenta estatística TRI em sintomas depressivos e nenhuma dispunha de uma amostra comparada a nossa em tamanho e com características próprias da diversidade cultural do nosso país. Por tanto, a metodologia do nosso estudo se mostra original, considerando a população a ser estudada e as técnicas estatísticas que foram empregadas.

2. HIPÓTESES DO ESTUDO

Os estudos epidemiológicos evidenciam a preponderância da prevalência dos sintomas depressivos e da depressão unipolar nas mulheres em relação aos homens (20), e as manifestações clínicas da depressão têm se mostrado diferentes entre os gêneros, com as mulheres mais frequentemente relatando alterações do apetite e do sono, fadiga, ansiedade e relatos compatíveis com hipocondria somática (29–31). Pesquisas que colaborem com o entendimento, ou mesmo ofereçam suporte teórico na tentativa de esclarecer estas diferenças merecem especial atenção. Devemos considerar ainda a necessidade de aperfeiçoarmos as estratégias para detectar e mensurar os sintomas depressivos, contemplando as características da população ou do gênero estudados.

As pesquisas na área da psiquiatria frequentemente avaliam características subjetivas de indivíduos como, por exemplo, os sintomas depressivos e ansiosos. Os dados são coletados através de questionários, cujos itens tentam identificar a presença ou ausência de certos sintomas associados à morbidade psiquiátrica de interesse. A TRI entende que a melhor avaliação da gravidade da sintomatologia mensurada sobre as respostas de um indivíduo a itens de múltipla escolha, não é feita pela interpretação da soma da pontuação obtida (escore bruto). Os itens que compõem o instrumento de avaliação expressam algum aspecto do traço latente em estudo.

Nosso estudo tem como hipótese que a análise da nossa amostra total e dividida por gênero, com o uso da TRI, pode nos ajudar a identificar os itens mais relevantes e ainda indicar diferenças entre os gêneros no funcionamento dos itens que

compõem o BDI-II. Os resultados de nossa pesquisa podem permitir no futuro, um maior entendimento das diferenças clínicas e epidemiológicas da depressão em homens e mulheres.

3. JUSTIFICATIVA

A importância dos estudos que objetivam a verificação do DIF é justificada porque cabe ao autor do teste ou ao autor do sistema de avaliação ou ao pesquisador verificar se existem itens com DIF para (1) que se possam buscar as causas, (2) evitar sua utilização no grupo em desvantagem e, finalmente, (3) controlar os fatores responsáveis pelo DIF para evitar construir novos itens com o mesmo problema (52).

A Teoria de Resposta ao Item constitui-se em um novo grupo de modelos psicométricos de construção, validação de instrumentos de aferição em desfechos em saúde. A presente proposta de Tese de Doutorado aplicou a TRI em uma população de estudantes de graduação concentrando-se nas seguintes justificativas:

3.1. O BDI-II tem se mostrado útil na mensuração de sintomas depressivos em vários contextos clínicos. Contudo, acreditamos que grupos diferentes (homens e mulheres) devem apresentar DIF em algum dos 21 itens que compõem o BDI-II e esta análise deve refletir a melhor descrição da psicopatologia do grupo avaliado.

3.2. A existência de DIF entre os gêneros, nos itens que compõem o BDI-II, pode estar associada a algumas variáveis como características sociais e econômicas. O Estudo destas possíveis associações pode ajudar a compreender as características clínicas e epidemiológicas da depressão.

3.3. Existe, em nível nacional e internacional, uma grande carência de estudos que utilizaram a estratégia estatística baseada na TRI para análise da psicometria dos sintomas depressivos (65).

3.4. A literatura científica valoriza as robustas estratégias estatísticas que exploram abordagens de mensuração dos fenômenos psicológicos. Neste contexto, consideramos de grande relevância o uso da TRI na determinação dos traços latentes na BDI- II.

Portanto, considerando que em nosso estudo os participantes foram divididos por gênero masculino e feminino para análise e naturalmente acreditávamos que os grupos iriam se diferir entre si, com uma variação que se pode considerar natural, e que as características próprias que os distinguem têm consequências nos diversos modos de elaboração e desenvolvimento de instrumentos de avaliação em depressão. Comisso, ao longo de nosso estudo identificamos as características dos 21 itens que compõem o BDI-II, e identificamos o item que sinalizou a existência de maior relevância em cada gênero.

4. OBJETIVOS

4.1 Objetivos Gerais

Determinar diferenças na apresentação dos sintomas psicopatológicos coletados pelo BDI-II entre homens e mulheres, com o uso da TRI, em uma amostra representativa de universitários brasileiros.

4.2 Objetivos Específicos

- Descrever o desempenho psicométrico do BDI-II, de acordo com a Teoria Clássica de Medida.
- Analisar o traço latente de sintomas depressivos dos itens do BDI-II, discriminando quais são os itens mais relevantes em diferentes níveis de gravidade do traço alvo.
- Identificar os itens que apresentam funcionamento diferencial de itens entre homens e mulheres.

5. MATERIAL E MÉTODO

5.1 Delineamento do Estudo I Levantamento Nacional Sobre o uso de álcool, tabaco e outras drogas entre universitários das 27 capitais brasileiras

A Secretaria Nacional de Políticas sobre Drogas (SENAD), órgão do governo federal responsável por coordenar a implementação da Política Nacional sobre Drogas (PNAD) e da Política Nacional sobre o Álcool (PNA), em parceria com o Grupo Interdisciplinar de Estudos de Álcool e Drogas da Faculdade de Medicina da Universidade de São Paulo (GREAFMUSP), realizaram o I Levantamento nacional Sobre o uso de álcool, tabaco e outras drogas entre universitários das 27 capitais brasileiras, incluindo o distrito federal, o qual recebeu adesão do corpo diretivo de 100 Instituições de Ensino Superior (IES), públicas e privadas. Os temas abordados neste estudo foram:

1. dados sociodemográficos;
2. prevalência e padrão de uso do tabaco e outras substâncias psicoativas;
3. prevalência e padrão de uso do álcool;
4. uso múltiplo de drogas;
5. uso de drogas na Universidade de São Paulo;
6. prevalência de comportamentos de risco;

7. saúde mental incluindo a BDI-II;
8. comparação do uso de drogas entre os universitários e outros segmentos sociais e
9. políticas públicas sobre drogas nas IES.

A população-alvo eram alunos matriculados em universidades nas capitais brasileiras (66). Esta amostra foi recrutada através de amostragem por conglomerados em dois estágios. A primeira etapa envolveu IES que ministram cursos de graduação em cada uma das 27 capitais, de acordo com a lista escolar fornecida pelo Ministério da Educação. As IES elegíveis foram selecionadas aleatoriamente da lista de instituições ordenadas, por meio da probabilidade proporcional ao tamanho (PPS) do número de alunos matriculados em cada IES. A estratificação foi baseada nas duas variáveis: a localização da capital e o tipo de financiamento da instituição (privada ou pública), totalizando 54 níveis. Das 114 faculdades e universidades selecionadas, 100 IES (87,7%) concordaram em participar após convite institucional.

Na segunda etapa, a turma de alunos foi considerada a unidade primária de amostragem dentro de cada IES selecionada, para favorecer um trabalho de campo viável. Cada elemento da turma foi definido como um grupo de alunos que frequentam um determinado curso. O número de turmas foi proporcional ao total de alunos da instituição, totalizando 654 turmas ou 70,6% da unidade primária de amostragem selecionada. Todos os alunos das turmas selecionadas foram convidados a se voluntariar para o estudo. A taxa de resposta dos participantes foi de 95,6% para os estudantes universitários que estavam nas aulas no momento da pesquisa (66).

5.2 Caracterização da População do Estudo

A amostra da pesquisa foi de 12.711 estudantes universitários. Destes, 34 não relataram o gênero. Assim, uma amostra final de 12.677 questionários válidos foi considerada para análise. Com isso, a população-alvo foi definida como os universitários regularmente matriculados no ano letivo de 2009, em cursos de graduação presencial, de Instituições de Ensino Superior (IES) públicas e privadas, das 27 capitais brasileiras. Todos os estudantes completaram um instrumento estruturado, anônimo, autoadministrado, consistindo de 98 perguntas fechadas sobre o uso de substâncias, em que o Inventário de Depressão Beck de 21 itens (BDI-II) foi incluído para o estudo de sintomas depressivos.

As características demográficas da amostra foram constituídas por 5.692 homens (44,9%) e 6.985 mulheres (55,1%), com média de idade de 23,9 anos ($DP = 6,9$; variação de 16-84 anos). Em média, a idade média das mulheres foi semelhante à dos homens ($M = 23,8$ vs. $24,0$, $p = 0,15$). A análise da distribuição da idade mostrou a maioria dos participantes com idade entre o final da adolescência e 30 anos (16-30 anos; $n = 10.887$, $M = 21,7$ e $DP = 3,1$). Mesmo os participantes com idade superior aos anos 30 (31-84 anos; $n = 1790$, $M = 38,7$ e $DP = 7,0$) também, em sua maioria, se concentravam em faixas etárias mais jovens (31-40 anos 66,8%; 41-50 anos 26%; 51-60 anos 6,6%; 61-70 anos 0,4% e 71-84 anos 0,2%).

Além disso, a cor da pele autorreferida como etnia no Brasil mostrou que a maioria da amostra era composta por estudantes brancos (55,5%), seguida por negros (6,8%), pardos (29,8%), asiáticos (2,9%), indianos (0,8%), outras não determinadas (3,0%) e 1,3% sem resposta. Quando perguntados sobre “Qual é a sua religião atual?” as respostas foram: 16,3% sem religião, 53,0% católicos, 18,4% evangélicos ou

protestantes, 11,8% outras religiões e 0,5% sem resposta. O estado civil apresentou as seguintes taxas: 80,5% solteira, 16,9% casada, 1,7% divorciada, 0,1% viúva e 0,7% sem resposta.

5.3 Características do instrumento Inventário de Depressão de Beck, versão II

A segunda versão do BDI é um instrumento de autoaplicação para avaliar a presença de sintomas depressivos e contém 21 questões autorrelatadas com afirmações sobre como o indivíduo se sentiu nos últimos 15 dias. Cada item apresenta quatro afirmações que correspondem a níveis crescentes de sintomas depressivos com pontuações entre 0 e 3. As pontuações totais possíveis variam entre 0 a 63. As questões 16 e 18 refletem as alterações de sono (insônia e hipersonia) e apetite (hiperfagia e hiporexia), respectivamente. Esse questionário tem sido amplamente utilizado em pesquisas médicas e psiquiátricas no mundo todo (67). Embora a reformulação do BDI-II tenha claros objetivos para se adequar aos critérios diagnósticos do DSM-IV, este instrumento não serve para fazer diagnóstico psiquiátrico, por não envolver avaliação clínica.

Para distinguir os níveis de gravidade estimaremos a intensidade dos sintomas depressivos dos participantes com os seguintes intervalos de pontuação: mínima (0-11), leve (12-19), moderada (20-35) e grave (36-63). O conteúdo da escala reflete os sintomas cognitivos, afetivos, somáticos e vegetativos da depressão (68).

5.4 Análise Estatística

Os dados foram tabulados no Microsoft Excel®, sendo exportados para os softwares SPSS versão 21 para Windows, o qual foi usado para analisar as estatísticas descritivas, e posteriormente exportados os softwares Mplus 6.12 (www.statmodel.com) e R versão 3.2.2 (R Core Team, 2015). O nível de significância foi estabelecido em 0,05 para testes bicaudais. Inicialmente, calculou-se a consistência interna com o coeficiente Alfa de Cronbach. Características da dispersão e homogeneidade dos dados do BDI-II também foram exploradas para cada item da escala e um escore total: média (M), desvio padrão (DP), coeficiente de variação de Pearson (CV) e a consistência interna obtida com o cálculo do Alfa de Cronbach se o item for excluído.

Para a execução da análise da TRI duas premissas são necessárias: unidimensionalidade (isto é, podemos aceitar apenas um fator subjacente nos dados) e independência local (os itens não devem ser correlacionados quando a variância compartilhada do traço latente é removida). Ambas as premissas foram testadas através de análises fatoriais confirmatórias (AFC), para examinar a estrutura de covariância da escala com o uso do software Mplus 6.12 (www.statmodel.com).

Para estimar os parâmetros de ajustamento do modelo aos dados da amostra foram utilizados índices de ajustamento absoluto: χ^2 (qui-quadrado) e RMSEA (Raíz quadrada média dos quadrados dos erros de aproximação). Para avaliar o ajuste incremental foram utilizados dois índices: CFI (índice de ajuste comparativo) e TLI (Índice de Tucker-Lewis) (69). Os seguintes pontos de corte foram considerados como indicativos de um ajuste adequado do modelo: RMSEA < 0,06, CFI e TLI $\geq$ 0,95 (70,71). Os resultados da AFC mostraram os dados adequadamente ajustados ao

modelo unidimensional: $\chi^2 (210) = 9054,21$, $p < 0,001$, CFI = 0,95, TLI = 0,94 e RMSEA = 0,06. A significância do teste do χ^2 foi ampliada devido ao grande tamanho da amostra. Esses indicadores favoreceram o uso do modelo de TRI (71).

Para uma escala ordinal politômica, o modelo de resposta gradual (45) foi adotado para descrever a relação entre as respostas observadas para o BDI-II e o traço latente subjacente θ (teta), como representativo da gravidade da depressão. Este modelo produz parâmetros de discriminação (a) e de gravidade (b_i) para cada item da escala que descreve sintomas da depressão. Itens da escala (i) com uma grande inclinação se mostram com maior capacidade de discriminação (a) dos sintomas depressivos. Para interpretação, Baker (2001) sugere que valores de parâmetros de discriminação variando de 0,01 a 0,24 são considerados muito baixos, de 0,25 a 0,64 baixos, de 0,65 a 1,34 moderados, de 1,35 a 1,69 de altura e acima de 1,7 de muito altos (72). O parâmetro dificuldade do item (b_i) é distribuído ao longo de uma dimensão latente θ (teta) para refletir o traço de gravidade em relação à ausência ou presença do estado patológico em estudo. Em ciências da saúde, o escore b de um dado item da escala (i) é usado para localizar sua posição no traço latente θ e quanto mais a variável se posiciona para o lado direito, mais se denota uma manifestação de maior gravidade da doença alvo (72). O pacote Mirt (73) do software livre R foi utilizado para estimar os parâmetros da TRI via modelo de Resposta Gradual de Samejima.

As análises de DIF foram calculadas no software R, onde se utilizou o pacote Lordif (74) juntamente com a interface gráfica em TCL/TK. As análises foram realizadas para identificar itens tendenciosos que mostrassem funcionamento diferente dos itens quando comparados por subgrupo, onde o gênero dos entrevistados foi dicotomizado na variável Grupo como homens ou mulheres. O pacote Lordif

permite analisar a possível existência de DIF por três métodos: Testes de Razão de Verossimilhança, Pseudo- R^2 e a Diferença entre os Parâmetros β_1 ($\Delta\beta_1$). Para isso, na análise da checagem de itens com DIF, utilizou-se o método híbrido de Regressão Logística Ordinal/TRI, que utiliza os escores da TRI como Critério de Correspondência para a regressão. Este oferece uma escolha mais adequada para a detecção de DIF, já que não utiliza a simples soma das pontuações como na Teoria Clássica dos Testes (72,75). Para realizar a equiparação dos thetas dos grupos, foi utilizado o método de purificação. Este consiste em, inicialmente, estimar os itens que provavelmente apresentam DIF e, em seguida, retirá-los da pontuação da variável latente (theta), depois, com a nova pontuação, estimam-se novamente os itens com DIF; as purificações são realizadas até que nenhum item com DIF contribua para a pontuação.

Procurando detectar a existência e o tipo de DIF, utilizou-se uma Regressão Logística Ordinal hierárquica de três passos, em que o primeiro verificou se o critério de correspondência (traço latente) predizia a resposta ao item; posteriormente, inseriu-se a variável Grupo juntamente com Critério de Correspondência; e, finalmente, no terceiro passo, foram consideradas as duas variáveis anteriores e a interação entre elas (Grupo e a Variável Correspondente). Quando a inserção de uma nova variável explicava significativamente a resposta ao item, sugere-se DIF; porém, como esse critério é vulnerável a grandes amostras, optou-se por comparar as diferenças entre os três passos por meio da estatística de verossimilhança χ^2 e Pseudo- R^2 (76).

Como o poder estatístico é associado ao tamanho da amostra, uma diferença muito pequena para efeitos práticos pode ser considerada significativa, dada uma amostra grande como a do presente estudo. Assim, tentando levar em consideração a

magnitude desta diferença, foi proposto como critério final o uso de estatísticas do pseudo- R^2 como medida de magnitude desta diferença (76).

Primeiro, se a diferença entre os modelos 1 e 3 mostrasse uma diferença $R^2 > 0,02$, o item marcado seria considerado como apresentando DIF (59). Segundo, se a diferença entre os modelos 2 e 3 mostrasse uma diferença de $R^2 > 0,02$, então o item seria considerado como mostrando "DIF não uniforme" porque o efeito varia de acordo com o nível de característica. Terceiro, se a diferença entre os modelos 1 e 2 mostrasse uma diferença $R^2 > 0,02$, o item seria considerado como apresentando "DIF uniforme", porque o efeito seria constante. Nos casos em que apenas os modelos 1 e 3 se diferenciam estatisticamente, não é possível definir se o item possui DIF uniforme ou não uniforme.

Apesar do método Pseudo- R^2 ter sido o caminho utilizado devido ao tamanho de nossa amostra, neste estudo, o último método para analisar ($\Delta\beta_1$) DIF também foi avaliado. Quando em uso da análise pelo $\Delta\beta_1$ $\{\Delta\beta_1 = [(\beta_{1\text{modelo2}} - \beta_{1\text{modelo1}}) / \beta_{1\text{modelo1}}]\}$ Crane et al em 2004(64), relataram que o $\Delta\beta_1$ menores que 0,1 (10%) indicaria ausência de DIF, mas também sugeriram 0,05 (5%) e até mesmo 0,01 (1%) como critérios mais intensos e menos permissivos.

Por fim, traçamos a curva característica do item (ICC) dos itens com DIF para ambos os sexos para inspecionar a direção da diferença, e ainda a curva característica do teste (TCC) para fornecer uma visão do impacto real dos itens com DIF na pontuação total.

5.5 Aspectos Éticos

No presente estudo realizamos uma análise de dados já coletados, sendo que os sujeitos entrevistados receberam explicação dos procedimentos do estudo e um termo de consentimento livre e esclarecido (TCLE) foi assinado pelos alunos, de acordo com a solicitação da Comissão de Ética para Análise de Projetos em Pesquisa - CAPPESq da Diretoria Clínica do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (sob o número de protocolo 1203/07, aprovado em 30/1/2008). Este projeto foi submetido ao Comitê de Ética da Faculdade de Medicina da Universidade de São Paulo – FMUSP e foi aprovado em 10/09/2015 com o número 48697015.2.0000.0065 do Certificado de Apresentação para Apreciação Ética (CAAE).

6. RESULTADOS

6.1 Confiabilidade e análise descritiva

A confiabilidade avaliada pelo alfa de Cronbach foi de 0,91 e não diferiu entre os sexos (homens = 0,90; mulheres = 0,91). As estatísticas descritivas de respostas obtidas de 12.677 participantes para o BDI-II são mostradas na Tabela 1, sendo que a pontuação média total foi de 6,99 (SD 7,56, CV 1,08). A média do escore total do BDI-II entre mulheres ($M = 7,85$, $DP = 8,08$) e homens ($M = 5,92$, $DP = 6,72$) apresentou diferença estatisticamente significativa ($p < 0,001$). Além disso, usando o ponto de corte recomendado de 13 (68), a proporção de provável depressão também foi significativamente diferente em relação ao sexo (homens = 14,2%; mulheres = 22,1%; $p < 0,001$).

Tabela 1: Características psicométricas do Inventário de Depressão de Beck-II para uma amostra de estudantes universitários ($N = 12,677$)

Item	<i>M</i>	DP	CV	<i>A</i>
1. Tristeza	0.21	0.46	2.05	0.90
2. Pessimismo	0.22	0.48	2.05	0.90
3. Sentimento de fracasso	0.23	0.57	2.25	0.90
4. Perda de prazer	0.26	0.52	1.83	0.90
5. Sentimento de culpa	0.33	0.53	1.53	0.90
6. Sentimentos de punição	0.22	0.62	2.33	0.90
7. Autoestima	0.24	0.63	2.50	0.90
8. Autocrítica	0.56	0.72	1.14	0.90
9. Ideias suicidas	0.08	0.32	3.57	0.90
10. Choro	0.36	0.77	1.91	0.90
11. Agitação	0.35	0.61	1.55	0.90
12. Perda de interesse	0.26	0.52	1.83	0.90
13. Indecisão	0.41	0.78	1.61	0.90
14. Desvalorização	0.17	0.52	2.67	0.90
15. Perda de energia	0.44	0.62	1.29	0.90
16. Alterações no sono	0.66	0.75	0.98	0.90
17. Irritabilidade	0.37	0.61	1.48	0.90
18. Mudanças no apetite	0.48	0.69	1.29	0.90
19. Dificuldade de concentração	0.46	0.72	1.38	0.90
20. Cansaço ou fadiga	0.51	0.70	1.20	0.90
21. Perda da libido	0.15	0.45	2.75	0.90
Total	6.99	7.56	1.08	0.91

M: Média

DP: Desvio Padrão

CV: Coeficiente de variação de Pearson = DP/*M*

α : Consistência interna obtida com o cálculo do Alfa de Cronbach se o item for excluído

A Tabela 2 apresenta as distribuições de frequência para cinco itens arbitrariamente selecionados: “tristeza” (nº 1), “sentimento de culpa” (nº 5), “choro” (nº 10), “perda de energia” (nº 15) e “perda de libido” (nº 21). As distribuições de frequência para os 21 itens são apresentadas em uma tabela na seção de anexos. Ao inspecionar a tabela, a grande maioria dos entrevistados endossou a categoria 1 dos itens, que denota “como de costume” ou “inalterada” manifestação de sintomas depressivos. As categorias sintomáticas de 2, 3 e 4 foram endossadas em frequência decrescente, intensidade ou sofrimento, para indicar a gravidade dos sintomas depressivos. Como esperado para esta amostra não clínica de estudantes, quanto mais grave o sintoma, menos ele foi escolhido pelos respondentes. Examinando as colunas para sexo masculino e feminino, uma diferença considerável foi observada para itens específicos. Por exemplo, as mulheres endossaram mais “tristeza”, “choro” e “perda de energia” do que os homens.

Tabela 2: Proporção da resposta do BDI-II (%) para as alternativas de categorias de escolha nos itens selecionados

Item	Amostra total N=12.677	Homens n = 5.692	Mulheres n = 6.985
1. Tristeza			
Categoria 1	81.3	85.7	77.6
Categoria 2	17.5	13.2	20.9
Categoria 3	0.7	0.5	0.8
Categoria 4	0.6	0.5	0.6
5. Sentimento de culpa			
Categoria 1	69.9	72.4	67.9
Categoria 2	28.1	25.9	30.0
Categoria 3	1.4	1.3	1.5
Categoria 4	0.6	0.5	0.6
10. Choro			
Categoria 1	77.6	88.4	68.9
Categoria 2	12.1	6.4	16.7
Categoria 3	6.5	0.8	11.2
Categoria 4	3.8	4.5	3.1
15. Perda de energia			
Categoria 1	61.6	66.3	57.7
Categoria 2	33.2	30.1	35.8
Categoria 3	4.5	3.2	5.7
Categoria 4	0.7	0.5	0.9
21. Perda da libido			
Categoria 1	88.0	91.9	84.8
Categoria 2	9.4	6.9	11.4
Categoria 3	2.0	0.9	2.9
Categoria 4	0.6	0.3	0.9

Nota: Em geral, a categoria 1 indica "como de costume"; categoria 2 "mais do que o habitual"; categoria 3 "o tempo todo"; e categoria 4 "sofrimento extremo". Embora a redação do item possa variar de acordo com o sintoma depressivo, as categorias são exibidas em nível crescente de gravidade.

A proporção de respostas para cada item não foi uniforme, cuja variação sugere uma probabilidade diferencial de taxa de endosso de acordo com a discriminação e o nível de dificuldade de um item no BDI-II. Além disso, esses resultados indicam diferenças notáveis no padrão de resposta relacionado ao sexo biológico nos sintomas depressivos autor relatados. A análise de itens e o funcionamento diferencial dos itens (DIF) são explorados a seguir para entender seu impacto na diferença entre os grupos de depressão, tanto pelo escore médio quanto pelo limiar de corte.

6.2 Parâmetros dos Itens pela Teoria da Resposta ao Item

A Tabela 3 mostra as estimativas de discriminação (a) e limiar de gravidade (b_i) de sintomas depressivos para todos os 21 itens. Os parâmetros de discriminação (a) variaram de 1,32 (nº 21 “perda de interesse em sexo”) a 3,31 (nº 14 “inutilidade”), sugerindo que todos os itens da escala apresentaram características discriminatórias de moderada a alta para o traço latente de “depressão relatada” (θ).

Tabela 3: Parâmetros estimados do modelo de resposta gradual (GRM) dos itens do BDI-II ($N = 12,677$)

Item	a	(EP)	b_1	(EP)	b_2	(EP)	b_3	(EP)
1. Tristeza	2.38	(0.06)	1.10	(0.05)	2.75	(0.12)	3.13	(0.15)
2. Pessimismo	2.20	(0.05)	1.12	(0.05)	2.57	(0.09)	3.35	(0.16)
3. Sentimento de fracasso	2.04	(0.05)	1.32	(0.05)	1.98	(0.07)	3.41	(0.14)
4. Perda de prazer	2.27	(0.05)	0.95	(0.05)	2.34	(0.08)	3.25	(0.15)
5. Sentimento de culpa	1.72	(0.04)	0.75	(0.03)	2.96	(0.08)	3.78	(0.13)
6. Sentimento de punição	1.74	(0.05)	1.46	(0.05)	2.39	(0.07)	2.63	(0.07)
7. Autoestima	2.86	(0.07)	1.24	(0.07)	1.60	(0.09)	2.54	(0.14)
8. Autocrítica	1.62	(0.03)	0.25	(0.03)	1.78	(0.04)	3.24	(0.08)
9. Ideias suicidas	2.12	(0.07)	1.89	(0.08)	3.22	(0.15)	3.60	(0.19)
10. Choro	1.52	(0.04)	1.16	(0.03)	1.93	(0.04)	2.73	(0.06)
11. Agitação	1.59	(0.04)	0.83	(0.03)	2.90	(0.06)	3.19	(0.08)
12. Perda de interesse	2.73	(0.06)	0.89	(0.05)	2.26	(0.11)	2.87	(0.15)
13. Indecisão	2.03	(0.04)	0.78	(0.04)	1.88	(0.06)	2.12	(0.06)
14. Desvalorização	3.31	(0.09)	1.31	(0.10)	1.80	(0.12)	2.72	(0.20)
15. Perda de energia	2.35	(0.05)	0.40	(0.04)	1.96	(0.07)	3.05	(0.13)
16. Alterações no sono	1.58	(0.03)	0.03	(0.03)	1.69	(0.04)	3.04	(0.07)
17. Irritabilidade	2.13	(0.04)	0.65	(0.04)	2.10	(0.07)	2.93	(0.11)
18. Mudanças no apetite	1.40	(0.03)	0.48	(0.03)	2.31	(0.04)	3.28	(0.07)
19. Dificuldade de concentração	2.07	(0.04)	0.57	(0.03)	1.60	(0.05)	2.84	(0.09)
20. Cansaço ou fadiga	2.24	(0.04)	0.32	(0.03)	1.75	(0.06)	2.52	(0.09)
21. Perda da libido	1.32	(0.04)	1.91	(0.04)	3.26	(0.07)	4.44	(0.12)

a: parâmetro de discriminação

b_i : parâmetros de localização do nível de gravidade no item

EP: erro padrão da estimativa do parâmetro

Todos os parâmetros de gravidade ou dificuldade (b_i) apresentaram valor positivo, no lado direito do θ . Os três parâmetros de dificuldade do item (b_1 , b_2 e b_3) variaram de 0,03 a 1,91, 1,60 a 3,26 e 2,12 a 4,44, respectivamente. A sequência ascendente dos limiares b_1 , b_2 e b_3 confirmou a direção apropriada das opções de resposta. Os parâmetros de localização do item BDI-II (limiar b_i) foram bastante consistentes ao longo de um gradiente crescente do escore em termos de traço latente θ .

Para o parâmetro gravidade do item b_1 , os itens mais fáceis (indicando menor gravidade) do BDI-II foram o nº 16 “alterações no sono”, o nº 8 “autocrítica” e o nº 20 “cansaço ou fadiga”. Os itens mais graves para b_1 foram nº 21 “perda da libido”, nº 9 “ideias suicidas” e nº 6 “sentimentos de punição”. Na mesma direção, para o parâmetro de limiar b_3 , os itens BDI-II mais fáceis eram “indecisão” nº 13, “cansaço ou fadiga” nº 20 e “autoestima” nº 7. Os itens de maior gravidade em b_3 foram nº 21 “perda da libido”, nº 5 “sentimento de culpa” e nº 9 “ideias suicidas”. Vale a pena notar que o pequeno valor do erro padrão (EP) em todos os parâmetros indicou a boa precisão das estimativas nesta grande amostra. Além disso, quanto maior a taxa de endosso de um dado item, menor ou mais preciso era o valor do EP.

Na Figura 5 está a representação gráfica do Modelo Resposta Gradual de Samejima para os 21 itens do BDI-II com 4 categorias de resposta medindo a intensidade de sintomas depressivos com as estimativas dos parâmetros a e b_i .

Figura 5: Curva Característica do Item para os 21 itens

A Curva Característica do Item apresenta graficamente, para cada um dos 21 itens, a relação do nível de intensidade de sintomas depressivos pela estimativa dos pontos de interseção (b_1 , b_2 e b_3) com a probabilidade de resposta a uma categoria de cada sintoma depressivo. Com o primeiro ponto de interseção (b_1), para cada um dos 21 itens, temos a posição que pode ser interpretada como uma medida da gravidade do sintoma depressivo na interseção entre a ausência de sintoma e o primeiro nível de gravidade do sintoma representado por um determinado item do BDI. O parâmetro c , conhecido como endossar um item ao acaso ou nas análises na área da educação como acerto por sorte (chute), não faz sentido em análises no campo da saúde.

As Curvas de Informação dos Itens na Figura 6 são representações gráficas da função de informação dos itens do BDI-II. Os itens desvalorização (14), perda de interesse (12) e autoestima (7) foram os mais informativos.

Figura 6: Curva de informação dos 21 itens

Como os parâmetros que indicam dificuldade dos 21 item forneceram informações desiguais sobre a distribuição dos limiares de escore, examinamos o padrão de resposta para subgrupos de gênero, por meio da análise de DIF.

6.3 Funcionamento Diferencial do Item

Na análise do Funcionamento Diferencial do Item (DIF) da escala BDI-II, a qual é composta por 21 itens politômicos com alternativas com escores que seguem a presença crescente de sintomas variando de 0 a 3, foi feito a equiparação entre os thetas de cada subgrupo a saber: 1 – Homens; 2 – Mulheres.

Para detectar o DIF de uma escala, três modelos de regressão logística ordinal para cada item foram criados: o Modelo 1 que supõe que não há presença de DIF; o Modelo 2 que supõe a presença de DIF uniforme; e o Modelo 3 que supõe a presença de DIF uniforme e não uniforme. A diferença entre os modelos foi realizada a partir de testes de razão de verossimilhança, o que possibilita a identificação de presença ou não de DIF.

A diferença entre os modelos 1 e 2 [$\Pr(\chi^2_{(1;2)})$] foi comparada com uma distribuição χ^2 com 1 grau de liberdade; caso a estatística fosse significativa ($p < 0.01$), o mesmo seria identificado como tendo DIF uniforme. De igual forma, considerando-se os modelos 2 e 3 [$\Pr(\chi^2_{(2;3)})$], pode-se identificar DIF não uniforme. A comparação do modelo 1 e 3 [$\Pr(\chi^2_{(1;3)})$] foi apenas uma forma de atestar presença de DIF, sem definir se é uniforme ou não uniforme.

A Tabela 4, apresenta os critérios utilizados para detectar o DIF dos itens, o primeiro com a Razão de Verossimilhança, o segundo com as diferenças entre os Pseudo- R^2 e o terceiro com os $\Delta\beta$ s resultantes dos modelos das regressões.

Tabela 4: Critérios Utilizados para Detectar o Funcionamento Diferencial dos Itens

Itens	Testes de Razão de			Diferença entre os			$\Delta(\beta 1)$
	Verossimilhança (p)			Pseudo- R^2			
	$\Pr(\chi^2$ (1;2))	$\Pr(\chi^2$ (1;3))	$\Pr(\chi^2$ (2;3))	R^2 (1;2)	R^2 (1;3)	R^2 (2;3)	
Item 1	0,000	0,000	0,000	0,004	0,005	0,000	0,005
Item 2	0,045	0,000	0,001	0,000	0,001	0,000	0,003
Item 3	0,000	0,000	0,192	0,002	0,002	0,000	0,013
Item 5	0,448	0,002	0,000	0,000	0,000	0,000	0,000
Item 6	0,000	0,000	0,919	0,002	0,002	0,000	0,013
Item 8	0,000	0,000	0,065	0,000	0,000	0,000	0,005
Item 9	0,000	0,000	0,015	0,002	0,003	0,000	0,013
Item 10	0,000	0,000	0,000	0,023	0,024	0,001	0,027
Item 11	0,000	0,000	0,009	0,000	0,001	0,000	0,004
Item 12	0,872	0,012	0,003	0,000	0,000	0,000	0,000
Item 13	0,000	0,000	0,293	0,004	0,004	0,000	0,004
Item 15	0,000	0,000	0,432	0,002	0,002	0,000	0,002
Item 16	0,256	0,002	0,000	0,000	0,000	0,000	0,001
Item 17	0,000	0,000	0,868	0,002	0,002	0,000	0,003
Item 18	0,000	0,000	0,704	0,009	0,009	0,000	0,003
Item 19	0,000	0,000	0,010	0,001	0,001	0,000	0,003
Item 20	0,000	0,000	0,648	0,002	0,002	0,000	0,003
Item 21	0,000	0,000	0,000	0,010	0,011	0,002	0,012

Nota. $\Pr(\chi^2_{(1;2)})$ – Comparação da razão de verossimilhança entre os passos um (01) e dois (02) da regressão. $\Pr(\chi^2_{(1;3)})$ – Comparação da razão de verossimilhança entre os passos um (01) e três (03) da regressão. $\Pr(\chi^2_{(2;3)})$ – Comparação da razão de verossimilhança entre os passos dois (02) e três (03) da regressão. $R^2_{(1;2)}$ – Diferença entre os Pseudo- R^2 dos passos um (01) e dois (02) da regressão. $R^2_{(1;3)}$ – Diferença entre os Pseudo- R^2 dos passos um (01) e três (03) da regressão. $R^2_{(2;3)}$ – Diferença entre os Pseudo- R^2 dos passos dois (02) e três (03) da regressão.

Com isso, por meio da razão de verossimilhança, 18 itens foram identificados com provável DIF:

- Para os itens 3, 6, 8, 9, 13, 15, 17, 18, 19 e 20 pode-se considerar o DIF não uniforme.
- Para os itens 2, 5, 12 e 16 pode-se considerar o DIF uniforme.
- Para os itens 1, 10, 11 e 21 pode-se considerar o DIF indeterminado.

O uso da diferença entre os parâmetros β_1 (parâmetros de regressão associados a cada termo explicativo) dos modelos 1 e 2 para a detecção de DIF uniforme também foi considerado. Crane, van Belle, e Larson (2004) relataram que o $\Delta\beta_1$ menor que 0,1 (10%) indicaria ausência de DIF, mas também sugeriram 0,05 (5%) e até mesmo 0,01 (1%) como critérios mais intensos e menos permissivos. Com isso, nós encontramos para os itens sentimento de fracasso (item 3), sentimento de punição (item 6), ideias suicidas (item 9) e perda da libido (item 21) valores de $\Delta\beta_1$ maiores que 0,01 sugestivos de DIF pelo critério mais permissivo (64).

Considerando-se que pelo elevado tamanho da nossa amostra os métodos anteriormente citados (testes de razão de verossimilhança e diferença entre os parâmetros β_1) podem superestimar o resultado, utilizou-se no presente estudo como medida de magnitude e classificação do DIF apenas a estatística Pseudo- R^2 . Desta forma, em relação ao DIF relacionado ao gênero, o método detectou o item nº 10 (“choro”) como um elemento problemático; com mulheres mais provavelmente endossando esse sintoma do que homens. Em análises, dos subgrupos divididos pelo gênero, implementadas por regressão logística ordinal (OLR), a diferenças entre os modelos 1 e 3 foi 0,024, o que indicou DIF global. Por outro lado, a diferença do

pseudo R^2 entre os modelos 2 versus 3 foi $<0,001$ para o item nº10 e indica a presença de DIF uniforme.

A curva característica do item (ICC) mostrou que os participantes do gênero feminino apresentam maior probabilidade de endossar o item 10 do que do gênero masculino (Figura 7). A curva de resposta para o item 10 é exibida em uma sequência ordenada ao longo do traço latente θ também é exibida na Figura 7. Curvas pontilhadas indicam que, dado um nível semelhante de traço latente, as mulheres tiveram uma chance maior de endossar esse item do que os homens. A partir de uma investigação adicional da ICC para os entrevistados do sexo masculino, é indicado que o subgrupo de homens apresentou uma sobreposição substancial para o limiar de b_1 e b_2 no item 10. Isso sugere que o limite para endossar o “choro” no subgrupo masculino poderia ser recalibrado.

Figura 7: O primeiro gráfico representa a pontuação que representa o construto avaliado (pontuação da teoria da resposta ao item) da amostra de estudantes universitários sobre o Funcionamento Diferencial dos Itens, para o item 10 (Choro). O segundo gráfico mostra a Curva Característica do Item para o item 10. As curvas mostram a probabilidade de endossar um item específico

A Figura 8 mostra a curva característica do teste (TCC) para todos os itens dos subgrupos masculino e feminino. Os números mostram o impacto geral do DIF para gênero nos TCCs para todos os 21 itens do BDI (ignorando o DIF) e o impacto individual do item 10 nortado pelo DIF no TCC. Em uma análise parcial do gráfico fica claro que não houve influência substancial de DIF no item 10 na pontuação total do domínio, por causa das inclinações e sobreposições similares nas pontuações entre as duas sub-amostras. Portanto, esses achados indicaram um impacto mínimo de DIF por gênero, embora o escore total esperado fosse maior para mulheres do que para homens. Para resumir, embora o item 10 mostrasse DIF, esse item apresentou impacto insignificante na pontuação total.

Figura 8: Impacto geral do Funcionamento Diferencial do Item (DIF) para gênero nas Curvas Características do Teste (TCC). O primeiro gráfico exibe o TCC de todos os itens e o segundo gráfico exibe o TCC de itens específicos do DIF

7. DISCUSSÃO

O presente estudo trouxe informações relativas as propriedades psicométricas do BDI-II e ao viés de gênero na avaliação do padrão de respostas dos sintomas depressivos, também denominado Funcionamento Diferencial do Item, em uma grande amostra representativa de estudantes universitários brasileiros. Iniciamos as nossas análises com base no arcabouço da Teoria Clássica do Teste, onde o BDI-II mostrou consistência interna suficiente para capturar a construção subjacente da “depressão maior”. Com isso, embora todos os 21 itens tenham provado serem capazes de discriminar bem o traço latente de "depressão", alguns itens foram examinados para avaliarmos a possibilidade de DIF, segundo a teoria da resposta ao item, em relação ao gênero dos respondentes. Os principais resultados indicaram que o sintoma auto relatado de "choro" poderia modificar a estimativa de depressão de acordo com o gênero. Assim, este estudo psicométrico por itens forneceu uma oportunidade para comparar a medida da depressão entre os grupos masculino e feminino.

O conceito subjacente para os dados apresentados neste trabalho se baseia no constructo depressão avaliado pelos 21 itens da BDI-II, sendo que nesta escala os itens são pontuados na direção prejudicada, refletindo a sintomatologia da depressão. A modelagem estatística com o uso da TRI oferece um caminho promissor para a busca de evidências de um funcionamento diferente dos itens entre diferentes subgrupos, sendo que o viés que uma variável pode produzir na avaliação do padrão

de resposta dos sintomas depressivos pode comprometer a consistência dos resultados, levando a conclusões imprecisas.

Quanto ao poder de discriminação (parâmetro a), a análise de resposta do item indicou que todos os itens do BDI-II apresentaram capacidade pelo menos moderada (72) para discriminar o traço latente ao longo do espectro de gravidade da “depressão”. Comparativamente, os itens somáticos (“perda da libido” e “mudanças no apetite”) foram menos informativos do que itens afetivos (“desvalorização” e “autoestima”). Portanto, todos os itens do BDI-II foram adequados para representar a depressão entre estudantes universitários.

Em relação ao nível de dificuldade (o parâmetro b_i), os itens mais endossados são considerados os menos graves (77–79). O excesso de sintomas depressivos abaixo do limiar nesta amostra de estudantes sugeriu que os itens do BDI-II eram difíceis de endossar, mas capazes de detectar sintomas depressivos em uma ampla faixa de gravidade. Se uma escala destinada a medir a depressão fosse composta apenas por itens que medem quadros depressivos leves, o instrumento teria grande dificuldade para distinguir entre casos moderados e graves de depressão. Isso se justificaria pelo fato de que em ambos os casos teríamos pontuações altas em todos os itens. Por outro lado, se os itens da escala detectam facilmente os sintomas somáticos da depressão, os fenótipos cognitivo-afetivos da depressão podem permanecer despercebidos (80–82). Em geral, os itens do BDI-II abrangem tipos somáticos ou tipos de depressão cognitivo-afetivos em uma única dimensão de gravidade.

Em teoria, os indivíduos com um determinado nível de gravidade da depressão também devem ter a mesma probabilidade de pontuar uma determinada resposta do item. Quando homens e mulheres com o mesmo nível de gravidade da depressão respondem sistematicamente a um item de maneiras diferentes, o item não está

funcionando corretamente para diferentes grupos de entrevistados (83). Portanto, a análise DIF examina se a probabilidade de endosso do item é igual ou não entre os subgrupos de gênero que são pareados no estado ou traço medido (por exemplo, depressão). No caso em que o DIF está ausente nos 21 itens do BDI-II, uma comparação significativa da distribuição dos escores latentes entre populações heterogêneas pode ser aceita. Se o DIF estiver presente e for ignorado, as estimativas de prevalência ou gravidade podem ser enviesadas.

No presente estudo, o item "choro" (nº 10) exibiu DIF de gênero significativo. Além disso, esse item demonstrou DIF uniforme com base na mudança do Pseudo R^2 . Maior endosso condicional foi observado para as mulheres no item 10. A direção do mau funcionamento do item indicou que uma amostra com mais mulheres do que homens poderia superestimar a prevalência e a gravidade da depressão devido ao 10º item.

Episódios de choro podem ser vistos como uma resposta adaptativa ao estresse, mas não devem ser automaticamente interpretados como um sinal ou sintoma de depressão. No presente estudo, os episódios de choro foram relatados por quase um em cada quatro entrevistados. Esse comportamento moderadamente discriminou a gravidade da depressão e apresentou um limiar modesto a ser expresso pelos alunos. No entanto, não há base empírica vigorosa na relação entre o choro e a depressão (84,85). A maioria dos instrumentos de avaliação e sistemas de diagnóstico de transtornos de humor não são confiáveis em como lidam com o choro como um sintoma (86). Para uma maior precisão diagnóstica, é importante que os critérios e instrumentos utilizados para avaliar a depressão reflitam adequadamente os sintomas depressivos comuns nos gêneros feminino e masculino.

Mais observado entre as mulheres, o item “choro” como resposta emocional ligada ao gênero tem sido reivindicado na maioria dos estudos com escalas depressivas para demonstrar DIF (75,87–93). Embora um DIF significativo de gênero também tenha sido encontrado em nossa investigação para o “chorar”, o significado desse sintoma na avaliação da depressão ainda é uma questão controversa. O presente estudo pode dizer que os limiares para “chorar” parecem ser bem diferentes entre adultos do sexo feminino e masculino. Enquanto as alunas eram mais propensas a chorar do que os colegas do sexo masculino, a pontuação geral não teve impacto na avaliação da gravidade da depressão. Possivelmente, nossa grande amostra equilibrada em termos de gênero pode ter ocultado evidências substanciais dessa diferença. Para evitar negligenciar ou subestimar a ocorrência de depressão, os pesquisadores devem ter cuidado ao interpretar os resultados dos estudos com a predominância de participantes do sexo feminino e vice-versa.

Pesquisas demonstraram que as mulheres têm taxas mais altas de depressão (20,21) e relatam sintomas depressivos mais graves do que os homens (94,95) e é importante entender melhor como e por que essas diferenças de gênero na depressão ocorrem para propor intervenções mais eficazes. Vários estudos examinaram se existem diferenças de gênero no endosso dos sintomas depressivos, item por item (96–103). Alguns estudos descobriram que as mulheres são mais propensas a relatar certos itens somáticos, incluindo alterações do apetite, alterações de peso e fadiga do que os homens (97–101,103,104). No entanto, os sintomas somáticos específicos identificados como potencialmente diferentes para homens e mulheres não são os mesmos entre os estudos, e a maioria dos itens de sintomas somáticos analisados não apresenta diferenças. Por outro lado, existem evidências de que as mulheres são mais propensas que os homens a relatar certos itens cognitivos / afetivos, como choro,

sentimento de culpa, sentimento de inutilidade, desinteresse sexual e pensamentos de morte, do que os homens (96,100,103). Entretanto, assim como os sintomas somáticos, padrões consistentes não emergem entre os estudos e, para a maioria dos sintomas cognitivos / afetivos examinados, não há diferenças. Possíveis razões que foram sugeridas para explicar estes achados incluem o uso de diferentes métodos para avaliar os sintomas de depressão e a grande variação no perfil das pessoas avaliadas (101).

Outras explicações podem ser mais promissoras. Por exemplo, tem sido sugerido que as diferenças de gênero na depressão podem estar relacionadas a 1) ansiedade prévia, que é mais comum entre mulheres do que homens (105), 2) a disposição maior das mulheres em admitir sua depressão (106), 3) a tendência de as mulheres refletirem sobre problemas mais do que os homens (96), 4) níveis mais altos de estresse e níveis mais baixos de satisfação associados aos papéis tradicionais do gênero masculino versus feminino (26) e 5) alterações hormonais próprias do gênero feminino (24).

Diferenças na mensuração de variáveis, como gênero, podem explicar parcialmente a discrepância entre estudos prévios de prevalência utilizando escalas de sintomas. Aproveitando-se do recente avanço psicométrico da TRI, a detecção de DIF para uma variável específica não indica, inevitavelmente, que um determinado item está polarizando as estimativas relatadas. O viés de medição é tipicamente identificado após evidências empíricas consideráveis terem sido coletadas e revisadas por especialistas para determinar o impacto potencial do DIF (107). No presente estudo, foi sugerido que o BDI-II parece ter um desempenho similar para homens e mulheres. Portanto, as diferenças de gênero nas estimativas de prevalência de sintomas depressivos são menos importantes do que se supunha anteriormente.

Pesquisas futuras usando conjuntos de dados epidemiológicos representativos são necessárias para replicar as principais descobertas relatadas no estudo atual sobre os efeitos de gênero na psicopatologia depressiva.

Em suma, estudos que avaliam a influência de gênero nos sintomas depressivos sugerem que a prevalência de depressão pode variar entre os grupos de participantes (108,109). A avaliação dos sintomas depressivos pelo BDI-II pode levar a uma superestimação dos sintomas entre as mulheres. Como o BDI-II foi corrigido de acordo com os critérios de depressão do sistema DSM-IV, o construto subjacente abrangido por este instrumento engloba alegações subjetivas ligadas a gênero (por exemplo, se sente triste ou vazio) ou observações feitas por outros (por exemplo, choroso). Como consequência, homens com depressão podem não ter sido adequadamente identificados na prática clínica e permaneceram sem tratamento ou com tratamento inadequado. O atual manual diagnóstico DSM-5 reconhece fenótipos heterogêneos de sintomas depressivos em um modelo dimensional de gravidade. Portanto, a abordagem de resposta a itens que considera a probabilidade de resposta do BDI-II em uma abordagem centrada na pessoa pode apoiar melhor a definição atual de depressão maior.

LIMITAÇÕES

A força da análise atual está relacionada ao seu poder estatístico e à precisão das estimativas, ambas derivadas do grande tamanho da amostra e da distribuição equilibrada por gênero dos participantes não clínicos. No entanto, a interpretação dos resultados deve ser considerada à luz de algumas limitações. Primeiro, apesar do tamanho da amostra, havia várias opções de resposta que foram endossadas em baixa

frequência (ou seja, menos de 5%). Assim, algumas estimativas de parâmetros podem ser imprecisas.

Em segundo lugar, os participantes fizeram o auto relato dos sintomas depressivos através dos itens do BDI-II. Assim, o viés de informação devido ao método de coleta de dados pode alterar os resultados finais. Estudos de validade prévia mostraram que o escore de autoavaliação do BDI-II foi correlacionado com a avaliação da gravidade em 65,4% dos casos de episódio depressivo maior do DSM-IV clinicamente diagnosticado (40).

Terceiro, os dados transversais deste estudo não permitiriam inferências adicionais sobre os fatores que afetam os diferentes estágios da gravidade da depressão, abrangendo o início, o estado ativo e o período de remissão da depressão. Um estudo com o design longitudinal pode superar essa deficiência.

A última preocupação está relacionada à generalização dos resultados. Embora tenhamos recrutado uma amostra representativa de estudantes universitários brasileiros (a maioria com idade próxima aos 20 anos), a influência do nível educacional ou regional nas propriedades psicométricas do BDI-II não foi avaliada no presente estudo. Portanto, os principais resultados deste estudo sobre o uso do BDI-II não devem ser estendidos à população geral de diferentes regiões, níveis de escolaridade e faixa etária. Estudos posteriores devem ser realizados utilizando uma amostra mais inclusiva, com participantes de faixas etárias distribuídas uniformemente, níveis educacionais diversos ou ainda recrutados em contextos clínicos.

8. CONCLUSÃO

A preponderância da depressão entre mulheres em estudos epidemiológicos tem motivado discussões para um maior entendimento das diferenças de gênero em sintomas depressivos. Algumas linhas de pesquisa sugerem que mulheres são mais propensas do que homens a relatar alterações de sono, peso ou apetite e dificuldades psicomotoras (31,81,99,100,106). Contudo, os estudos que abordam este tema não encontraram padrões consistentes de diferenças de gênero e ainda não permanece claro se um grupo tem maior probabilidade de relatar um determinado sintoma simplesmente pelo fato de estarem mais propensos a ter depressão (103).

Os principais resultados da presente investigação tentaram esclarecer, em uma amostra representativa de estudantes universitários no Brasil, a adequação dos itens do BDI-II como uma medida de avaliação da gravidade da depressão no gênero masculino e feminino, levando em conta as marcantes diferenças epidemiológicas em termos de gênero. A detecção correta de DIF nos sintomas depressivos, comparados por gênero, em amostras de estudantes pode ter implicações preventivas para a comunidade, porque elas possivelmente estão passando pelo período de pico do início da depressão (110,111). Mostramos, por meio da análise de respostas aos itens, que a variável gênero exerce influência na manifestação da depressão.

Embora o item “choro” tenha apresentado DIF relacionado ao gênero na escala de autor relato, seu impacto na escala foi insignificante. Quando o BDI-II é aplicado a uma amostra da população com características demográficas distorcidas ou tamanho de amostra insuficiente, os resultados podem ser alterados da mesma forma. Portanto, é importante que os critérios e instrumentos utilizados para avaliar a depressão possam

refletir adequadamente os sintomas comuns e as experiências de depressão relacionadas a gênero. Os resultados da presente investigação fornecem suporte adicional para a validade do uso do BDI-II ao avaliar a depressão em contextos acadêmicos.

9. ANEXOS

Anexo A

Tabela suplementar: Proporção da resposta do BDI-II (%) para as alternativas de categorias de escolha nos 21 itens

Item	Amostra total N=12.677	Homens n = 5.692	Mulheres n = 6.985
1. Tristeza			
Categoria 1	81.3	85.7	77.6
Categoria 2	17.5	13.2	20.9
Categoria 3	0.7	0.5	0.8
Categoria 4	0.6	0.5	0.6
2- Pessimismo			
Categoria 1	81.0	82.6	79.7
Categoria 2	16.9	15.4	18.2
Categoria 3	1.6	1.6	1.7
Categoria 4	0.5	0.5	0.5
3- Sentimento de fracasso			
Categoria 1	84.3	84.2	84.3
Categoria 2	9.4	10.3	8.7
Categoria 3	5.8	5.0	6.4
Categoria 4	0.5	0.5	0.6
4- Perda de prazer			
Categoria 1	77.2	80.0	74.9
Categoria 2	19.8	17.5	21.7
Categoria 3	2.4	1.9	2.8
Categoria 4	0.5	0.5	0.6
5. Sentimento de culpa			
Categoria 1	69.9	72.4	67.9
Categoria 2	28.1	25.9	30.0
Categoria 3	1.4	1.3	1.5
Categoria 4	0.6	0.5	0.6
6- Sentimentos de punição			
Categoria 1	85.3	85.2	85.4
Categoria 2	10.2	10.9	9.6
Categoria 3	1.3	1.1	1.5
Categoria 4	3.2	2.9	3.4
7- Autoestima			
Categoria 1	85.5	87.8	83.7
Categoria 2	6.2	6.0	6.4
Categoria 3	6.9	5.4	8.2
Categoria 4	1.3	0.8	1.7
8- Autocrítica			
Categoria 1	56.0	56.5	55.5
Categoria 2	33.1	33.4	32.9
Categoria 3	9.4	9.2	9.5

Categoria 4	1.5	1.0	2.0
9- Ideias suicidas			
Categoria 1	93.2	93.6	92.8
Categoria 2	6.1	5.7	6.5
Categoria 3	0.4	0.4	0.3
Categoria 4	0.3	0.3	0.4
10- Choro			
Categoria 1	77.6	88.4	68.9
Categoria 2	12.1	6.4	16.7
Categoria 3	6.5	0.8	11.2
Categoria 4	3.8	4.5	3.1
11- Agitação			
Categoria 1	71.0	74.7	67.9
Categoria 2	25.1	21.9	27.8
Categoria 3	2.2	2.0	2.4
Categoria 4	1.7	1.5	1.9
12- Perda de interesse			
Categoria 1	77.0	79.6	74.9
Categoria 2	20.4	18.2	22.3
Categoria 3	1.8	1.6	2.0
Categoria 4	0.7	0.7	0.8
13- Indecisão			
Categoria 1	71.6	76.7	67.5
Categoria 2	20.7	18.6	22.4
Categoria 3	2.3	1.8	2.8
Categoria 4	5.4	2.9	7.4
14- Desvalorização			
Categoria 1	88.1	90.4	86.2
Categoria 2	7.0	6.5	7.4
Categoria 3	4.3	2.7	5.6
Categoria 4	0.6	0.5	0.7
15- Perda de energia			
Categoria 1	61.6	66.3	57.7
Categoria 2	33.2	30.1	35.8
Categoria 3	4.5	3.2	5.7
Categoria 4	0.7	0.5	0.9
16- Alterações no sono			
Categoria 1	48.5	51.6	45.9
Categoria 2	38.9	36.7	40.7
Categoria 3	10.4	9.9	10.9
Categoria 4	2.2	1.8	2.5
17- Irritabilidade			
Categoria 1	68.7	73.0	65.1
Categoria 2	26.6	23.6	29.1
Categoria 3	3.6	2.7	4.3

Categoria 4	1.2	0.7	1.5
18- Mudanças no apetite			
Categoria 1	61.6	69.1	55.4
Categoria 2	31.3	26.2	35.4
Categoria 3	4.9	3.4	6.2
Categoria 4	2.2	1.3	2.9
19- Dificuldade de concentração			
Categoria 1	65.8	70.0	62.4
Categoria 2	23.4	21.8	24.7
Categoria 3	9.3	7.0	11.2
Categoria 4	1.4	1.2	1.6
20- Cansaço ou fadiga			
Categoria 1	59.1	63.6	55.4
Categoria 2	33.2	31.0	34.9
Categoria 3	5.6	4.0	6.9
Categoria 4	2.1	1.4	2.8
21. Perda da libido			
Categoria 1	88.0	91.9	84.8
Categoria 2	9.4	6.9	11.4
Categoria 3	2.0	0.9	2.9
Categoria 4	0.6	0.3	0.9

Nota: Em geral, a categoria 1 indica "como de costume"; categoria 2 "mais do que o habitual"; categoria 3 "o tempo todo"; e categoria 4 "sofrimento extremo". Embora a redação do item possa variar de acordo com o sintoma depressivo, as categorias são exibidas em nível crescente de gravidade.

Anexo B

Anexo C

Anexo D**Artigo 1**

**Response pattern of depressive symptoms among college
students: What lies behind items of the Beck Depression
Inventory-II?**

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Response pattern of depressive symptoms among college students: What lies behind items of the Beck Depression Inventory-II?

Antonio Reis de Sá Junior^{a,b}, Arthur Guerra de Andrade^{b,c}, Laura Helena Andrade^b,
Clarice Gorenstein^{b,d}, Yuan-Pang Wang^{b,*}

^a Department of Health Sciences, Federal University of Santa Catarina, Araranguá, SC, Brazil

^b Institute & Department of Psychiatry (LIM-23), University of São Paulo Medical School, São Paulo, SP, Brazil

^c Department of Neuroscience, Medical School, ABC Foundation, Santo André, SP, Brazil

^d Department of Pharmacology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil

ARTICLE INFO

Keywords:
Depressive symptoms
Beck Depression Inventory-II
Classical test theory
Item response theory
College students

ABSTRACT

Background: This study examines the response pattern of depressive symptoms in a nationwide student sample, through item analyses of a rating scale by both classical test theory (CTT) and item response theory (IRT).
Method: The 21-item Beck Depression Inventory-II (BDI-II) was administered to 12,711 college students. First, the psychometric properties of the scale were described. Thereafter, the endorsement probability of depressive symptom in each scale item was analyzed through CTT and IRT. Graphical plots depicted the endorsement probability of scale items and intensity of depression. Three items of different difficulty level were compared through CTT and IRT approach.

Results: Four in five students reported the presence of depressive symptoms. The BDI-II items presented good reliability and were distributed along the symptomatic continuum of depression. Similarly, in both CTT and IRT approaches, the item 'changes in sleep' was easily endorsed, 'loss of interest' moderately and 'suicidal thoughts' hardly. Graphical representation of BDI-II of both methods showed much equivalence in terms of item discrimination and item difficulty. The item characteristic curve of the IRT method provided informative evaluation of item performance.

Limitation: The inventory was applied only in college students.

Conclusion: Depressive symptoms were frequent psychopathological manifestations among college students. The performance of the BDI-II items indicated convergent results from both methods of analysis. While the CTT was easy to understand and to apply, the IRT was more complex to understand and to implement. Comprehensive assessment of the functioning of each BDI-II item might be helpful in efficient detection of depressive conditions in college students.

1. Introduction

The second version of Beck Depression Inventory (BDI-II) is, worldwide, one of the most influential measures for assessing the presence and severity of depressive symptoms (Beck et al., 1996). Validation investigations and normative data of the BDI-II in large student samples have been widely published in relevant literature (Amarson et al., 2008; Storch et al., 2004; Whisman et al., 2013; Whisman and Richardson, 2015).

Traditionally, the validity of BDI for measuring the severity of depression has been determined by classical test theory (CTT), where individual summed scores are compared with cut-off thresholds in a

sample (Crocker and Algina, 2008). Because self-rated depressive symptoms may be misrepresented in specific health conditions (Lerdal et al., 2014), the scaling process of the BDI-II is highly dependent on the sample characteristics (Hambleton and Jones, 1993). For example, fatigue symptoms may be over-reported in medical settings. Conversely, the meaning of fatigue in competitive academic life needs to be better described, because university is a potential environment for the prevention and treatment of depression (Cuijpers et al., 2016; Reavley et al., 2010). Many young adults enrolled in higher education are passing through the peak age period for depression onset (Ibrahim et al., 2013; Zivin et al., 2009). Hence, investigations into the capacity of the BDI-II to detect depressive symptoms in an analogous student

* Correspondence to: Núcleo de Epidemiologia Psiquiátrica (LIM-23), Departamento & Instituto de Psiquiatria, Faculdade de Medicina, Universidade de São Paulo, Rua Dr. Ovídio Pires de Campos, 785, CEP 05403-010 São Paulo, SP, Brazil.
E-mail address: gsap.jebao@hotmail.com (Y.-P. Wang).

<https://doi.org/10.1016/j.jad.2018.02.064>

Received 19 May 2017; Received in revised form 1 February 2018; Accepted 25 February 2018
Available online 03 March 2018

0165-0327/ © 2018 Elsevier B.V. All rights reserved.

sample may strengthen its utility as a screening tool for depression in this population.

Interpretation of the severity level of depression using the item counting practice of the CIT framework may be misleading, as all the items that comprise the scale are assumed to contribute equally to the same underlying construct. The total score of the BDI-II may be artificially inflated due to symptoms of underlying illness and type of environment, rather than of depression itself. For example, whilst common depressive states distinguish heterogeneous illnesses, serious suicidal thoughts are rarely endorsed in milder depressive states. This suggests that item-level is easily reported in all levels of depression, but only severely depressed patients will report suicidal ideation. Thus, the endorsement probability of an item is reliant on the health status, as expressed by the difficulty level of an item on the scale. Because adequate scale items must be able to discriminate all levels of construct, i.e., from absence of illness to the most severe pathology (Tavres et al., 2007), the capacity of individual BDI-II items in forecasting cases of depression is under-researched. Furthermore, since the findings are both sample and scale dependent, there is a serious logical obstacle to the measurement performance for which the instrument is designed (Hamilton and James, 1993).

In past decades, item response theory (IRT) has been increasingly used to overcome the shortcomings and inconsistencies of the CIT approach. The IRT-based model aims to calculate the probability of getting an expected response to an item along a continuum of the latent trait (e.g., level of ability or depression severity). Using an item-wise approach, this information is expressed along a latent dimension θ (Greek theta), much like a standardized score. Independent of the questions used, the estimated latent trait represents a close measure of the true construct free of errors (Hamilton and James, 1993). Therefore, the estimated parameters are sample invariant.

Unlike conventional CIT, there is less ambiguity about the meaning of a response in IRT parameters – non-depressed people do not answer certain items and depressed people do. Nevertheless, IRT parameters approximately resemble their traditional CIT counterparts in some instances. For example, items with higher item-total correlations correspond to higher discriminations (a_i), and items with lower endorsement rates have more positive location parameters (b_i) in the latent dimension. Estimates of the IRT parameters can reveal which items show higher discrimination values to distinguish between divergent latent variables, for example, depression. On other hand, the item difficulty of a rating scale is distributed in θ to reflect the trait of severity in relation to the absence or presence of the pathological condition. A positive value of θ indicates that a person is more depressed than average, and a negative value of θ means less depressed than average. These basic aspects of the IRT parameters can be visually shown in an S-shaped curve known as the item characteristic curve (ICC). We refer readers to comprehensive discussions of its basic properties in relevant publications (Coppolani et al., 2014; Emberton and Raine, 2000; Rabe and Rodriguez, 2016; Rostner, 2010; Thomas, 2011).

Although IRT models have been applied in high-quality measurement instruments such as the BDI (Browne et al., 2013; Lendel et al., 2014), comprehensive comparisons between the item analysis through both the CIT and IRT approaches have been relatively neglected in health literature, in terms of discriminant capacity and item difficulty. While the CIT model is most adopted in the early stage of the content validation of a measure, the IRT model requires a larger sample size to refine item functioning (Coppolani et al., 2014). Furthermore, the standard error (SE) of item parameters in IRT models could estimate the latent trait (e.g., the severity of depressive symptoms), as well as the discrimination parameters. Straightforward psychometric comparisons between these two apparently opposing psychometric methods in depression scales is virtually nonexistent or misleading (Naplan and Macmillan, 2018).

The aims of this study were: a) to investigate the psychometric properties of the BDI-II for detection of depressive symptoms in a large

sample of college students, and b) to analyze the item functioning of the BDI-II through CIT and IRT approaches. Hypotheses and misinterpretations of the results on the applicability of the BDI-II in college students are discussed.

2. Materials and methods

2.1. Ethics statement

The Ethics Committee of the University of São Paulo Medical School approved the present study. All participants signed an informed consent prior to participating in the data collection.

2.2. Sampling frame

The target population was college students enrolled in a nationwide survey on the use of alcohol, tobacco and other drugs among college students in the 27 Brazilian state capitals (Andrade et al., 2010). The participants were recruited through two-stage cluster sampling. The first stage involved higher education institutions (HEIs) that deliver undergraduate courses in each of the state capitals, in accordance with the school list provided by the Brazilian Ministry of Education. HEIs located in small cities were excluded from the sampling plan. Eligible HEIs were randomly selected from the list of ordered institutions, by means of the probability proportional to size (PPS) of the number of enrolled students in each HEI. The stratification was based on two variables: the location of the capital and the type of institution funding (either private or public), totaling 54 levels. Of the selected 114 colleges and universities, 100 HEIs (87.7%) agreed to participate after institutional invitation.

In the second stage, the class of students was considered the primary sampling unit (PSU) within each selected HEI, for the sake of feasible follow-up. Each element of class was defined as a group of students attending a given course. The number of classes was proportional to the total number of students in the institution, totaling 484 classes or 70.6% of the selected HEI. All students in the selected classes were invited to volunteer for the study. The participant response rate was 95.4% for the college students who were in class at the time of the survey (Andrade et al., 2010).

2.3. Participants

The initial sample of the survey was 12,721 college students. Of these, 10 respondents were excluded because they claimed to use “relevo”, a dummy drug. “Relevo” has been used in some drug surveys to detect over-reporting of drug use (e.g., Miller and Plant, 1994; Plant and Miller, 2000). College students may inflate the rates of substance use by endorsing fake or fictitious drug “relevo” use. The answers of these participants were considered untrustworthy and their questionnaires were discarded. Hence, a final sample of 12,711 valid questionnaires was considered for analysis. All students completed a structured, anonymous, self-administered instrument consisting of 96 closed questions on use of substances, wherein the 21-item BDI-II was included. The demographic characteristics of the sample were 7002 women (58.1%) and 5709 men (44.9%) and the mean age was 23.9 years ($SD = 6.9$). The mean age of women was similar to men ($M = 24.8$ vs. 23.8 , $p = 0.18$). The age and total score showed no evidence of correlation and the mean total score of the BDI-II showed no difference (women $M = 7.86$ vs. men $M = 5.92$; $p < 0.001$).

2.4. The Beck Depression Inventory-II (BDI-II)

The BDI-II is a 21-item self-report questionnaire for assessing the severity of depressive symptoms; the items comprise four ordinal categories (from 0 through 3). Absence (or “no”) of depressive symptom in each item is scored as “0” and presence of symptom is

scored between 1 and 2, with the higher score indicating more intense symptom severity. The possible total score ranges from 0 to 63 (Bark et al., 1961).

The timeframe of the BDI-II covers the 2 weeks required to meet the DSM-IV criteria for major depressive disorder (American Psychiatric Association, 1994). Substantial evidence of reliability has been reported in different language versions of the BDI-II (Wang et al., 2013), including Brazilian-Portuguese (Gomes-Oliveira et al., 2012). According to the classic test-retest method, the percentages of depression levels were as follows: minimal 0–13 = 40.5%; mild 14–19 = 13.4%; moderate 20–28 = 5.2% and severe 29–63 = 3.7%. Furthermore, the classic scoring method showed a high proportion of respondents scoring zero and 20.1% of respondents have scored above the 13/14 threshold of mild depression (Bark and Stanley, 2006).

After dichotomizing the BDI-II items as absent or present (score 0 vs. 1–6), we were able to detect 79.9% of asymptomatic students reporting any depressive complaints in past 15 days. Worth mentioning, Aaron Beck and colleagues (1996) also have dichotomized the scale items as 'zero vs. non-zero' scores in their manual, purposing to show the endorsement rate of depressive symptomatology in both clinical and student samples. Possible total scores ranged from 0 to 21 for this new binary scoring method.

3.1. Statistical analyses

Both CIT and IRT methods were applied to inspect the BDI-II data. First, the characteristics of BDI-II data dispersion and homogeneity were explored for each scale item and total score: mean (M), standard deviation (SD), Pearson's coefficient of variation (CV), and Cronbach's alpha (α) coefficient of internal consistency.

Next, under the CIT approach, two types of item analyses calculated the biserial point correlation (r_{pb}) and the proportion of endorsement rate (%). r_{pb} was used as an indicator of the discrimination ability or item selectivity (Ullas and Hopkins, 1999) and %, the proxy of item difficulty or severity (Cudeck and Algina, 2000). The correlation r_{pb} refers to the relationship of dependency between the 'observed' values of each BDI-II item with its 'expected' values. If the observed value was much higher than the expected value, this may indicate dependency in the data. Alternatively, if the observed value was much lower than expected, this may indicate "unrelated" noise. According to Kim (2006) and Kucukarslan (2006), r_{pb} correlations between 0.2 and 0.4 might be considered acceptable. Furthermore, an observed value between 0 and 0.14 indicates a weak, between 0.15 and 0.25 moderate, between 0.26 and 0.38 good, and > 0.39 very good correlation.

On the other hand, the % index varies between 0 and 1.0: where an easy item shows a value close to 1.0, moderately difficult near 0.5, and high difficulty near 0 (Kucukarslan et al., 2000; Kim and Rastah, 2006). Extreme values, such as very difficult and/or very easy items, are not useful for identification of severe and mild depressive psychopathology.

For CIT-derived graphical interpretation, the total BDI-II score (y-axis) in the sample was stratified through z-score transformation in five interval groups of percentile: 0–20, 20–40, 40–60, 60–80, and 80–100. The endorsement rate shows the presence of related depressive symptoms in relation to their probability of endorsement (y-axis). I.e., the high mean (based on the binary scoring method) BDI-II items would be the less endorsed ones, and vice-versa. Analyses were performed using SPSS 20.0 (www.spss.com) and the graphics were produced in Excel spreadsheet. The level of significance was set at $p < 0.05$ for two-tailed tests.

For the IRT latent trait approach, the two-parameter logistic (2-PL) model was chosen to calculate item discrimination (a_i) and item difficulty (b_i). The magnitude of the a_i parameter indicates how well a scale item (i) can distinguish the construct under assessment. In health sciences, the b_i score of a given scale item (i) is located on the right side of the latent variable θ to denote more severe manifestation of the target

Table 1
Psychometric characteristics of the Beck Depression Inventory-II for college students ($N = 12,711$).

Item	M	SD	CV	α
1. Sadness	0.19	0.39	2.08	0.90
2. Frustration	0.19	0.39	2.05	0.90
3. Feet Heavy	0.16	0.36	2.25	0.90
4. Loss of pleasure	0.23	0.42	1.83	0.90
5. Guilty feelings	0.30	0.45	1.53	0.90
6. Frustration feelings	0.15	0.35	2.33	0.90
7. Self-blame	0.14	0.33	2.39	0.90
8. Self-criticism	0.14	0.33	2.39	0.90
9. Irrational thoughts	0.07	0.25	3.67	0.90
10. Crying	0.22	0.41	1.81	0.90
11. Anger	0.29	0.45	1.55	0.90
12. Loss of interest	0.23	0.42	1.83	0.90
13. Indifference	0.20	0.40	2.00	0.90
14. Worthlessness	0.13	0.33	2.57	0.90
15. Loss of energy	0.20	0.40	2.00	0.90
16. Changes in sleep	0.11	0.31	2.82	0.90
17. Irritability	0.21	0.40	1.90	0.90
18. Changes in appetite	0.20	0.40	2.00	0.90
19. Concentration difficulty	0.24	0.42	1.75	0.90
20. Tiredness or fatigue	0.41	0.49	1.20	0.90
21. Loss of interest in sex	0.13	0.33	2.57	0.90
Total	0.46	0.33	0.74	0.90

M: Mean.

SD: Standard deviation.

CV: Pearson's coefficient of variation = SD/M.

α : Cronbach's alpha coefficient of internal consistency if the item is deleted.

The mean scores are based on the dichotomous scoring method.

Since these item parameters were displayed graphically as item characteristic curves (ICC) in the latent dimension θ , in an ICC, a scale item (i) with a steep slope provides more information or discrimination ability (a_i) than a flat slope for the detection of the underlying construct of BDI-II. According to Belive (2003), an a_i value > 1.0 is considered highly discriminant. Assumptions of unidimensionality and local independence for the latent trait model were checked, respectively, through values of eigenvalues of factorial model and between-item correlations controlled for a single factor. The software Mplus 5.0 (www.statmodel.com) calculated IRT parameters and generated the ICC graphical plots.

3.2. Results

Table 1 shows the psychometric analysis obtained from 12,711 respondents for BDI-II, the mean total score was 0.46 (SD = 0.33, CV = 0.74). The great majority of students ($n = 10,152$; 79.9%) have endorsed at least one depressive symptom in this rating scale. With the exception of "changes in sleep" (#16, $M = 0.11$), all BDI-II items presented a $M < 0.5$. Items such as "tiredness or fatigue" (#20) "self-criticism" (#8) also presented a high mean score in this student sample (M around 0.4). Intermediate mean scores were items such as "loss of pleasure" (#4), "crying" (#10), "anger" (#11), "loss of interest" (#12) and "indifference" (#13); the M ranged from 0.2 to 0.3. The lowest mean scores were observed for "irrational thoughts" (#9), "loss of interest in sex" (#21) and "worthlessness" (#14); the M was around 0.1.

The lowest dispersion items were "changes in sleep" (#16), "self-criticism" (#8) and "tiredness or fatigue" (#20); CV < 1.0. Intermediate figures of dispersion were items such as "loss of interest" (#12), "crying" (#10), "loss of pleasure" (#4), and "guilty feelings" (#5); CV around 2.0 was observed. The highest dispersion items were "irrational thoughts" (#9), "loss of interest in sex" (#21), and "worthlessness" (#14), with CV > 2.5, indicating double response bias of score precision. For the total scale, the Cronbach's alpha coefficient was 0.90, revealing an adequate internal consistency. Item-total correlation

Table 2

Estimation of item difficulty and item discrimination of BDI-II between classical test theory and item response theory models.

Item	Classical test theory		Item response theory	
	η_{ijk}	η_{ik}	a_i (SD)	b_i (SD)
1. Sadness	0.61	19	2.39 (0.86)	1.18 (0.63)
2. Frustration	0.59	14	2.34 (0.86)	1.11 (0.63)
3. Fastidious	0.54	18	2.05 (0.86)	1.39 (0.63)
4. Loss of pleasure	0.52	20	2.13 (0.86)	0.95 (0.63)
5. Tiredly thoughts	0.57	30	1.63 (0.84)	0.79 (0.63)
6. Frustration thoughts	0.50	18	1.69 (0.85)	1.30 (0.63)
7. Self-criticism	0.43	14	2.12 (0.85)	1.38 (0.63)
8. Self-criticism	0.50	14	1.61 (0.84)	0.35 (0.63)
9. Irrational thoughts	0.42	37	1.68 (0.87)	1.06 (0.63)
10. Crying	0.53	33	1.64 (0.84)	1.18 (0.63)
11. Agitation	0.37	26	1.65 (0.84)	0.41 (0.63)
12. Loss of interest	0.46	38	2.71 (0.87)	0.29 (0.63)
13. Indifference	0.41	30	2.81 (0.88)	0.75 (0.63)
14. Worthlessness	0.41	13	2.44 (0.84)	1.38 (0.63)
15. Loss of energy	0.46	30	2.61 (0.86)	0.40 (0.63)
16. Changes in sleep	0.57	31	1.85 (0.84)	– 0.38 (0.63)
17. Irritability	0.43	31	2.05 (0.86)	0.66 (0.63)
18. Changes in appetite	0.33	38	1.38 (0.80)	0.49 (0.63)
19. Concentration difficulty	0.45	34	2.31 (0.86)	0.66 (0.63)
20. Thoughts of suicide	0.45	41	2.61 (0.87)	0.33 (0.63)
21. Loss of interest in sex	0.59	13	1.35 (0.84)	1.30 (0.63)

η_{ijk} internal point correlation.

η_{ik} difficulty index.

a_i discrimination parameter of the item i .

b_i severity (difficulty) parameter of the item i .

SD: standard error.

demonstrated that deleting any item would slightly affect the internal consistency of the construct, with Cronbach's alpha ranging from 0.89 to 0.90.

Item indicators of the BDI-II are compared in Table 2, from the perspective of both CTT and IRT. The CTT discrimination (r_{ijk}) and difficulty index (η_{ik}) were estimated one-by-one, in parallel columns, using the IRT discrimination (a_i) and difficulty parameter (b_i).

The mean discrimination index for total scores of the CTT approach was $r_{ijk} = 0.55$ and the mean difficulty index was $\eta_{ik} = 26$ (not shown), suggesting that despite overall good discrimination of the scale items of depressive states, the BDI-II items were rarely endorsed.

All items presented high correlation ($r_{ijk} > 0.35$, range 0.33–0.64), indicating good discrimination of the items of the underlying construct of depression. The item “loss of interest in sex” showed the lowest discrimination ($r_{ijk} = 0.30$), and “loss of interest” and “loss of energy” presented the highest discrimination ($r_{ijk} = 0.65$). Notably, the following items also presented high discrimination capacity for underlying depression ($r_{ijk} > 0.60$): “sadness”, “worthlessness”, “loss of pleasure”, “self-criticism”, “irritability”, “indifference”, “concentration difficulty”, and “tiredness or fatigue”.

Respondents could not easily endorse most BDI-II items. The item “changes in sleep” (#16) was the most endorsed or the easiest depressive symptom to be reported by students ($\eta_{ik} = 51$). Next, 12 items presented moderate endorsement difficulty (η_{ik} between 20 and 30) and 8 items were highly difficult to endorse ($\eta_{ik} < 20$). The most difficult item was “suicidal thoughts” (#20), $\eta_{ik} = 7$.

In order to further compare the CTT and IRT models, three symptoms of increasing levels of difficulty were chosen to illustrate the results obtained from both methods. The easiest item “changes in sleep” (#16), the moderately difficult item “loss of interest” (#12), and the most difficult item “suicidal thoughts” (#20) are graphically displayed in Fig. 1. The comparative response pattern of each item can be inspected through CTT item response slopes, with the y-axis indicating the probability of endorsement and x-axis the severity level or the range of the total BDI-II score. All items start off with low severity and the slope moves up over the next stratified level of severity.

Fig. 1. Proportion of endorsement of item exhibits thought, loss of interest and changes in sleep, by classical test theory approach.

The majority of respondents (around 90%) have endorsed #16 in the 20–100 percentile range of the total score, i.e., in most levels of symptomatic intensity for depression. Around 50% of respondents endorsed #12 in the 20–40 range (or mild symptomatic severity) and 50% in the 40–60 range (or moderate symptomatic severity) of the total score. Finally, only 20% endorsed #20 in the 20–40 range, 60% in the 40–60 range, and 78% in the 60–80 range, and 90% in 80–100 range of severity. In other words, while #9 is a good discriminator across the entire continuum of the attribute (depression), #12 discriminates better at the lower than the upper end of the attribute. Summarizing, #16 was easily endorsed in all levels of severity, #12 in the moderate level, and #20 in the severe level of depression.

Regarding the 2-PL IRT model to assess BDI-II items, the results showed that the item difficulty parameter ranged between $b_i = -0.53$ and 1.38 along the latent dimension θ of depression severity (Table 2). In this student sample, “changes in sleep” (#16), “self-criticism” (#8), “tiredness or fatigue” (#20), “loss of energy” (#15), “changes in appetite” (#18) were easily endorsed symptoms or of low severity, all with $b_i < 0.5$ of depression construct θ . The items indicating a higher level of severity of depression (difficult endorsement) were: “sadness”

Fig. 2. Item characteristic curves for item suicidal thoughts, loss of interest and changes in sleep, by item response theory approach.

(#1), "pessimism" (#2), "crying" (#10), "self-harm" (#4), "worthlessness" (#14), "guilt feelings" (#3), "punishment feelings" (#6), "loss of interest in sex" (#21), and "suicidal thoughts" (#9), all with a $b_1 > 1.0$. The remaining items, such as "concentration difficulty", "irritability", "indecisiveness", "guilty feeling", "agitation", "loss of interest", were placed between these two extremes of endorsement, with b_1 ranging between 0.5 and 1.0.

Concerning the relationship between the difficulty parameters of CTT and IRT, or b_1 , the correlation was statistically significant ($r = -0.96$, $p < 0.001$), being marked and substantial. The Pearson's confidence interval of 95% based on 1000 bootstrap samples fell between -0.92 and -0.99 .

Item discrimination parameters ranged between $a_1 = 1.96$ – 3.44 . The top five highly discriminating items included "worthlessness" ($a_1 = 3.44$), "self-harm" ($a_1 = 3.18$), "loss of interest" ($a_1 = 3.71$), "sadness" ($a_1 = 2.38$), and "loss of pleasure" ($a_1 = 2.32$).

The graphical plot of the 2-PL IRT model provides the slopes of both item discrimination (a_1) and item difficulty (b_1) in IRT for the three selected depressive symptoms #16, #12 and #9 (Fig. 2). The S-shaped curves show varying degrees of monotonicity and represent the probability of endorsing in a certain category or higher, as well as how the probability changes as a function of the latent trait level. Accordingly, the easiest #16 was placed on the left side of 0, #12 around the zero, and the most difficult #9 the right side of 0. Regarding the relationship between discrimination parameters by CTT and IRT, or r_{a_1} , and a_1 , the correlation also was positive and significant ($r = 0.89$; $p < 0.006$). Notably, the narrow SE of parameter estimates produced from our large sample (Table 3) indicates that the calculated scores were quite accurate for the IRT model.

4. Discussion

Taking advantage of a nationwide anonymous sample of 12,711 college students, the present study has investigated the response pattern of depressive symptoms by means of the EDI-II. The great majority of students in the current study reported some depressive symptoms, thus its rapid detection through web-friendly scale is justifiable from the perspective of college health. Item-wise performance of the EDI-II was evaluated by means of both the CTT and the IRT framework, because the symptomatic presentation of depressive disorders in different populations may be heterogeneous. We have examined agreements and contradictions on the use of the EDI-II regarding CTT-IRT equivalences, in terms of item ability to discriminate and depict severity of depression. To the best of our knowledge, this type of comparison has never been reported in the literature for the EDI-II. The results of the EDI-II data with binary scored items and no guessing have suggested ways to deal with misconceptions about the alleged deficiencies of CTT relative to IRT (Bayliss and Maroulidou, 2011). Both methods have provided

comprehensive results for the interpretation of depressive symptoms among college students.

Since the EDI-II was not originally designed for diagnosing depressive disorders in non-clinical samples (Bauer and Slutske, 2006), its utility in identifying the presence of depressive symptoms in students is of utmost importance (Hendell et al., 1987). First of all, depressive symptoms in these college students could be easily detected in a high proportion of participants (around 80%). Second, the self-reporting format of the EDI-II was well accepted by students, with low participation refusal (4.4%) and negligible invalid answers (0.0008%). Third, the internal consistency of the EDI-II was good, being reliably homogeneous for all depression-related items in the scale. Therefore, the universal use of this self-reporting inventory in a higher education environment seems feasible for implementation, aimed at screening and monitoring the students' mental health.

Remarkably, the total score of the present student sample was well below than that reported for psychiatric patients in the original item analysis of the tool (Bauer and Slutske, 2006). This skewed score distribution suggested that most items showed a considerable floor effect or little representative convenience for use in non-clinical samples (Huang and Wang, 2006), hampering the score interpretation (Thomado and Thomado-Ciclet, 2010). After rewording the EDI-II items as no/yes or non-symptomatic/symptomatic, the endorsement rate of participants greatly improved to almost 80%. Individual item scores provided varied endorsement rates. The importance of comprehensive detection of depression in student samples is based on the following: (1) an average, mild symptoms of variable clinical significance were reported by one in four students; (2) the early detection and prevention of depression in students might have an evident impact on their academic performance and drop-out rate; (3) while somatic-negative symptoms were frequently reported by students, reports of suicidal ideation were rare. Consequently, investigations into the capacity of a user-friendly, self-report device such as EDI-II to detect depressive symptoms among students may strengthen its utility as a screening and monitoring tool for mental health in academic settings.

The observed endorsement rate (76%) and the latent trait (difficulty parameter b_1) are considered as indicators for intensity of depressive symptoms, respectively for CTT and IRT model. Figs. 1 and 2 show graphical representation of the comparison of the two severity concepts. Combined assessment of the instrument performance in both the CTT and IRT models (Bayliss and Maroulidou, 2011) suggests that the probability of endorsement (76%) corresponds well to the severity parameter (b_1) of the latent trait under investigation. When the results of item correlation (r_{a_1}) and discrimination parameters (a_1) were compared through CTT and IRT, acceptable equivalences also emerged for this sample (Bayliss and Maroulidou, 2011). Since our purpose was to obtain precise measurements of item characteristics with stable items, using a large sample also made more sense for fitting the IRT

than the CIT model.

Somatic symptoms such as changes in sleep, changes in appetite, tiredness, and loss of energy were the most common complaints. Sizable cognitive manifestations of concentration difficulty and irritability were also frequently reported. Self-evaluation stood out as the most common negative symptom hindering students' achievements. However, these complaints have poorly discriminated the severity of depression. Regardless of their high frequency, these somatic-cognitive symptoms in students might be of a transient nature and hold little hint of depressive states. In a college population, somatic symptoms such as inability to sleep, irritability, tiredness, lack of energy, and appetite or weight change need be examined in stressful academic life (Carrington et al., 2014; Kessler et al., 2013). Probably, an independent complaint of each single symptom might not indicate the need for immediate treatment.

Sadness and loss of pleasure were endorsed by around one in five students, whereas their ability to discriminate different levels of depressive severity was acceptable. These symptoms are core criteria for the diagnosis of DSM-5 major depressive disorders and may suggest the need for counseling or therapeutic intervention (American Psychiatric Association, 2013).

Heavily endorsed symptoms of suicidal thought, loss of interest in sex, and worthlessness were disclosed by around one in ten students. These extreme complaints necessitate immediate attention from families, peers, friends and academic tutors, and must be examined in detail in conjunction with other depressive manifestations. They constitute a red flag for severe states of academic vulnerability, environmental difficulty or severe depressive conditions. Although having suicidal thoughts was the least reported item among college students, substantial correlation still demonstrated its contribution to the variance structure of the scale (Dancer and Stanley, 2006; Berk et al., 1979; Wang et al., 2013). Loss of libido may also be influenced by other causes (e.g., stress, low testosterone), but we have not excluded non-depressive reasons for loss of interest in sex in this analogous sample of young adults.

Considering the results above, both traditional CIT and modern IRT approaches were useful to assess the construct soundness of measures of depression, in this particular type of population and specific circumstances (Fabrizio and De Laet, 1987). Furthermore, the graphical presentation of the CIT indicated that easy items ("changes in sleep") were able to separate only a limited proportion of depressed patients with a high total score. Conversely, moderately difficult ("loss of interest") and very difficult items ("suicidal thoughts") could discriminate with higher confidence all severity levels of depression. This equivalent reasoning also can be visualized in the ICC of the 2-PL IRT model along the latent dimension of depression (θ) in different discrimination ability (a_i) and increasing severity (b_i). While the item "changes in sleep" presented a negative value, on the left side of θ , "loss of interest" was intermediate, and "suicidal thoughts" was on the extreme right of θ . Therefore, both methods were sufficiently applicable for describing the item performance of this scale.

Results comparing CIT and IRT parameters showed strong correlation between b_i and a_i . While these results provide empirical evidence on the complementarity between the two statistical approaches regarding the estimate of item difficulty, the CIT method allowed a rough graphical overview of item analysis when compared with IRT (Lord, 1980; Hershfield and Jones, 1983). Consequently, both CIT and IRT can provide useful frameworks for assessing the measurement properties of psychometric tools such as the EDI-II. Insights into which items should be assigned to a scale can improve its performance through item-level analysis, in particular, for the student samples.

In summary, this study has brought closer the CIT-IRT equivalence relation by employing the widely used set of homogeneous dichotomous items of depression. The correlation analysis ($r_{b,a}$) and parameter for discrimination (a_i) showed that all EDI-II items comprised a good representation of the same continuum of depressive states. It is

also clear from the proportion of endorsement (b_i) and the severity parameter (a_i) that most of the items encompassing the scale were of moderate difficulty, as indicated by their right-sided position along the latent dimension of depression. Therefore, this set of items seems to be able to identify depressive symptoms of sufficient intensity in this student sample. CIT models can be furnished using the reverse application of either of these approaches from corresponding item response models. Both methodologies are authoritative frameworks for last end scale construction and development in the educational, behavioral, and social sciences.

4.1. Limitations

Our study contributes to recommending the utility of EDI-II as a potential measure to document the presence of depressive symptoms in an analogous sample of population evaluated. However, some limitations should be borne in mind in the interpretation of the results.

The present study evaluated a large non-clinical sample and all participants were college students within a restricted age bracket. Even though the results were adequate, direct generalization of latent trait of depressive symptoms for community dwellers or a clinical population should be avoided. In future, our key results may be replicated in more inclusive samples, both gender groups, as well as infrequent and serious depressive states (Jorge et al., 2013).

A clinical assessment was not performed in this study; individuals were only evaluated on the use of drugs in terms of alcohol and psychoactive substance use (Andrade, 2012).

There may possibly be item functioning differences by gender in the manifestation of depression, e.g., "crying" is a symptom known to be embarrassing in some male populations (Castro et al., 2013; Choi et al., 2011). The gender-specific response pattern of depressive symptoms by self-report tools will be explored in future analysis.

5. Conclusions

Depression is often associated with negative educational, economic and social outcomes, increasing the rates of withdrawal from college prior to completion. Early detection and effective treatment of depressive disorders during the college years might foster positive effects in future human capital, either from societal or patients' perspective. Several depressive symptoms are frequently overlooked in the student population, delaying their help seeking or proper treatment. Therefore, our findings expand the understanding of manifestation of depressive symptoms in college students, when an easy self-report measure such the EDI-II was applied. Minimal education and clinicians might refer suspected cases of depression for detailed assessment by mental health specialists.

The present study provided comprehensive insights on item functioning of the popular EDI-II, using CIT analysis to compare and supplement its psychometric properties with IRT methods. Both approaches were informative with respect to which clusters of depressive symptoms merit more attention from clinicians and academic mentors. In order to be propagated internationally, the EDI-II deserves further assessment by coupling the CIT and IRT approaches. Further investigation on measurement characteristics of prevalent parameters of depressive symptoms using modern psychometric methods, such as item-based and person-based approaches, may contribute to the construction of far-reaching rating scales.

Acknowledgments

ARS is awarded with Ph.D. fellowship from Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES/PROEX). The National Research Council (CNPq) supports LEA and CG.

Role of funding source

The National Anti-Drug Secretariat (SENAD), Brazilian Ministry of Health, supported data collection of this research. The agency has no further influence on the results reported herein, the decision to disseminate, the analytical strategy and the contents of this article.

References

- American Psychiatric Association, 1994. *American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders*, Fourth edition. Am. Psychiatric Assoc. American Psychiatric Association, 2013. *American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Washington, DC.
- Andrade, A.G. de, 2012. Use of alcohol and other drugs among Brazilian village residents: effects of gender and age. *Rev. Bras. Psiquiatr.* 34, 294–300. <http://dx.doi.org/10.1596/S0034-8924.2012.00492>.
- Andrade, A.G., Pereira, P.C.A.V., Oliveira, L.O., 2016. In: *National survey on the use of alcohol, tobacco and other drugs among village residents in the 27 Brazilian States*. Brasília: SENAD-MS.
- Assato, P.O., Oliveira, D.J., Baid, J., Gonçalves, J.R., 2009. The Beck Depression Inventory-II (BDI-II): Psychometric properties in Brazilian students and patients population. *Psichiatr. J. Psychiatry* 21, 269–275. http://dx.doi.org/10.1007/978-94-007-1000-0_10.
- Baker, J.B., 2002. The Status of Item Response Theory. *BMJ: Good Practice on Assessment and Evaluation*. <http://dx.doi.org/10.1136/bmj.324.7243.0001>.
- Baker, A., Brown, M., 1986. Beck Depression Inventory-II manual. J. Health Psychol. <http://dx.doi.org/10.1177/026182618601100401>.
- Baker, A.T., Bond, A.J., Brown, R.E., Brown, G., 1979. *Cognitive therapy of depression*. Chgo. Univ. Press.
- Baker, A.T., Ward, C.H., Marchbanks, M., Meek, A., Schacht, J., 1981. Beck Depression Inventory. San Antonio, TX Psychol. Corp.
- Brown, G., Meyer, R.R., Zuckerman, J., 2013. Measuring individual algorithm change on the Beck Depression Inventory-II through IRT-based strategies. *Psychometrika* 78, 499–521. <http://dx.doi.org/10.1007/s11336-012-924-0>.
- Chapman, L.L., Jovan-Lewis, A., Hays, R.D., 2014. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcome measures. *Clin. Ther.* 36, 549–563. <http://dx.doi.org/10.1016/j.clinther.2014.04.005>.
- Castro, A.M., de, J., Ode, M., Thomas, V.J.L., Ribeiro, J., Castro, E.M., de, J., Ode, M., Thomas, V.J.L., Ribeiro, J., 2015. *Permanente diferencial de item no inventário de Depressão Beck*. *Rev. Bras. Epidemiol.* 18, 54–69. <http://dx.doi.org/10.1590/S1519-29662015000010005>.
- Cronbach, L., Allyn, J., 2009. Chapter 7 - Procedures for estimating reliability. In: *Introduction to Classical and Modern Test Theory*. <http://dx.doi.org/10.1177/0146165908300010>.
- Collins, P., Cohen, L.A., Burt, D.M., Xue, H.M., Anisovich, S.P., Bradbury, T., Bender, R.C., 2016. Psychometric properties of depression in college students: a meta-analysis. *Depression Anxiety* 33, 400–414. <http://dx.doi.org/10.1002/dea.22961>.
- Cronbach, L.L., Gage, M.J., Kiehl, J.A., 2014. Overlap of anxiety and depression in children and adolescents: 10 years after. *Psychol. Bull.* 140, 634–645. <http://dx.doi.org/10.1037/xap0000075>.
- Gui, M., Sheng, J.M., Andrade, A.G., 2011. A model for psychiatric questionnaire with alternating items. *Stat. Methods Med. Res.* 26, 451–470. <http://dx.doi.org/10.1177/0962280210369424>.
- Kasselman, S.J., Niles, J.P., 2008. Item Response Theory for Psychologists. *Measurement Applications Ready for Use*.
- Kravitz, R.L., Rogers, R.L., Mowbray, M.J., 2006. The Level of Difficulty and Discrimination Index of the Beck Depression Inventory and BDI-II. *Measurement* 10, 100–104. <http://dx.doi.org/10.1177/1063426905280001>.
- Olson, G.V., Hopkins, K.D., 1996. *Statistical Methods in Education and Psychology*. Allyn and Bacon.
- Osorio-Gutierrez, M.H., Gonzalez, C., Melo, R.L., Andrade, L.H., Wang, Y.P., 2012. Validation of the Brazilian Portuguese Version of the Beck Depression Inventory-II in a community sample. *Rev. Bras. Psiquiatr.* 34, 289–294. <http://dx.doi.org/10.1596/S0034-8924.2012.00492>.
- Hornblum, R.L., Jones, R.W., 1998. Comparison of classical test theory and item response theory and their applications to test development. *TESS - Test Top. Educ. Meas.*
- Hong, E., Wong, H.C., 2006. Rasch rating scale modeling of the Korean version of the Beck Depression Inventory. *Behav. Psychol. Meas.* <http://dx.doi.org/10.1177/0013164406282828>.
- Reichen, A.L., Kelly, L.J., Adams, C.L., Goodwin, C., Tan, C., Kohn, J., et al., 2013. A systematic review of studies of depression prevalence in university students. *J. Psychol.* 147, 391–408. <http://dx.doi.org/10.1080/00220272.2012.711015>.
- Reichen, A.L., Kelly, L.J., Kelly, A.T., Hansen, C.L., Rogers, R.L., 2007. Test and retest reliability regarding use of the Beck Depression Inventory. *Cogn. Ther. Res.* <http://dx.doi.org/10.1007/s10608-006-9094-9>.
- Rosin, R.G., Perelman, M., Schepers, M.A., Redding, A.M., Winick, H.-U., 2013. Twelve-month and lifetime prevalence and lifetime episode rate of panic and agoraphobia in the United States. *Int. J. Geriatr. Psychiatry* 28, 154–164. <http://dx.doi.org/10.1002/gps.2828>.
- Lord, F.M., 1980. *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates.
- Miller, J.M., Kohn, M., 2009. Drinking, smoking, and illicit drug use among 15 and 16 year olds in the United Kingdom. *BMJ* 339, 964–967. <http://dx.doi.org/10.1136/bmj.b2705>.
- Plant, R., Jellison, R., 2008. Drug use has declined among teenagers in the United Kingdom. *Br. Med. J.* 336, 1836–1837.
- Rasch, G., 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1961. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1962. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1963. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1964. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1965. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1966. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1967. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1968. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1969. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1970. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1971. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1972. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1973. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1974. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1975. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1976. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1977. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1978. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1979. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1981. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1982. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1983. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1984. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1985. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1986. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1987. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1988. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1989. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1990. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1991. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1992. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1994. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1995. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1996. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1997. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1998. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 1999. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2000. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2001. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2002. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2003. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2004. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2005. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2006. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2007. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2008. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2009. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2010. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2011. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2012. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2013. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2014. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2015. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2016. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2017. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Rasch, G., 2018. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.

Anexo E

Artigo 2

**Can gender and age impact on response pattern of
depressive symptoms among college students? A
differential item functioning analysis**

Can gender and age impact on response pattern of depressive symptoms among college students? A differential item functioning analysis

Antonio R. De Sa-Junior^{1, 2*}, Graziela Liebel³, Arthur G. Andrade^{2, 4}, Laura H. Andrade², Clarice Gorenstein^{2, 5}, Yuan-Pang Wang²

¹Department of Health Sciences, Federal University of Santa Catarina, Brazil, ²Institute & Department of Psychiatry, Faculty of Medicine, University of São Paulo, Brazil, ³Department of Collective Health, Universidade Federal de Juiz de Fora, Brazil, ⁴Department of Neuroscience, Faculdade de Medicina do ABC, Brazil, ⁵Department of Pharmacology, Institute of Biomedical Sciences, University of São Paulo, Brazil

Submitted to Journal:
Frontiers in Psychiatry

Specialty Section:
Mood and Anxiety Disorders

Article type:
Original Research Article

Manuscript ID:
415719

Received on:
25 Jul 2018

Revised on:
29 Nov 2018

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contribution statement

Acquisition of the study was made by AGA, YPW, LHA, CG, ARS, YPW and CG designed the study. AGA, YPW, CG, ARS provided the data. ARS and YPW designed the analyses, and ARS and GL performed them. ARS, CG and YPW contributed to the interpretation of the data. ARS, GL, CG and YPW wrote the first version of the manuscript. All authors critically revised and approved the final version of the article.

Keywords

Depression, gender, age, college students, Beck Depression Inventory - Second Edition, item response theory, Differential Item Functioning

Abstract

Word count: 274

Background: Self-reported depressive complaints among college students might indicate different degrees of severity of depressive states. Through the framework of item response theory, we aim to describe the pattern of responses to items of the Beck Depression Inventory-II (BDI-II), in terms of endorsement probability and discrimination along the continuum of depression. Potential differential item functioning of the scale items of the BDI-II is investigated, by gender and age, to compare across sub-groups of students.

Methods: The 21-item BDI-II was cross-sectionally administered to a representative sample of 12,677 Brazilian college students. Reliability was evaluated based on Cronbach's alpha coefficient. Severity (bi) and discrimination (a) parameters of each BDI-II item were calculated through the graded response model. The influence of gender and age were tested for differential item functioning (DIF) within the item response theory-based approach.

Results: The BDI-II presented good reliability ($\alpha = 0.91$). Women and younger students significantly presented higher likelihood of depression (cut-off > 13) than men and older counterparts. In general, participants endorsed more easily cognitive-somatic items than affective items of the scale. "Quality feelings", "suicidal thoughts", and "loss of interest in sex" were the items that most highly indicated depression severity (b a 3.66). However, all BDI-II items showed moderate-to-high discrimination (a a 1.23) for depressive status. While two items were flagged for DIF, "crying" and "loss of interest in sex", respectively for gender and age, the global weight of these items on the total score was negligible.

Conclusions: Although respondents' gender and age might present influence on response patterns of depressive symptoms, the measures of self-reported symptoms have not inflated severity scores. These findings provide further support to the validity of using BDI-II for assessing depression in academic contexts and highlight the value of considering gender- and age-related common symptoms of depression.

Ethics statements

(Authors are required to state the ethical considerations of their study in the manuscript, including for cases where the study was exempt from ethical approval procedures)

Does the study presented in the manuscript involve human or animal subjects: Yes

Please provide the complete ethics statement for your manuscript. Note that the statement will be directly added to the manuscript file for peer-review, and should include the following information:

- Full name of the ethics committee that approved the study
- Consent procedure used for human participants or for animal owners
- Any additional considerations of the study in cases where vulnerable populations were involved, for example minors, persons with disabilities or endangered animal species

As per the Frontiers authors guidelines, you are required to use the following format for statements involving human subjects: This study was carried out in accordance with the recommendations of [name of guidelines], [name of committee]. The protocol was approved by the [name of committee]. All subjects gave written informed consent in accordance with the Declaration of

Helsinki.

For statements involving animal subjects, please use:

This study was carried out in accordance with the recommendations of 'name of guidelines, name of committee'. The protocol was approved by the 'name of committee'.

If the study was exempt from one or more of the above requirements, please provide a statement with the reason for the exemption(s).

Ensure that your statement is phrased in a complete way, with clear and concise sentences.

The Ethics Committee for the Analysis of Research Projects at the University of São Paulo Medical School approved the present study. All participants provided written informed consent prior to participating in the data collection.

In review

1 Can gender and age impact on response pattern of depressive
2 symptoms among college students? A differential item
3 functioning analysis

4

5 Antonio Reis de Sá Junior^{1,2}, Graziela Liebel³, Arthur Guerra de Andrade^{2,4}, Laura
6 Helena Andrade², Clarice Gorenstein^{2,5}, Yuan-Pang Wang^{2,*}

7 Affiliation:

- 8 1. Department of Health Sciences, Federal University of Santa Catarina, Araranguá**
9 SC, Brazil.
10 2. Institute & Department of Psychiatry (LIM-23), University of São Paulo
11 Medical School, São Paulo SP, Brazil.
12 3. Department of Collective Health, Federal University of Juiz de Fora, Juiz de
13 Fora, MG, Brazil.
14 4. Department of Neuroscience, Medical School, ABC Foundation, Santo André
15 SP, Brazil.
16 5. Department of Pharmacology, Institute of Biomedical Sciences, University of
17 São Paulo, São Paulo SP, Brazil.

18 Correspondence: Dr. Yuan-Pang Wang, Núcleo de Epidemiologia Psiquiátrica (LIM-
19 23), Departamento & Instituto de Psiquiatria, Faculdade de Medicina, Universidade de
20 São Paulo, Rua Dr. Ovídio Pires de Campos, 785, CEP 05403-010, São Paulo, SP,
21 Brasil.

22 E-mail address: gnap_inbox@hotmail.com

23 Running Title: Gender and age impact on depression

24

25 Statistics:

26 Number of words: 5782

27 Number of figures: 3 Tables and 2 Figures

28

29 ABSTRACT

30 **Background:** Self-reported depressive complaints among college students might
 31 indicate different degrees of severity of depressive states. Through the framework of
 32 item response theory, we aim to describe the pattern of responses to items of the Beck
 33 Depression Inventory-II (BDI-II), in terms of endorsement probability and
 34 discrimination along the continuum of depression. Potential differential item
 35 functioning of the scale items of the BDI-II is investigated, by gender and age, to
 36 compare across sub-groups of students.

37 **Methods:** The 21-item BDI-II was cross-sectionally administered to a representative
 38 sample of 12,677 Brazilian college students. Reliability was evaluated based on
 39 **Cronbach's alpha** coefficient. Severity (b_i) and discrimination (a_i) parameters of each
 40 BDI-II items were calculated through the graded response model. The influence of
 41 gender and age were tested for differential item functioning (DIF) within the item
 42 response theory-based approach.

43 **Results:** The BDI-II presented good reliability ($\alpha = 0.91$). Women and younger students
 44 significantly presented a higher likelihood of depression (cut-off > 13) than men and
 45 older counterparts. In general, participants endorsed more easily cognitive-somatic
 46 items than affective items of the scale. "Guilty feelings", "suicidal thoughts", and "loss
 47 of interest in sex" were the items that most likely indicated depression severity ($b \geq$
 48 3.60). However, all BDI-II items showed moderate-to-high discrimination ($a \geq 1.32$) for
 49 depressive state. While two items were flagged for DIF, "crying" and "loss of interest in
 50 sex", respectively for gender and age, the global weight of these items on the total score
 51 was negligible.

52 **Conclusions:** Although respondents' gender and age might present influence on
 53 response pattern of depressive symptoms, the measures of self-reported symptoms have
 54 not inflated severity scores. These findings provide further support to the validity of
 55 using BDI-II for assessing depression in academic contexts and highlight the value of
 56 considering gender- and age-related common symptoms of depression.

57
 58 **KEYWORDS:** Depression, gender, age, college students, Beck Depression Inventory,
 59 item response theory, differential item functioning.

60

61 INTRODUCTION

62 Depression is among the most prevalent psychiatric disorders worldwide and has been
63 associated with some demographic determinants, including gender and age. The
64 prevalence of depression is higher among women when compared to men, with evident
65 sex ratio of 2:1 of diagnosed depression after adolescence (1,2). The underlying
66 mechanism of this gender difference is not fully understood. Possibly, the reported
67 variation is due to the joint effect of biological factors, including genetic, physiologic
68 and endocrine determinants, and social or demographic factors, such as gender-specific
69 roles and coping strategies (3).

71 Depression scales are commonly assumed to measure the same attributes for both males
72 and females. Insufficient evidence of measurement bias compromise the conclusions
73 based on gender group. Studies regarding gender differences in depression usually
74 depend on the mean comparisons estimated by depression scales. However, the mean
75 differences in depression can be attributed to a true difference, measurement bias, or a
76 combination of both. It appears that few studies have investigated the gender-related
77 measurement bias of the BDI-II score when making a comparison of gender (4).

79 Regarding age, late adolescence is a transitional period during which young adults
80 experience stressful socio-biological transformations. Notably, psychiatric disorders are
81 common among college-aged individuals (5). The first onset of depression in adulthood,
82 particularly among college students around their 20s, appears to be higher than in those
83 non-college adults in their 30s or 40s (6–8). Frequently, undergraduate students have to
84 cope with part-time work and family demands in addition to academic requirements.
85 Also, searching for a formal job in the working world is an additional distress for many
86 students before completing graduation. Thus, the observed gender and age-related
87 differences of emotional symptoms among college students may indicate which group
88 of individuals would most benefit from prevention and timely management of
89 depression.

91 Traditionally, the diagnosis of depression is reached through standardized interviews
92 and complemented with assessment scales. Most instruments for symptomatic
93 assessment of depression are built on the ground of classical test theory (CTT), where
94 the item scores are summed to indicate severity and a same weight is attributed to each
95 scale item. In this classical approach, the probability of item endorsement changes in
96 accordance with the characteristics of participants that take a specific test, obstructing
97 the comparability of results (9,10).

99 For instance, the 21-item Beck Depression Inventory-II (BDI-II) is one of the most
100 commonly used self-report measures of severity of depressive symptoms (11). The
101 popularity of the BDI-II has been attributed to its rigorous psychometric evaluation
102 (12). The BDI-II may be scored either by summing a number of equal-weighted items
103 or fitting a measurement latent model, yielding a cumulative score as a continuous
104 variable. Nevertheless, the classical method cannot provide direct guidance regarding
105 the ability of BDI-II to assess depressive symptomatology at various points of the
106 severity continuum of depressive syndrome. Individuals reporting suicidal thoughts may
107 indicate more severe states of depression than those with sleeplessness. Common
108 symptoms in academic environments, like fatigue or loss of energy, changes in sleep
109 and appetite, might be misinterpreted as reliable mood-related indicators of severity
110 (13). In medical contexts, while some individuals report more somatic symptoms,

others may predominantly complain of cognitive-affective symptoms (14). Item analysis of the scale (15) may examine the confounding effect of the somatic symptoms on the measurement of depressive symptoms in people diagnosed with a chronic disease (16,17)

In recent decades, research demonstrated that diagnostic and screening instruments could benefit from refined statistical approaches, such as the item response theory (IRT) models (10). This family of psychometric methods is considered an appropriate and robust approach to address the measurement properties of rating scales, for example, those built for measuring depression (15). The basic assumption in IRT asserts that the latent trait θ (theta, in Greek) of a participant is independent of the content of a test (18). The IRT model provides information on the relationship between responses to a set of items (probability of endorsement/non-endorsement of depressive symptoms, or parameter b_i) and the value of a person's latent trait (difficulty or the severity of depression). Each item can be associated with a discrimination parameter (a_i), which denotes the extent to which responses on the item reliably indicate differences between persons' overall severity scores. High item discrimination suggests the symptom has a stronger association with the underlying depression dimension θ (19–21).

The IRT can examine item bias of a psychometric tool across diverse population groups. The differential item functioning (DIF) tests how similar is the between-group item perform (22). The measurement equivalence is a necessary condition for meaningful comparison among different sub-groups. The variety of existing depression scales, the methodological differences of assessment, and the diversity of group variables examined for DIF across the reviewed studies make the synthesis of findings on depression a challenging task (13,19,20,23).

Hypothetical DIF can be generated based on the fact that women present higher scores on self-reported depression measures than men (24,25) and there are also age-related differences in the manifestation of depressive symptoms (26). Consequently, some symptoms would reflect a more severe depressive state than other symptoms among college students. As the BDI-II is intended to be a clinically useful measure of depression, it is expected that some items might discriminate between individuals with high and low levels of depression. Therefore, group-level differences are expected, since the assessment of rates or severity of depression might reflect DIF with respect to expression of symptoms by gender or age sub-groups.

In the current study, the item performance of the BDI-II was examined in a representative sample of Brazilian college students. The underlying construct considered herein is "self-reported depression". The aim of the current paper was to present item-level psychometric findings of the BDI-II for college students, in the form of IRT parameters and DIF analysis regarding gender and age. Strengths and limitations of results found in item analyses are discussed.

MATERIALS AND METHODS

Ethics Statement

The Ethics Committee for the Analysis of Research Projects at the University of São Paulo Medical School approved the present study. All participants provided written informed consent prior to participating in the data collection.

Sampling frame

The target population was students enrolled in the “I Nationwide survey on the use of alcohol, tobacco, and other drugs among college students in the 27 Brazilian state capitals” (27). This sample was recruited through two-stage cluster sampling. The first stage involved higher education institutions (HEIs) that deliver undergraduate courses in each of 27 state capitals, in accordance with the school list provided by the Brazilian Ministry of Education. Eligible HEIs were randomly selected from the list of ordered institutions, by the means of the probability proportional to size (PPS) of the number of enrolled students in each HEI. The stratification was based on the two variables: the location of capital and the type of institution funding (either private or public), totaling 54 levels. Of the selected 114 colleges and universities, 100 HEIs (87.7%) agreed to participate after institutional invitation.

In the second stage, the class of students was considered the primary sampling unit (PSU) within each selected HEI, for the sake of feasible fieldwork. Each element of the class was defined as a group of students attending a given course. The number of classes was proportional to the total number of students in the institution, totaling 654 classes or 70.6% of selected PSU. All students of the selected classes were invited to volunteer for the study. The participants’ response rate was 95.6% for the college students who were in classes at the time of the survey (27).

Participants

The initial sample of the survey was 12,721 college students. Of these, 10 respondents were excluded because they claimed to have used the dummy drug “relewin” and 34 did not report gender. Hence, a final sample of 12,677 valid questionnaires (99.7%) was considered for analysis.

All students completed a structured, anonymous, self-administered instrument consisting of 98 closed questions on the use of substances, wherein the 21-item Beck Depression Inventory-II (BDI-II) was included. The demographic characteristics of the sample were 5,692 men (44.9%) and 6,985 women (55.1%), with the mean age of 23.9 years (SD = 6.9; range 16-84 years). On average, the mean age of women was similar to men ($M = 23.8$ vs. 24.0 , $p = 0.15$). Because the distribution of age was left-skewed, with the greatest majority of participants aged in late adolescence and 20s, the age group was dichotomized as younger participants (16–30 years old; $n = 10,887$, $M = 21.7$ and $SD = 3.1$) and older participants (31–84 years old; $n = 1,790$, $M = 38.7$ and $SD = 7.0$). Older participants were divided as follows: 31–40 years old 66.8%; 41–50 years old 26%; 51–60 years old 6.6%; 61–70 years old 0.4% and 71–84 years old 0.2%. Additionally, the self-reported skin color – as the proxy for race/ethnicity in Brazil – showed that most of the sample consisted of white students (55.5%), followed by black (6.8%), mulatto (29.8%), Asian (2.9%), Indian (0.8%), other mixed colors (3.0%) and 1.3% no reply. When asked about “What is your present religion?” the answers were: 16.3% with no religion, 53.0% Catholic, 18.4% Evangelical or Protestant, 11.8% others

religions and 0.5% no reply. Marital status had the following rates: 80.5% single, 16.9% married, 1.7% divorced, 0.1% widowed and 0.7% no reply.

The Beck Depression Inventory-II (BDI-II)

The BDI-II is a 21-item self-report questionnaire for assessing the severity of depressive symptoms, which items consist of four ordinal categories (from 1 through 4). Absence (or "as usual") of depressive symptom in each item is scored as "0" and an increasing presence of symptoms is endorsed between 1 and 3. High scores, either in an item or in the scale, indicates more intense symptom severity. Possible total score ranges from 0 to 63. The scale content reflects the cognitive, affective, somatic, and vegetative symptoms of depression (11).

The timeframe of the BDI-II covers the 2 weeks requirement to meet the DSM-IV criteria for major depressive disorder (28). Substantial evidence of reliability has been reported in different language versions of the BDI-II (11), including the Brazilian-Portuguese (29). The BDI-II has been validated for psychiatric and medical patients, students, and adolescent samples (11,12,30). Considerable evidence of construct validity has been demonstrated in college student samples (31–33).

Statistical analyses

First, the **Cronbach's alpha (α)** coefficient of internal consistency was calculated. Characteristics of BDI-II data dispersion and homogeneity were explored for each scale item and a total score: mean (M), standard deviation (SD), **Pearson's coefficient of variation (CV)** and **Cronbach's α** if the item is deleted (item-total analysis).

There are two assumptions for IRT: unidimensionality (i.e., there is only one underlying factor within the data) and local independence (items should not be correlated when the shared variance of the latent trait is removed). Both assumptions were tested through confirmatory factor analyses (CFA), to examine the covariance structure of the scale. To evaluate model fits, one absolute close-fit index, that is, root mean squared error of approximation (RMSEA) and two incremental close-fit indexes, that is, comparative fit index (CFI) and Tucker-Lewis index (TLI) were used (34). The following cut-off points were considered as indicative of an adequate model fit: $RMSEA < 0.06$, CFI and $TLI \geq 0.95$ (35,36). The results of CFA approximately fitted data to unidimensional model: $\chi^2_{(210)} = 9054.21$, $p < 0.001$, $CFI = 0.95$, $TLI = 0.94$, and $RMSEA = 0.06$. The **significance of χ^2** test was magnified due to the large sample size. These indicators favored the use of IRT model (37).

For a polytomic ordinal scale, the graded response model (38) was adequate to describe the relationship between the observed responses to the BDI-II and the underlying latent trait θ (theta), as the proxy of depression severity. This model yields discrimination (a) and severity or threshold (b) parameters for each criterion. Scale items (i) with a large slope (a) provide more information or discrimination for the depressive symptoms. The **item difficulty (b_i) is distributed along a latent dimension θ (theta) to reflect the trait of severity in relation to absence or presence of the pathological condition**. In health sciences, the b score of a given scale item (i) used to locate in the right side of the latent **variable θ to denote more severe manifestation of the target illness** (21). For interpretation, Baker (21) suggests that discrimination parameter values a ranging from

0.01 to 0.24 are considered very low, 0.25 to 0.64 low, 0.65 to 1.34 moderate, 1.35 to 1.69 high, and more than 1.7 very high.

Detection of differential item functioning (DIF) was performed in the framework of iterative hybrid ordinal logistic regression (OLR) (39). This state-of-the-art approach incorporates IRT-derived trait scores θ as the estimator of the “ability” to score an item, rather than the observed BDI-II domain cumulative score as in OLR. The generic term “ability” can be used hereby to represent the trait score as measured by the test, either as the observed sum score on the BDI-II or the latent variable (the b_i parameter). The dependent variable of regression models was the item response (0–3) and the independent variable was the IRT-derived trait score, i.e., the level of depression on the BDI-II scale.

Analyses were conducted to identify biased items showing gender and age difference. Respondents’ gender was dichotomized as men or women and the age as younger or older participants. For each item, a null model and three nested models were estimated hierarchically with additional explanatory variables. In the null model 0, only the intercept was considered in the equation. In the baseline model 1, the trait score θ was used as the predictor of the item response in addition to the intercept. In model 2, both trait score θ and gender/age predicted the item response (DIF). In model 3, an interaction term was included between the trait score θ and gender/age.

Comparisons were performed to assess the magnitude and the type of DIF for flagged items, by the means of likelihood-ratio χ^2 and pseudo R^2 statistics for comparing three nested logistic regression models. First, if the difference between models 1 vs. 3 showed an R^2 difference > 0.02 , the flagged item was considered to show DIF (39). Second, if the difference between models 2 vs. 3 showed an R^2 difference > 0.02 , then the item was considered to show ‘non-uniform DIF’ because the effect varies conditional to the trait level. Third, if the difference between models 1 vs. 2 showed an R^2 difference > 0.02 , the item was considered to show ‘uniform DIF’, because the effect was constant. Final comparison tested whether DIF was uniform, i.e., consistent impact across endorsement probability or non-uniform, i.e., the probability that varied according to levels of endorsement.

We plotted the item characteristic curve (ICC) of the DIF items for both genders and age to inspect the direction of the difference. Finally, the test characteristic curve (TCC) provides insight into the actual impact of DIF items on the total score.

While descriptive analyses and CFA were computed through Mplus 6.12 (www.statmodel.com), the estimates of the IRT model and DIF analyses were calculated through R 3.5.0 (cran.r-project.org). The *lordif* library was implemented to compute the regression models. The level of significance was set as 0.05 for two-tailed tests.

RESULTS

Reliability and descriptive analysis

Cronbach’s α reliability was 0.91 and did not differ between genders (men = 0.90; women = 0.91) or age (younger and older participants = 0.91). Descriptive statistics of

responses obtained from 12,677 respondents for BDI-II are shown in Table 1, the mean total score was 6.99 (SD 7.56, CV 1.08). The average BDI-II total score between women ($M = 7.85$, $SD = 8.08$) and men ($M = 5.92$, $SD = 6.72$) or younger participants ($M = 7.17$, $SD = 7.64$) and older participants ($M = 5.83$, $SD = 6.91$) were significantly different ($p < 0.001$). Also, using the recommended cut-off score of 13 (Beck et al., 1996), the proportion of probable depression was significantly different in relation to gender (men = 14.2%; women = 22.1%; $p < 0.001$) and age (younger participants = 19.2%; older participants = 14.7%; $p < 0.001$).

Table 2 presents the frequency distributions for five arbitrarily selected items: "sadness" (#1), "guilty feelings" (#5), "crying" (#10), "loss of energy" (#15) and "loss of interest in sex" (#21). The frequency distributions for 21 items are presented in a table as a supplementary file. Inspecting the table, the great majority of respondents endorsed the category 1 of items, which denotes "as usual" or "unchanged" manifestation of depressive symptoms. The symptomatic categories of 2, 3, and 4 were endorsed in decreasing frequency, intensity or distress, to indicate the severity of depressive symptoms. As expected for this non-clinical sample of analogous students, the more severe the symptom, the less it was chosen by the respondents. Examining the columns for gender and age groups, a considerable difference was observed for specific items. For example, women endorsed more "sadness", "crying" and "loss of energy" than men did. On the other hand, while "crying" was more likely endorsed by younger participants, "loss of interest in sex" was more reported by older counterparts.

The proportion of responses to each item was not uniform, which variation suggests a differential probability of endorsement rate according to discrimination and difficulty level of an item in the BDI-II. Moreover, these results indicate noteworthy differences in the gender- and age-related response pattern of self-reported depressive symptoms. Item analysis and differential item functioning (DIF) are explored below to understand their impact on the between-group difference of depression, both by the mean score and by the cut-off threshold.

Item analysis

Table 3 shows discrimination (a) and severity threshold (b_i) estimates of depressive symptoms for all 21 items. The discrimination (a) parameters ranged from 1.32 (#21 "loss of interest in sex") to 3.31 (#14 "worthlessness"), suggesting that all scale items presented moderate-to-high discriminatory characteristics for the latent trait of "self-reported depression" (θ).

All severity or difficulty parameters (b_i) presented positive value, on the right side of the θ . The first (b_1), second (b_2), and third (b_3) threshold parameters ranged from 0.03 to 1.91, 1.60 to 3.26 and 2.12 to 4.44 respectively. The ascending sequence of the thresholds b_1 , b_2 , and b_3 confirmed the appropriate direction of response options. The BDI-II item location parameters (b_i threshold) were fairly consistent over an increasing gradient of the score in terms of latent trait θ .

For the threshold parameter b_1 , the easiest BDI-II items were #16 "changes in sleep", #8 "self-criticalness", and #20 "tiredness of fatigue". The more severe items for b_1 were #21 "loss of interest in sex", #9 "suicidal thoughts", and #6 "punishment feelings". In the same direction, for the threshold parameter b_3 , the easiest BDI-II items were #13 "indecisiveness", #20 "tiredness of fatigue", and #7 "self-dislike". The more severe

items on b_3 were #21 "loss of interest in sex", #5 "guilty feelings", and #9 "suicidal thoughts". Worth noting, the small value of standard error (SE) of all parameters indicated the good precision of the estimates in this large sample. Furthermore, the higher the endorsement rate of a given item, the smaller or more precise was the value of the SE.

Because the threshold parameters provided uneven information of score threshold distribution, we examined the pattern of response for gender and age sub-groups, by the means of DIF analysis.

Differential item functioning

Regarding gender-related DIF, the *lordif* method detected the item #10 ("crying") as a problematic element; with women more likely endorsed this symptom than men. For age-related DIF, item #21 ("loss of interest in sex") also indicated a potential difference in the response pattern; with younger participant less likely endorsed this symptom than older counterparts.

In gender and age analyses implemented by ordinal logistic regression (OLR), the pseudo R^2 differences between models 1 vs. 3 were respectively 0.024 and 0.027, which indicated overall DIF. On the other hand, the pseudo R^2 differences between models 2 vs. 3 were < 0.001 for both items and indicated the presence of uniform DIF. The complete table of the OLR models for all BDI-II items is not shown, but it is available upon request to the authors.

The item characteristic curve (ICC) showed that participants from women and older sub-groups present higher likelihood of endorsing items #10 and 21 than men and younger participants, respectively (Figure 1). The response curves for both items are displayed in an ordered sequence along the latent trait θ . Dotted curves indicate that, given a similar latent trait level, the women and older participants had a higher chance to endorse these items than men and younger counterparts. From a further investigation of the ICC for male respondents, it is indicated that the men sub-group presented a substantial overlap for b_1 and b_2 threshold on item #10. This suggests that the threshold for endorsing "crying" in men's sub-group needs be re-calibrated.

Figure 2 displays the test characteristic curve (TCC) for all items and gender/age sub-groups. The figures show the overall impact of DIF for gender and age on TCCs for all 21 BDI items (ignoring DIF) and the individual impact of DIF-anchored item #10 and #21 on TCC. Inspecting the plots, it is clear that there was no substantial influence of DIF in items #10 and #21 on the total domain score, because of the similar slopes and overlap in scorings between the two sub-samples. These findings indicated a minimal impact of DIF by gender or age, even though the expected total score was higher for women than men and older adults than younger adults. To summarize, although items #10 and #21 showed DIF, these items only had a negligible impact on the total score.

DISCUSSION

The present study has brought remarkable information on gender and age bias on the response pattern of depressive symptoms in a large representative sample of college

students. According to classical test theory framework, the BDI-II has shown sufficient internal consistency to capture the underlying construct of "major depression". While all 21 items have proved to discriminate well the latent trait of "depression", some items were examined with respect to respondent's characteristics. The major findings indicated that self-reported symptoms of "crying" and "loss of interest in sex" might modify the estimate of depression. Thus, this item-wise psychometric study provided an opportunity to compare the measurement of depression across different groups.

Regarding discrimination power (the a parameter), the item response analysis has indicated that all BDI-II items have presented at least moderate ability (21) to discriminate the latent trait along the severity continuum of "depression". In comparison, the somatic items ("loss of interest in sex" and "changes in appetite") were less informative than affective items ("worthlessness" and "self-dislike"). Therefore, all BDI-II items were suitable to depict depression among college students.

Regarding the item difficulty (the b_i parameter), the most endorsed items were the least severe ones (13,40,41). The excess of sub-threshold depressive symptoms (total score < 13) in this student sample suggested that the BDI-II items were hard to endorse but capable of detecting depressive symptoms within a wide range of severity. If a scale purposing to measure depression were composed only of items measuring mild depressive states, the instrument would have great trouble for distinguishing between moderate and severe cases of depression: both would be characterized by high scores on all items. On the other hand, if a scale's items easily detect somatic symptoms of depression, cognitive-affective phenotypes of depression might remain unnoticed (42–44). In general, the BDI-II items cover either somatic types or cognitive-affective types of depression in a single dimension of severity.

In theory, individuals with a given depression severity level also should have the same likelihood of scoring a given item response. When men and women or younger and older adults with the same level of depression severity systematically respond to an item in different ways, the item is malfunctioning for different groups of respondents (45). Therefore, the DIF analysis examines whether or not the likelihood of item endorsement is equal across gender and age sub-groups that are matched on the state or trait measured (e.g., depression). When DIF is absent across 21 items of the BDI-II, meaningful comparison of the distribution of latent scores across heterogeneous populations might be accepted. If the DIF is present and ignored, estimates of prevalence or severity might be biased.

In the present study, item "crying" (#10) exhibited significant gender DIF and item "loss of interest in sex" (#21) exhibited significant age DIF. Also, these items were found to show uniform DIF based on the pseudo R^2 change. A high conditional endorsement was observed for women on #10 and older participants on #21. The direction of item malfunctioning indicated that a sample with more women than men might overestimate the prevalence and severity of depression due to #10. Conversely, a younger-participant sample might under-estimate cases of depression due to #21.

Episodes of crying can be viewed as an adaptive coping response to stress, but it should not be automatically interpreted as a sign or symptom of depression. In the present study, crying spells were reported by almost one in four respondents. This behavior has moderately discriminated the severity of depression and displayed a modest threshold to be expressed by the students. However, there is no rigorous empirical foundation on the

relationship between crying and depression (46,47). Most assessment instruments and diagnostic systems for mood disorders are unreliable in how they handle crying as a symptom (48). For a higher diagnostic accuracy, it is important that the criteria and instruments used to assess depression adequately reflect female and male common depressive symptoms.

Most observed among women, the item "crying" as a gender-bounded emotional response has been claimed in most studies with depressive scales to demonstrate DIF (19,20,49–54). Although significant gender DIF also was found in our investigation for "crying", the meaning of this symptom to the assessment of depression remains a matter of debate. We have shown that the thresholds for "crying" seem to be quite different among female and male adults. While female students were more easily tearful than male colleagues, the overall score did not impact on the severity assessment of depression. Possibly, our gender-balanced large sample might have concealed substantial evidence of this difference. In order to prevent over- or under-estimate of depression, researchers must beware when interpreting the results of studies with the dominance of female participants, and vice-versa.

Sexuality, both libido and sexual functioning, is viewed as an important component of emotional and physical intimacy that most people experience throughout their lives. Depressive symptoms have been associated with low sexual desire in 50% to 60% of untreated patients with diagnosed depression (55). For instance, Toosi et al (56) showed lower endorsement of the "loss of interest in sex" were depressive symptoms associated with ageing. Many medical diseases become increasingly prevalent with ageing, blurring boundaries between depression symptoms and physical complaints. As a consequence, depression may remain under-detected because both types of symptoms may co-occur and interact as physical-mental comorbidities (57,58). Other sources of confusion include unwillingness among elderly individuals to not disclose their sexual life, stigma to psychiatric labels (59), and the belief that depression is simply expected when they get older (60).

When measured as a continuous variable through sum scoring, factor analysis or item response model, levels of depression generally increase with age (61–63). However, many health care professionals are reluctant to address this issue during a routine encounter, without an active inquiry of embarrassing topics (64). In this way, it is not surprising that sexual dysfunction remains an under-recognized problem in clinical practice. Likewise to "crying", the impact of "loss of interest in sex" should be understood in light of the age distribution of the sample.

Differences in the measurement of variables such as gender or age might partially explain the discrepancy between previous prevalence studies using symptom scales. Taking advantage of recent psychometric advance of IRT, the detection of DIF for a specific variable did not inevitably indicate that a given item is biasing the reported estimates. Measurement bias is typically identified after considerable empirical evidence has been gathered and reviewed by experts to determine the potential impact of the DIF (23). In the current study, it was suggested that the BDI-II appears to perform similarly for men and women or for younger and older adults. Therefore, gender or age-based differences in prevalence estimates of depressive symptoms are less important than previously supposed. Future research using representative

epidemiological datasets is warranted to replicate the key findings reported in the current study on the effects of gender or age in depressive psychopathology.

To summarize, studies assessing the influence of gender or age on depressive symptoms has suggested that the prevalence of depression can vary across groups of participants (65,66). The assessment of depressive symptoms by BDI-II may lead to an overestimation of symptoms among women, as well as to lower reported rates of depression among younger adults. Some items of the BDI-II were amended to match the criteria for DSM-IV depression in 1996, therefore the underlying construct of the inventory encompasses either gender-bounded subjective complaints (e.g., feels sad or empty) or observations made by others (e.g., appears tearful). As a consequence, men with depression may have not been properly identified in clinical practice and remained under-treated. In the same direction, age-related changes in the sexual drive may also be mixed-up with depression-related symptoms, effects of biological transformation and non-disclosure due to social stigma. The current DSM-5 system recognizes heterogeneous phenotypes of depressive symptoms in a dimensional model of severity. Hence, item response approach that accounts for the response probability of the BDI-II in a person-centered approach may better support the current definition of major depression.

Limitations

The strength of the current analysis is related to its statistical power and precision of estimates, both derived from the large sample size and gender-balanced distribution of non-clinical participants. However, the interpretation of the results should be considered in light of some limitations. First, despite the sample size, there were a number of response options that were endorsed at low frequency (i.e., less than 5%). Thus, some parameter estimates may be imprecise. Also, clinical diagnosis of major depression and physical evaluation were not feasible to undertake in such a huge sample.

Second, the non-clinical participants have self-reported the increasing severity level of depressive symptoms through the items of BDI-II. Thus, information bias due to the data collection method might alter the final results. A previous validity study has shown that the self-rated score of the BDI-II was correlated with the assessment of severity in 65.4% of cases of clinically diagnosed DSM-IV major depressive episode (29).

Third, the cross-sectional data of this study would not allow additional inference on factors affecting the different stages of depression severity, covering the onset, the active state, and remission period of depression. A longitudinal design might overcome this deficiency. Fourth, other demographics captured or analysis within-subjects' differences could have also impacted the levels of severity of depressive symptoms, however, they have not been investigated in the current study.

The last concern is related to the generalizability of the findings. Although we have recruited a representative sample of Brazilian college students (the majority in their 20's), the influence of educational level on the BDI-II's psychometric properties was not evaluated in the present study. Therefore, the key findings of this study on the use of the BDI-II should not be extended to the general population or different schooling and age bracket. Further studies should be carried out using a more inclusive sample, with

participants from evenly distributed age brackets, diverse educational levels, and recruited from clinical settings.

Conclusion

Key results of the present investigation have indicated the appropriateness of the BDI-II items as an assessment measure of the severity of depression in a college population. The correct detection of depressive symptoms in student samples may have preventive implications for the community because they are passing by the peak period of onset of depression (6,67). The item response analysis has indicated that gender and age are variables that hold potential influence on the manifestation of depression.

Although the items "crying" and "interest in sex" have shown noticeable differential item functioning on self-report scales, the impact was trivial. When the BDI-II is applied to a population sample with skewed demographic characteristics or insufficient sample size, the results might be altered likewise. Therefore, it is important that the criteria and instruments used to assess depression can adequately reflect gender- and age-related common symptoms and experiences of depression. The findings of the present investigation provide further support for the validity of using BDI-II for assessing depression in academic contexts.

Acknowledgments

ARS is awarded Ph.D. fellowship from Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES/PROEX). The National Research Council (CNPq) supports LHA and CG.

Author Contributions

Acquisition of the study was made by AGA, YPW, LHA, CG. ARS, YPW, and CG designed the study. AGA, YPW, CG, ARS provided the data. ARS and YPW designed the analyses, and ARS and GL performed them. ARS, CG, and YPW contributed to the interpretation of the data. ARS, GL, CG, and YPW wrote the first version of the manuscript. All authors critically revised and approved the final version of the article

Funding

The National Anti-Drug Secretariat (SENAD), Brazilian Ministry of Health, supported data collection of this research. The agency has no further influence on the results reported herein, the decision to disseminate, the analytical strategy and the contents of this article.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

586 REFERENCES

- 587 1. Hasin DS, Goodwin RD, Stinson FS, Grant BF. Epidemiology of major
588 depressive disorder: results from the National Epidemiologic Survey on
589 Alcoholism and Related Conditions. *Arch Gen Psychiatry* (2005) 10:1097-106.
590 doi:10.1001/archpsyc.62.10.1097
- 591 2. Kuehner C. Gender differences in unipolar depression: An update of
592 epidemiological findings and possible explanations. *Acta Psychiatr Scand* (2003)
593 3:163-74. doi:10.1034/j.1600-0447.2003.00204.x
- 594 3. Afifi M. Gender Differences in Mental Health. *Singapore Med J* (2007) 5:385-
595 91. PMID: 17453094
- 596 4. Wu PC, Huang TW. Gender-related invariance of the Beck Depression Inventory
597 II for Taiwanese adolescent samples. *Assessment* (2014) 21:218-26. doi:
598 10.1177/1073191112441243.
- 599 5. Blanco C, Okuda M, Wright C, Hasin DS, Grant BF, Liu SM, et al. Mental health
600 of college students and their non-college-attending peers: Results from the
601 national epidemiologic study on alcohol and related conditions. *Arch Gen*
602 *Psychiatry* (2008) 12:1429-37. doi:10.1001/archpsyc.65.12.1429
- 603 6. Bayram N, Bilgel N. The prevalence and socio-demographic correlations of
604 depression, anxiety and stress among a group of university students. *Soc*
605 *Psychiatry Psychiatr Epidemiol* (2008) 8:661-72. doi:10.1007/s00127-008-0345-
606 x
- 607 7. Ibrahim AK, Kelly SJ, Adams CE, Glazebrook C. A systematic review of studies
608 of depression prevalence in university students. *J Psychiatr Res* (2013) 3:391-
609 400. doi:10.1016/j.jpsychires.2012.11.015
- 610 8. Young CB, Fang DZ, Zisook S. Depression in Asian-American and Caucasian
611 undergraduate students. *J Affect Disord* (2010) 1-3:379-82.
612 doi:10.1016/j.jad.2010.02.124
- 613 9. Crocker L, Algina J. Introduction to classical and modern test theory. *Introd to*
614 *Class Mod Test Theory* (1986)
- 615 10. Hambleton R, Swaminathan H. Item Response Theory. Principles and
616 applications. *Bost Kluwer Nijhoff Publ* (1985)
- 617 11. Beck A, Steer R, Brown G. *Beck Depression Inventory-II*. (1996).
- 618 12. Wang Y-P, Gorenstein C. Assessment of depression in medical patients: A
619 systematic review of the utility of the beck depression inventory-II. *Clinics*
620 (2013) 9:1274-87. doi:10.6061/clinics/2013(09)15
- 621 13. **de Sá Junior AR, de Andrade AG, Andrade LH, Gorenstein C, Wang YP.**
622 **Response pattern of depressive symptoms among college students: What lies**
623 **behind items of the Beck Depression Inventory-II? *J Affect Disord* (2018)**
624 **234:124-30. doi:10.1016/j.jad.2018.02.064**
- 625 14. Thombs BD, Ziegelstein RC, Pilote L, Dozois DJA, Beck AT, Dobson KS, et al.

- 626 Somatic symptom overlap in Beck Depression Inventory-II scores following
627 myocardial infarction. *Br J Psychiatry* (2010) 197:61–6.
628 doi:10.1192/bjp.bp.109.076596
- 629 15. Embretson SE, Reise SP. Multivariate Applications Books Series. Item response
630 theory for psychologists. (2000). Mahwah, NJ, US: Lawrence Erlbaum
631 Associates Publishers.
- 632 16. Jones SM, Ludman EJ, McCorkle R, Reid R, Bowles EJ, Penfold R, Wagner EH.
633 A differential item function analysis of somatic symptoms of depression in
634 people with cancer. *J Affect Disord* (2015) 170:131–7. doi:
635 10.1016/j.jad.2014.09.002.
- 636 17. Hu J, Ward MM. Screening for depression in arthritis populations: an assessment
637 of differential item functioning in three self-reported questionnaires. *Qual Life*
638 *Res* (2017) 26:2507–17. doi: 10.1007/s11136-017-1601-x.
- 639 18. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome
640 measures for clinical trials in neurology: problems, solutions, and
641 recommendations. *Lancet Neurol* (2007) 12:1094–105 doi:10.1016/S1474-
642 4422(07)70290-9
- 643 19. Reeve BB. *Item and scale-level analysis of clinical and non-clinical sample*
644 *responses to the MMPI-2 depression scales employing item response theory*. The
645 University of North Carolina; Chapel Hill (2000).
- 646 20. Yang FM, Jones RN. Center for Epidemiologic Studies-Depression scale (CES-
647 D) item response bias found with Mantel-Haenszel method was successfully
648 replicated using latent variable modeling. *J Clin Epidemiol* (2007) 11:1195–200.
649 doi:10.1016/j.jclinepi.2007.02.008
- 650 21. Baker FB. *The basics of item response theory*. ERIC Clearinghouse on
651 Assessment and Evaluation. (2001). doi:10.1111/j.1365-2362.2010.02362.x.
- 652 22. Teresi JA, Fleishman JA. Differential item functioning and health assessment.
653 *Qual Life Res* (2007) 16 (Suppl. 1):33–42. doi:10.1007/s11136-007-9184-6
- 654 23. Teresi JA, Ocepok-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN,
655 et al. Analysis of differential item functioning in the depression item bank from
656 the Patient Reported Outcome Measurement Information System (PROMIS): An
657 item response theory approach. *Psychol Sci Q* (2009) 2:148–80. PMID:
658 20336180; PMCID: PMC2844669.
- 659 24. Clancy K, Gove W. Sex Differences in Mental Illness: An Analysis of Response
660 Bias in Self-Reports. *Am J Sociol* (1974) 80:205–16. doi:10.1086/225767
- 661 25. Tibblin G, Bengtsson C, Furunes B, Lapidus L. Symptoms by age and sex. The
662 population studies of men and women in Gothenburg, Sweden. *Scand J Prim*
663 *Health Care* (1990) 8:9–17.
- 664 26. Dawson RS. Depression in Children and Adolescents: The Pediatrician at the
665 Front Lines. *Pediatr Ann* (2018) 7:e261–65. doi: 10.3928/19382359-20180618-
666 01.

27. Andrade AG, Duarte PAV, Oliveira LG. 1st Nationwide Survey on the Use of Alcohol, Tobacco and Other Drugs among College Students in the 27 Brazilian State Capitals. *Brasil: SENAD* (2010)282. Available at: <http://www.justica.gov.br/central-de-conteudo/politicas-sobre-drogas/relatorios-politicas-sobre-drogas/levantamentodrogasuniversitarios-english-version.pdf/@download/file> [Accessed June 27, 2018]
28. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition. *Am Psychiatr Assoc* (1994)
29. Gomes-Oliveira MH, Gorenstein C, Neto FL, Andrade LH, Wang YP. Validation of the Brazilian Portuguese Version of the Beck Depression Inventory-II in a community sample. *Rev Bras Psiquiatr* (2012) 34:389-94. doi:10.1016/j.rbp.2012.03.005
30. Whisman MA, Richardson ED. Normative Data on the Beck Depression Inventory - Second Edition (BDI-II) in College Students. *J Clin Psychol* (2015) 71:898-907. doi:10.1002/jclp.22188
31. Storch EA, Roberti JW, Roth DA. Factor structure, concurrent validity, and internal consistency of the beck depression inventory?second edition in a sample of college students. *Depress Anxiety* (2004) 19:187-89. doi:10.1002/da.20002
32. Arnason PO, Olason DB, Benedict J, Sigurdsson JF. The Beck Depression Inventory Second Edition (BDI-II): Psychometric properties in Icelandic student and patient populations. *Nord J Psychiatry* (2008) 62:360-65. doi:10.1080/08039480801962681
33. Whisman MA, Judd CM, Whiteford NT, Gelhorn HL. Measurement Invariance of the Beck Depression Inventory-Second Edition (BDI-II) Across Gender, Race, and Ethnicity in College Students. *Assessment* (2013) 20:419-28. doi:10.1177/1073191112460273
34. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model* (1999) doi:10.1080/10705519909540118
35. Hu LT, Bentler PM. Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. *Psychol Methods* (1998) 3:424-53. doi:10.1037/1082-989X.3.4.424
36. Vandenberg RJ, Lance CE. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organ Res Methods* (2000) 3: 4-70. doi:10.1177/109442810031002
37. Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale NJ, Lawrence Erlbaum Associates.
38. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Suppl.* (1969)
39. Choi SW, Gibbons LE, Crano PK. **lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic**

- Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw* (2011) 8:1-30. doi:10.18637/jss.v039.i08
40. Lord FM. *Applications of item response theory to practical testing problems*. L. Erlbaum Associates (1980).
41. Hambleton, R. K. JRW. Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educ Meas Issues Pract* (1993) 12:38-47. doi:10.1097/01.mlr.0000245426.10853.30
42. Wenzel A, Steer RA, Beck AT. Are there any gender differences in frequency of self-reported somatic symptoms of depression? *J Affect Disord* (2005) 89:177-81. doi:10.1016/J.JAD.2005.06.009
43. Delisle VC, Beck AT, Dobson KS, Dozois DJA, Thombs BD. Revisiting Gender Differences in Somatic Symptoms of Depression: Much Ado about Nothing? *PLoS One* (2012) 7:e32490. doi:10.1371/journal.pone.0032490
44. Silverstein B, Edwards T, Gamma A, Ajdacic-Gross V, Rossler W, Angst J. The role played by depression associated with somatic symptomatology in accounting for the gender difference in the prevalence of depression. *Soc Psychiatry Psychiatr Epidemiol* (2013) 48:257-63. doi:10.1007/s00127-012-0540-7
45. Holland PW, Wainer H. *Differential item functioning*. (1993).
46. Vingerhoets AJJM, Rottenberg J, Cevaal A, Nelson JK. Is there a relationship between depression and crying? A review. *Acta Psychiatr Scand* (2007) 115:340-51. doi:10.1111/j.1600-0447.2006.00948.x
47. Bylgma LM, Grahnin A, Vingerhoets AJJM. The neurobiology of human crying. *Clin Auton Res* (2018) doi:10.1007/s10286-018-0526-y
48. Davis D, Lamberti J, Ajans ZA. Crying in depression. *Br J Psychiatry* (1969) 115:597-8.
49. Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *J Clin Epidemiol* (2000) 53:285-9. doi:10.1016/S0895-4356(99)00151-1
50. Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educ Psychol Meas* (2003) 63: 65-74. doi:10.1177/0013164402239317
51. Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item Bias in the Center for Epidemiologic Studies Depression Scale: Effects of Physical Disorders and Disability in an Elderly Community Sample. *J Gerontol B Psychol Sci Soc Sci* (2000) 55:273-82. doi:10.1093/geronb/55.5.P273
52. Pickard AS, Dalal MR, Bushnell DM. A comparison of depressive symptoms in stroke and primary care: Applying Rasch models to evaluate the Center for Epidemiologic Studies-Depression scale. *Value Heal* (2006) 9:59-64. doi:10.1111/j.1524-4733.2006.00082.x

- 749 53. Teresi JA, Golden RR. Latent structure methods for estimating item bias, item
750 validity and prevalence using cognitive and other geriatric screening measures.
751 *Alzheimer Dis Assoc Disord* (1994) 8 (Suppl. 1):S291-8
- 752 54. **Bulhões C, Ramos E, Severo M, Dias S, Barros H.** Measuring depressive
753 symptoms during adolescence: what is the role of gender? *Epidemiol Psychiatr*
754 *Sci* (2017) 28:1–11. doi:10.1017/S2045796017000312
- 755 55. Basson R, Weijmar Schultz W. Sexual sequelae of general medical disorders.
756 *Lancet* (2007) 369: 409–24. doi:10.1016/S0140-6736(07)60197-4
- 757 56. Toosi F, Rahimi C, Sajjadi S. Psychometric Properties of Beck Depression
758 Inventory-II for High School Children in Shiraz City, Iran. *Int J Sch Heal* (2017)
759 3: e41069. doi:10.5812/intjsh.41069
- 760 57. Blazer D, Burchett B, Service C, George LK. The association of age and
761 depression among the elderly: An epidemiologic exploration. *J Gerontol* (1991)
762 46: M210-5. doi:10.1093/geronj/46.6.M210
- 763 58. Drayer RA, Mulsant BH, Lenze EJ, Rollman BL, Dew MA, Kelleher K, et al.
764 Somatic symptoms of depression in elderly patients with medical comorbidities.
765 *Int J Geriatr Psychiatry* (2005) 10:973–82. doi:10.1002/gps.1389
- 766 59. **Chew-Graham C, Kovandžić M, Gask L, Burroughs H, Clarke P, Sanderson H, et**
767 **al.** Why may older people with depression not present to primary care? Messages
768 from secondary analysis of qualitative data. *Health Soc Care Community* (2012)
769 20:52–60. doi:10.1111/j.1365-2524.2011.01015.x
- 770 60. Law J, Laidlaw K, Peck D. Is depression viewed as an inevitable consequence of
771 age? The “understandability phenomenon” in older people. *Clin Gerontol* (2010)
772 33:194–209. doi:10.1080/07317111003773627
- 773 61. Glaesmer H, Riedel-Heller S, Braehler E, Spangenberg L, Lupp M. Age- and
774 gender-specific prevalence and risk factors for depressive symptoms in the
775 elderly: a population-based study. *Int Psychogeriatr* (2011) 23: 1294–300.
776 doi:10.1017/S1041610211000780
- 777 62. **Arias-de la Torre J, Vilagut G, Martín V, Molina AJ, Alonso JJ.** Prevalence of
778 major depressive disorder and association with personal and socio-economic
779 factors. Results for Spain of the European Health Interview Survey 2014–2015.
780 *Affect Disord* (2018) 239:203–207. doi: 10.1016/j.jad.2018.06.051
- 781 63. Hegeman JM, Kok RM, van der Mast RC, Giltay EJ. Phenomenology of
782 depression in older compared with younger adults: meta-analysis. *Br J Psychiatry*
783 (2012) 200:275–81. doi: 10.1192/bjp.bp.111.095950.
- 784 64. Curi M, Singer JM, Andrade DF. A model for psychiatric questionnaires with
785 embarrassing items. *Stat Methods Med Res* (2011) 20:451–70.
786 doi:10.1177/0962280210369424
- 787 65. Baldwin G, Shean GD. A taxometric study of the center for epidemiological
788 studies depression scale. *Genet Soc Gen Psychol Monogr* (2006) 132:101–28.
789 doi:10.3200/MONO.132.2.101-128

- 790 66. Jorm AF. Does old age reduce the risk of anxiety and depression? A review of
791 epidemiological studies across the adult life span. *Psychol Med* (2000) 30:11-22.
792 doi:10.1017/S0033291799001452
- 793 67. Chen L, Wang L, Qiu XH, Yang XX, Qiao ZX, Yang YJ, et al. Depression
794 among Chinese University Students: Prevalence and Socio-Demographic
795 Correlates. *PLoS One* (2013) 8:e58379. doi:10.1371/journal.pone.0058379
796

In review

797 Table 1: Psychometric characteristics of the Beck Depression Inventory-II for college
798 students ($N = 12,677$)

Item	<i>M</i>	SD	CV	α
1- Sadness	0.21	0.46	2.05	0.90
2- Pessimism	0.22	0.48	2.05	0.90
3- Past failure	0.23	0.57	2.25	0.90
4- Loss of pleasure	0.26	0.52	1.83	0.90
5- Guilty feelings	0.33	0.53	1.53	0.90
6- Punishment feelings	0.22	0.62	2.33	0.90
7- Self-dislike	0.24	0.63	2.50	0.90
8- Self-criticalness	0.56	0.72	1.14	0.90
9- Suicidal thoughts	0.08	0.32	3.57	0.90
10- Crying	0.36	0.77	1.91	0.90
11- Agitation	0.35	0.61	1.55	0.90
12- Loss of interest	0.26	0.52	1.83	0.90
13- Indecisiveness	0.41	0.78	1.61	0.90
14- Worthlessness	0.17	0.52	2.67	0.90
15- Loss of energy	0.44	0.62	1.29	0.90
16- Changes in sleep	0.66	0.75	0.98	0.90
17- Irritability	0.37	0.61	1.48	0.90
18- Changes in appetite	0.48	0.69	1.29	0.90
19- Concentration difficulty	0.46	0.72	1.38	0.90
20- Tiredness or fatigue	0.51	0.70	1.20	0.90
21- Loss of interest in sex	0.15	0.45	2.75	0.90
Total	6.99	7.56	1.08	0.91

799 *M* Mean

800 SD Standard deviation

801 CV Pearson's coefficient of variation = SD/M

802 α Cronbach's alpha coefficient of internal consistency if the item is deleted

803

804

805 Table 2: BDI-II response proportion (%) for responses categories on selected items

	Overall	Men	Women	Younger	Older
Item	N=12,677	n = 5,692	n = 6,985	n = 10,887	n = 1,790
1. Sadness					
Category 1	81.3	85.7	77.6	80.3	86.9
Category 2	17.5	13.2	20.9	18.3	12.3
Category 3	0.7	0.5	0.8	0.8	0.3
Category 4	0.6	0.5	0.6	0.6	0.5
5. Guilty feelings					
Category 1	69.9	72.4	67.9	69.3	73.9
Category 2	28.1	25.9	30.0	28.7	24.8
Category 3	1.4	1.3	1.5	1.4	1.1
Category 4	0.6	0.5	0.6	0.6	0.2
10. Crying					
Category 1	77.6	88.4	68.9	76.7	83.5
Category 2	12.1	6.4	16.7	12.6	8.8
Category 3	6.5	0.8	11.2	7.0	3.9
Category 4	3.8	4.5	3.1	3.7	3.8
15. Loss of energy					
Category 1	61.6	66.3	57.7	61.9	59.4
Category 2	33.2	30.1	35.8	32.6	36.4
Category 3	4.5	3.2	5.7	0.47	3.9
Category 4	0.7	0.5	0.9	0.8	0.3
21. Loss of interest in sex					
Category 1	88.0	91.9	84.8	89.6	78.3
Category 2	9.4	6.9	11.4	8.0	17.7
Category 3	2.0	0.9	2.9	1.9	3.0
Category 4	0.6	0.3	0.9	0.5	1.1

806 Note: In general, category 1 indicates "as usual"; category 2 "more than usual"; category 3 "all the time";
807 and category 4 "extreme distress". Although the item wording may vary according to the depressive
808 symptom, the categories are displayed in increasing level of severity.
809

810 Table 3: Estimated graded response model (GRM) parameters of the items of the BDI-
811 II ($N = 12,677$).

Item	a	(SE)	b_1	(SE)	b_2	(SE)	b_3	(SE)
1. Sadness	2.38	(0.06)	1.10	(0.05)	2.75	(0.12)	3.13	(0.15)
2. Pessimism	2.20	(0.05)	1.12	(0.05)	2.57	(0.09)	3.35	(0.16)
3. Past failure	2.04	(0.05)	1.32	(0.05)	1.98	(0.07)	3.41	(0.14)
4. Loss of pleasure	2.27	(0.05)	0.95	(0.05)	2.34	(0.08)	3.25	(0.15)
5. Guilty feelings	1.72	(0.04)	0.75	(0.03)	2.96	(0.08)	3.78	(0.13)
6. Punishment feelings	1.74	(0.05)	1.46	(0.05)	2.39	(0.07)	2.63	(0.07)
7. Self-dislike	2.86	(0.07)	1.24	(0.07)	1.60	(0.09)	2.54	(0.14)
8. Self-criticalness	1.62	(0.03)	0.25	(0.03)	1.78	(0.04)	3.24	(0.08)
9. Suicidal thoughts	2.12	(0.07)	1.89	(0.08)	3.22	(0.15)	3.60	(0.19)
10. Crying	1.52	(0.04)	1.16	(0.03)	1.93	(0.04)	2.73	(0.06)
11. Agitation	1.59	(0.04)	0.83	(0.03)	2.90	(0.06)	3.19	(0.08)
12. Loss of interest	2.73	(0.06)	0.89	(0.05)	2.26	(0.11)	2.87	(0.15)
13. Indecisiveness	2.03	(0.04)	0.78	(0.04)	1.88	(0.06)	2.12	(0.06)
14. Worthlessness	3.31	(0.09)	1.31	(0.10)	1.80	(0.12)	2.72	(0.20)
15. Loss of energy	2.35	(0.05)	0.40	(0.04)	1.96	(0.07)	3.05	(0.13)
16. Changes in sleep	1.58	(0.03)	0.03	(0.03)	1.69	(0.04)	3.04	(0.07)
17. Irritability	2.13	(0.04)	0.65	(0.04)	2.10	(0.07)	2.93	(0.11)
18. Changes in appetite	1.40	(0.03)	0.48	(0.03)	2.31	(0.04)	3.28	(0.07)
19. Concentration difficulty	2.07	(0.04)	0.57	(0.03)	1.60	(0.05)	2.84	(0.09)
20. Tiredness or fatigue	2.24	(0.04)	0.32	(0.03)	1.75	(0.06)	2.52	(0.09)
21. Loss of interest in sex	1.32	(0.04)	1.91	(0.04)	3.26	(0.07)	4.44	(0.12)

812 Note. a is the discrimination parameter; the b_i are location parameters; SE is standard error of the
813 parameter estimate.

814

815

Figure 1: The first (a) and third (c) plots represent the true score (item response theory score) of the college student sample on differential item functioning (DIF), for item #10 and item #21. The second (b) and fourth (d) plots show the item characteristic curves for items #10 and #21. The curves show the probability of endorsing a particular item.

823

824

825 Figure 2: Overall impact of differential item functioning (DIF) for gender and age on
 826 the test characteristic curves (TCC). The first (a) and third (c) plots display the TCC of

all items and the second (b) and fourth (d) plots display the TCC of specific DIF items
(i.e., item #10 for gender and item #21 for age).

In review

10. REFERÊNCIAS

1. Estadísticas sanitarias MundialEs 2011 [Internet]. [cited 2019 Jan 18]. Available from: https://www.who.int/whosis/whostat/ES_WHS2011_Full.pdf
2. Waraich P, Goldner EM, Somers JM, Hsu L. Prevalence and Incidence Studies of Mood Disorders: A Systematic Review of the Literature. *Can J Psychiatry*. 2004 Feb 23;49(2):124–38.
3. Mueller TI, Leon AC, Keller MB, Solomon DA, Endicott J, Coryell W, et al. Recurrence after recovery from major depressive disorder during 15 years of observational follow-up. *Am J Psychiatry*. 1999 Jul; 156(7):1000–6.
4. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, et al. Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. *Arch Gen Psychiatry*. 1992 Oct;49(10):809–16.
5. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The Epidemiology of Major Depressive Disorder. *JAMA*. 2003 Jun 18;289(23):3095.
6. Posternak MA, Solomon DA, Leon AC, Mueller TI, Shea MT, Endicott J, et al. The naturalistic course of unipolar major depression in the absence of somatic therapy. *J Nerv Ment Dis*. 2006 May;194(5):324–9.
7. Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *JAMA*. 1989 Aug 18;262(7):914–9.

8. Evans DL, Charney DS, Lewis L, Golden RN, Gorman JM, Krishnan KRR, et al. Mood disorders in the medically ill: scientific review and recommendations. *Biol Psychiatry*. 2005 Aug 1;58(3):175–89.
9. Penninx BW, Geerlings SW, Deeg DJ, van Eijk JT, van Tilburg W, Beekman AT. Minor and major depression and the risk of death in older persons. *Arch Gen Psychiatry*. 1999 Oct;56(10):889–95.
10. Wulsin LR, Vaillant GE, Wells VE. A systematic review of the mortality of depression. *Psychosom Med*. 1999;61(1):6–17.
11. Lloyd KR, Jenkins R, Mann A. Long-term outcome of patients with neurotic illness in general practice. *BMJ*. 1996 Jul 6;313(7048):26–8.
12. Johnson J, Weissman MM, Klerman GL. Service utilization and social morbidity associated with depressive symptoms in the community. *JAMA*. 1992 Mar 18;267(11):1478–83.
13. Andrade L, Walters EE, Gentil V, Laurenti R. Prevalence of ICD-10 mental disorders in a catchment area in the city of São Paulo, Brazil. *Soc Psychiatry Psychiatr Epidemiol*. 2002 Jul 8;37(7):316–25.
14. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, de Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med*. 2011 Jul 26;9:90.
15. Hardeveld F, Spijker J, De Graaf R, Nolen WA, Beekman ATF. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr Scand*. 2009 Dec 11;122(3):184–91.
16. Torpey DC, Klein DN. Chronic depression: update on classification and

- treatment. *Curr Psychiatry Rep.* 2008 Dec;10(6):458–64.
17. Möller HJ. Suicide, suicidality and suicide prevention in affective disorders. *Acta Psychiatr Scand Suppl.* 2003;(418):73–80.
 18. Patten SB, Williams JVA, Lavorato DH, Bulloch AGM, D'Arcy C, Streiner DL. Recall of recent and more remote depressive episodes in a prospective cohort study. *Soc Psychiatry Psychiatr Epidemiol.* 2012 May 1;47(5):691–6.
 19. Colman I, Ataullahjan A. Life Course Perspectives on the Epidemiology of Depression. *Can J Psychiatry.* 2010 Oct;55(10):622–32.
 20. Kuehner C. Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. *Acta Psychiatr Scand.* 2003 Sep;108(3):163–74.
 21. Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. *Br J Psychiatry.* 2000 Dec;177:486–92.
 22. Klose M, Jacobi F. Can gender differences in the prevalence of mental disorders be explained by sociodemographic factors? *Arch Womens Ment Health.* 2004 Apr 1;7(2):133–48.
 23. Kendler KS, Gardner CO, Prescott CA. Toward a comprehensive developmental model for major depression in women. *Am J Psychiatry.* 2002 Jul;159(7):1133–45.
 24. Nolen-Hoeksema S, Larson J, Grayson C. Explaining the gender difference in depressive symptoms. *J Pers Soc Psychol.* 1999 Nov;77(5):1061–72.
 25. Kendler KS, Gardner CO, Prescott CA. Toward a comprehensive developmental model for major depression in men. *Am J Psychiatry.* 2006

- Jan;163(1):115–24.
26. Hammarström A, Lehti A, Danielsson U, Bengs C, Johansson EE. Gender-related explanatory models of depression: a critical evaluation of medical articles. *Public Health*. 2009 Oct;123(10):689–93.
 27. Kendler KS, Thornton LM, Prescott CA. Gender differences in the rates of exposure to stressful life events and sensitivity to their depressogenic effects. *Am J Psychiatry*. 2001 Apr;158(4):587–93.
 28. Pearlin LI. The sociological study of stress. *J Health Soc Behav*. 1989 Sep;30(3):241–56.
 29. Frank E, Carpenter LL, Kupfer DJ. Sex differences in recurrent depression: are there any that are significant? *Am J Psychiatry*. 1988 Jan;145(1):41–5.
 30. Young MA, Scheftner WA, Fawcett J, Klerman GL. Gender differences in the clinical features of unipolar major depressive disorder. *J Nerv Ment Dis*. 1990 Mar;178(3):200–3.
 31. Silverstein B. Gender difference in the prevalence of clinical depression: the role played by depression associated with somatic symptoms. *Am J Psychiatry*. 1999 Mar;156(3):480–2.
 32. Hirschfeld RMA, Russell JM. Assessment and Treatment of Suicidal Patients. *N Engl J Med*. 1997 Sep 25;337(13):910–5.
 33. Rochlen AB, Paterniti DA, Epstein RM, Duberstein P, Willeford L, Kravitz RL. Barriers in Diagnosing and Treating Men With Depression: A Focus Group Report. *Am J Mens Health*. 2010 Jun 11;4(2):167–75.
 34. Martin LA, Neighbors HW, Griffith DM. The experience of symptoms of

- depression in men vs women: analysis of the National Comorbidity Survey Replication. *JAMA psychiatry*. 2013 Oct 1;70(10):1100–6.
35. Williams JW, Noël PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA*. 2002 Mar 6;287(9):1160–70.
 36. Dunn G, Pickles A, Tansella M, Vázquez-Barquero JL. Two-phase epidemiological surveys in psychiatric research. *Br J Psychiatry*. 1999 Feb;174:95–100.
 37. Pignone M, Gaynes BN, Rushton JL, Mulrow CD, Orleans CT, Whitener BL, et al. Screening for Depression. Rockville (MD): Agency for Healthcare Research and Quality (US); 2002 Apr. Report No.: 02-S002.
 38. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry*. 1961 Jun;4:561–71.
 39. Rush AJ. At the Research Front: Cognitive Therapy for Depression. *Aust New Zeal J Psychiatry*. 1979 Mar 26;13(1):13–6.
 40. Gomes-Oliveira MH, Gorenstein C, Lotufo Neto F, Andrade LH, Wang YP. Validation of the Brazilian Portuguese version of the Beck Depression Inventory-II in a community sample. *Rev Bras Psiquiatr*. 2012 Dec;34(4):389–94.
 41. Nunes CHS da S, Primi R. Impacto do tamanho da amostra na calibração de itens e estimativa de escores por teoria de resposta ao item. Vol. 4, *Avaliação Psicológica*. 2005. 141-53.
 42. Doucette A, Wolf AW. Questioning the measurement precision of psychotherapy research. *Psychother Res*. 2009 Jul 22;19(4–5):374–89.

43. Meijer RR, Baneke JJ. Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychol Methods*. 2004 Sep;9(3):354–68.
44. Pasquali L. *Psicometria : teoria dos testes na psicologia e na educação*. Vozes; 2003.
45. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*. 1970.
46. Emons WHM, Meijer RR, Denollet J. Negative affectivity and social inhibition in cardiovascular disease: evaluating type-D personality and its assessment using item response theory. *J Psychosom Res*. 2007 Jul;63(1):27–39.
47. Purpura DJ, Wilson SB, Lonigan CJ. Attention-deficit/hyperactivity disorder symptoms in preschool children: examining psychometric properties using item response theory. *Psychol Assess*. 2010 Sep;22(3):546–58.
48. Cohen AS, Kim S-H, Baker FB. Detection of Differential Item Functioning in the Graded Response Model. *Appl Psychol Meas*. 1993 Dec 26;17(4):335–50.
49. Downing SM, Haladyna TM. Test Item Development: Validity Evidence From Quality Assurance Procedures. *Appl Meas Educ*. 1997 Jan;10(1):61–82.
50. Mellenbergh GJ. Item bias and item response theory. *Int J Educ Res*. 1989 Jan 1;13(2):127–43.
51. Chang C-H, Reeve BB. Item Response Theory and its Applications to Patient-Reported Outcomes Measurement. *Eval Health Prof*. 2005 Sep 15;28(3):264–82.
52. Andriola WB. Descrição dos Principais Métodos para Detectar o

- Funcionamento Diferencial dos Itens (DIF). *Psicol Reflexão e Crítica*. 2001;14(3):643–52.
53. Kim S-H, Cohen AS. Detection of Differential Item Functioning Under the Graded Response Model With the Likelihood Ratio Test. *Appl Psychol Meas*. 1998 Dec 27;22(4):345–55.
 54. Mazor KM, Hambleton RK, Clauser BE. Multidimensional DIF Analyses: The Effects of Matching on Unidimensional Subtest Scores. *Appl Psychol Meas*. 1998 Dec 27;22(4):357–67.
 55. Camilli G, Shepard LA. Methods for identifying biased test items. Sage Publications; 1994. 174.
 56. Verkuilen J. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach by P. de Boeck and M. Wilson and Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models by A. Skrondal and S. Rabe-Hesketh. *Psychometrika*. 2006 Jun 11;71(2):415–8.
 57. Woods CM. Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behav Res*. 2009 Feb 24;44(1):1–27.
 58. Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. 1993.
 59. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic

- Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw.* 2011 Mar 1;39(8):1-30.
60. Swaminathan H, Rogers HJ. Detecting Differential Item Functioning Using Logistic Regression Procedures. *J Educ Meas.* 1990 Dec 1;27(4):361–70.
 61. Zumbo BD, Wild WR. A Handbook on the Theory and Methods of Differential Item Functioning (DIF) Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores. 1999.
 62. Jodoin MG, Gierl MJ. Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Appl Meas Educ.* 2001 Oct;14(4):329–49.
 63. Kim S-H, Cohen AS, Alagoz C, Kim S. DIF Detection and Effect Size Measures for Polytomously Scored Items. *J Educ Meas.* 2007 Jul 1;44(2):93–116.
 64. Crane PK, Van Belle G, Larson EB, Crane PK, Van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. *Stat Med Stat Med.* 2004;23:241–56.
 65. Brouwer D, Meijer RR, Zevalkink J. Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychother Res.* 2013 Sep;23(5):489–501.
 66. Andrade AG, Duarte PCAV, Oliveira LG. 1st Nationwide Survey on the Use of Alcohol, Tobacco and Other Drugs among College Students in the 27 Brazilian State Capitals. Brasília: SENAD. 2010;282.
 67. Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the Validity of the Beck

- Depression Inventory. *Psychopathology*. 1998;31(3):160–8.
68. Beck A, Steer R, Brown G. Beck Depression Inventory II manual. Beck Depression Inventory-II. 1996.
 69. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model. A Multidisciplinary Journal*. 1999; 6:1, 1-55.
 70. Hu L, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychol Methods*. 1998;3(4):424–53.
 71. Vandenberg RJ, Lance CE. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organ Behav Hum Decis Process*. 2000 Sep;83(1):107-140.
 72. Baker F. Item Response Theory. *Eric.Ed.Gov*. 2001
 73. Hu J, Ward MM. Screening for depression in arthritis populations: an assessment of differential item functioning in three self-reported questionnaires. *Qual Life Res*. 2017 Sep 17;26(9):2507–17.
 74. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol*. 2007 Dec;6(12):1094-105
 75. Yang FM, Jones RN. Center for Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *J Clin Epidemiol*. 2007

- Nov;60(11):1195-200.
76. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res.* 2007;
 77. de Sá Junior AR, de Andrade AG, Andrade LH, Gorenstein C, Wang YP. Response pattern of depressive symptoms among college students: What lies behind items of the Beck Depression Inventory-II? *J Affect Disord.* 2018 Jul;234:124-130.
 78. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educ Meas Issues Pract.* 1993; 12:38–47.
 79. Lord FM. Applications of item response theory to practical testing problems. L. Erlbaum Associates; 1980. 274 p.
 80. Delisle VC, Beck AT, Dobson KS, Dozois DJA, Thombs BD. Revisiting Gender Differences in Somatic Symptoms of Depression: Much Ado about Nothing? Kemp AH, editor. *PLoS One.* 2012 Feb 24;7(2):e32490.
 81. Wenzel A, Steer RA, Beck AT. Are there any gender differences in frequency of self-reported somatic symptoms of depression? *J Affect Disord.* 2005 Dec 1;89(1–3):177–81.
 82. Silverstein B, Edwards T, Gamma A, Ajdacic-Gross V, Rossler W, Angst J. The role played by depression associated with somatic symptomatology in accounting for the gender difference in the prevalence of depression. *Soc Psychiatry Psychiatr Epidemiol.* 2013 Feb 3;48(2):257–63.
 83. Holland P, Wainer H. Differential item functioning. Differential item

- functioning. 2012.
84. Bylsma LM, Gračanin A, Vingerhoets AJJM. The neurobiology of human crying. *Clin Auton Res*. 2018 Apr 23;
 85. Vingerhoets AJJM, Rottenberg J, Cevaal A, Nelson JK. Is there a relationship between depression and crying? A review. *Acta Psychiatr Scand*. 2007 May;115(5):340–51.
 86. Davis D, Lamberti J, Ajans ZA. Crying in depression. *Br J Psychiatry*. 1969 May;115(522):597–8.
 87. Reeve BB. Item and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory. The University of North Carolina; Chapel Hill; 2000.
 88. Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *J Clin Epidemiol*. 2000 Mar 1;53(3):285-9.
 89. Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educ Psychol Meas*. 2003; 63: 65-74.
 90. Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item Bias in the Center for Epidemiologic Studies Depression Scale: Effects of Physical Disorders and Disability in an Elderly Community Sample. *J Gerontol B Psychol Sci Soc Sci*. 2000 Sep;55(5):273-82.
 91. Pickard AS, Dalal MR, Bushnell DM. A comparison of depressive symptoms in stroke and primary care: Applying Rasch models to evaluate the Center for

- Epidemiologic Studies-Depression scale. *Value Heal.* 2006;9(1):59–64.
92. Teresi J, Golden R. Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. *Alzheimer Dis Assoc Disord.* 1994;8 Suppl 1:291–8.
 93. Bulhões C, Ramos E, Severo M, Dias S, Barros H. Measuring depressive symptoms during adolescence: what is the role of gender? *Epidemiol Psychiatr Sci.* 2017 Jun 28;1–11.
 94. Tibblin G, Bengtsson C, Furunes B, Lapidus L. Symptoms by age and sex. The population studies of men and women in Gothenburg, Sweden. *Scand J Prim Health Care.* 1990 Mar;8(1):9–17.
 95. Clancy K, Gove W. Sex Differences in Mental Illness: An Analysis of Response Bias in Self-Reports. *Am J Sociol.* 1974; 80:205-16.
 96. Dekker J, Koelen JA, Peen J, Schoevers RA, Gijsbers-Van Wijk C. Gender Differences in Clinical Features of Depressed Outpatients: Preliminary Evidence for Subtyping of Depression? *Women Health.* 2008 Feb 19;46(4):19–38.
 97. Ernst C, Angst J. The Zurich Study. *Eur Arch Psychiatry Clin Neurosci.* 1992 Mar;241(4):222–30.
 98. Kockler M, Heun R. Gender differences of depressive symptoms in depressed and nondepressed elderly persons. *Int J Geriatr Psychiatry.* 2002 Jan 1;17(1):65–72.
 99. Kornstein SG, Schatzberg AF, Thase ME, Yonkers KA, McCullough JP, Keitner GI, et al. Gender differences in chronic major and double depression.

- J Affect Disord. 2000 Oct;60(1):1–11.
100. Marcus SM, Kerber KB, Rush AJ, Wisniewski SR, Nierenberg A, Balasubramani GK, et al. Sex differences in depression symptoms in treatment-seeking adults: confirmatory analyses from the Sequenced Treatment Alternatives to Relieve Depression study. *Compr Psychiatry*. 2008;49(3):238–46.
 101. Middeldorp CM, Wray NR, Andrews G, Martin NG, Boomsma DI. Sex differences in symptoms of depression in unrelated individuals and opposite-sex twin and sibling pairs. *Twin Res Hum Genet*. 2006 Oct 1;9(5):632–6.
 102. Romans SE, Tyas J, Cohen MM, Silverstone T. Gender Differences in the Symptoms of Major Depressive Disorder. *J Nerv Ment Dis*. 2007 Nov;195(11):905–11.
 103. Uebelacker LA, Strong D, Weinstock LM, Miller IW. Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychol Med*. 2009 Apr 30;39(4):591–601.
 104. Khan AA, Gardner CO, Prescott CA, Kendler KS. Gender Differences in the Symptoms of Major Depression in Opposite-Sex Dizygotic Twin Pairs. *Am J Psychiatry*. 2002 Aug 1;159(8):1427–9.
 105. Breslau N, Schultz L, Peterson E. Sex differences in depression: a role for preexisting anxiety. *Psychiatry Res*. 1995 Sep 8;58(1):1–12.
 106. Young MA, Fogg LF, Scheftner WA, Keller MB, Fawcett JA. Sex differences in the lifetime prevalence of depression: does varying the diagnostic criteria

- reduce the female/male ratio? *J Affect Disord.* 1990 Mar;18(3):187–92.
107. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychol Sci Q.* 2009;51(2):148-180.
 108. Baldwin G, Shean GD. A taxometric study of the center for epidemiological studies depression scale. *Genet Soc Gen Psychol Monogr.* 2006 May;132(2):101-28.
 109. Jorm AF. Does old age reduce the risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychol Med.* 2000 Jan;30(1):11-22.
 110. Bayram N, Bilgel N. The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Soc Psychiatry Psychiatr Epidemiol.* 2008 Aug;43(8):667-72
 111. Chen L, Wang L, Qiu XH, Yang XX, Qiao ZX, Yang YJ, et al. Depression among Chinese University Students: Prevalence and Socio-Demographic Correlates. Mitchell AJ, editor. *PLoS One.* 2013 Mar 13;8(3):e58379.

11. APÊNDICES

Apêndice A

Artigo 1

Psychological and Cognitive Profile of Hypertensive and Diabetic Patients

Psychological and Cognitive Profile of Hypertensive and Diabetic Patients

Robson Bonoto Teixeira, MSc,* João Carlos Bouzas Martins, PhD, DSc,* Antônio Reis de Sá-Junior, MD,*
Cristiane Junqueira de Carvalho, MD,* Carlos Gabriel Lade, MSc,* Albert A. Rizvanov, PhD, DSc,†
Andrey P. Kiyasov, MD, PhD, DSc,‡ Marat A. Mukhamedyarov, MD, PhD,‡
András Palotás, MD, PhD,‡§ and Luciana Moreira Lima, PharmD, PhD*

Abstract: Chronic disorders such as hypertension and diabetes mellitus are often associated with depressive and anxiety symptoms, as well as cognitive decline. Once developed, psychological support is essential for improving the quality of life. This study is aimed at identifying impaired mental health in connection with these systemic metabolic disorders. A total of 34 patients were included in this cross-sectional study: 17 hypertensive individuals with a mean age of 59 ± 10 years, and 17 diabetic patients aged 54 ± 10 years. The following psychometric tests were used: Mini-Mental State Examination (MMSE), Beck Depression Inventory (BDI), Beck Anxiety Inventory (BAI), and self-reporting questionnaire (SRQ-20). A large number of patients with high blood pressure or diabetes was associated with mental health problems (82% or 65%, respectively; $p = 0.246$). Affective disorder, especially moderate to severe depression, was seen mainly in diabetic patients (76%), whereas hypertensive individuals had higher prevalence of anxiety (64%). There was no cognitive impairment in this middle-aged population. This study shows a high proportion of depression and anxiety symptoms in patients with hypertension or diabetes mellitus, reinforcing the importance of psychiatric support for appropriate control of these metabolic disorders.

Key Words: Anxiety, cognitive function, depression, diabetes, hypertension, mental health

(J Nerv Ment Dis 2015;203: 781–785)

There is evidence that high levels of systolic blood pressure and diastolic blood pressure are significantly related to cognitive decline, both in younger people as well as the elderly, and young adults with very high blood pressure have increased risks of dementia in old age (Nimmiya et al., 2011).

The association between systemic hypertension (SH), anxiety, and depression is not yet fully established in literature. However, neurofibrillary tangles and brain atrophy have been evidenced in autopsy of patients with this clinical condition, which can also be responsible for increased production of free radicals linked to the mechanisms of cerebrovascular damage (Reitz et al., 2007).

Hypertensive individuals are at risk of developing vascular dementia 3 times higher than the normotensive ones, and this proportion increases to 6 times with the presence of SH associated with diabetes mellitus (DM) (Posner et al., 2002). In diabetic patients, prevalence of anxiety and depression is greater than in the general population (Komsuoglu Celikyurt et al., 2014). Innumerable changes related to

DM can lead to mental disorders, including cerebral microvascular disease, the effects of hyperglycemia on the circuits and neural networks, hyperinsulinism, functional deficit of insulin in the brain, increased oxidative stress, and the characteristic DM dyslipidemia. The presence of cardiovascular disease, kidney failure, obesity, metabolic syndrome, and sedentary lifestyle can also contribute to mental disorders in diabetic patients (Strachan, 2011).

Memory loss is one of the most frequently reported cognitive deficits in diabetic patients, causing inappropriate behavior (Palta et al., 2014), which can interfere with self-care during treatment and promote the permanence of hyperglycemia. The hyperglycemic state can cause cognitive deficits leading to macrovascular and microvascular diseases in the brain, such as stroke and white matter lesions, associated with neurodegeneration markers, leading to atrophy of the medial lobe. Additionally, follow-up studies have shown that DM is associated with accelerated development of cerebral atrophy (Qiu et al., 2014). On the other hand, anxiety and depression may also contribute to weight gain (De Wit et al., 2010) and SH and/or DM development, which is one more reason for the counseling of hypertensive and diabetic patients.

This study aimed to describe the mental, cognitive decline as well as depression and anxiety states in hypertensive and diabetic patients treated at a center of secondary care of a city in Minas Gerais, candidates for participation in physical training program.

METHODS

Study Design

This is a descriptive, cross-sectional study developed as part of the "Global assessment of cardiovascular risk and mental status of patients assisted by Vçosa Hiperdia Center after exercise program," a partnership between the Federal University of Vçosa with the Hiperdia Minas Program. The study was fully performed on the premises of Hiperdia Center in Vçosa, MG. The study was approved by the Ethics in Research Committee of the Federal University of Vçosa and registered in Brazil Platform with number 33979214.3.0000.5153 protocol number 832.149/2014.

Sample

The study included a total of 34 patients, 17 patients with type 2 DM and 17 patients with resistant SH, characterized by high blood pressure that remains above the levels considered ideal, despite the concomitant use of 3 different classes of antihypertensive drugs, one needing to be diuretic in their correct dosages (Daugherty et al., 2012).

All volunteers were eligible to participate in the supervised exercise program Hiperdia Center in Vçosa, MG, for the registration and monitoring of hypertensive and/or diabetic patients assisted by the Unified Health System that meet the SES Resolution no. 2606 of December 7, 2010, which set out the guiding criteria for Integrated Reference Centers Viva Vida and Hiperdia Minas. The following exclusion criteria were considered: nonparticipation in the supervised exercise

*Federal University of Vçosa, Vçosa, Minas Gerais, Brazil; †Kazan Federal University and ‡Kazan State Medical University, Kazan, Russia; and §Aldopio-Med (private medical practice and research center), Szeged, Hungary.

Send reprint requests to Luciana Moreira Lima, PharmD, PhD, Departamento de Medicina e Enfermagem, Universidade Federal de Vçosa, Av. Peter Henry Rolfs, Vçosa, Minas Gerais, BR-36570-000, Brazil. E-mail: luciana.lima@ufv.br; and András Palotás, MD, PhD, Aldopio-Med, Kossuth Lajos sq. 23, H-6722 Szeged, Hungary. E-mail: palotas@aldopio-med.eu

Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 0022-3018/15/203 10–0781

DOI: 10.1097/NMD.0000000000000367

Apêndice B

Artigo 2

Improved cognitive, affective and anxiety measures in patients with chronic systemic disorders following structured physical activity

Improved cognitive, affective and anxiety measures in patients with chronic systemic disorders following structured physical activity

Diabetes & Vascular Disease Research
1–10
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1479164115602651
dvr.sagepub.com

Robson Bonoto Teixeira¹, João Carlos Bouzas Marins¹,
Antonio Reis de Sá Junior¹, Cristiane Junqueira de Carvalho¹,
Tiago Augusto da Silva Moura¹, Carlos Gabriel Lade¹,
Albert A Rizvanov², Andrey P Kiyasov²,
Marat A Mukhamedyarov³, Andrey L Zefirov³, András Palotás^{2,4}
and Luciana Moreira Lima¹

Abstract

Mental illnesses are frequent co-morbid conditions in chronic systemic diseases. High incidences of depression, anxiety and cognitive impairment complicate cardiovascular and metabolic disorders such as hypertension and diabetes mellitus. Lifestyle changes including regular exercise have been advocated to reduce blood pressure and improve glycaemic control. The purpose of this project was to evaluate the effect of physical training on the most prevalent corollary psychiatric problems in patients with chronic organic ailments. This longitudinal study assessed the mental health of hypertensive (age: 57 ± 8 years) and/or diabetic (age: 53 ± 8 years) patients using minimal state examination, Beck's depression inventory, Beck's anxiety inventory and self-reporting questionnaire-20 before and after a 3-month supervised resistance and aerobic exercise programme comprising structured physical activity three times a week. Clinically relevant improvement was observed in the Beck's depression inventory and Beck's anxiety inventory scores following the 12-week training (61%, $p = 0.001$, and 53%, $p = 0.02$, respectively). Even though statistically not significant ($p = 0.398$), the cognitive performance of this relatively young patient population also benefited from the programme. These results demonstrate positive effects of active lifestyle on non-psychotic mental disorders in patients with chronic systemic diseases, recommending exercise as an alternative treatment option.

Keywords

Anxiety, chronic disease, cognitive dysfunction, depression, diabetes mellitus, hypertension, mental health, physical exercise

Introduction

Regular physical activity has been considered as one of the mainstream in treating systemic arterial hypertension (SAH) and diabetes mellitus (DM).^{1,2} Benefits such as better glycaemic and cardiorespiratory control,³ increased insulin sensitivity,⁴ improved lipid profile⁵ and weight loss are well characterized.⁶ Acute⁷ and chronic⁸ blood pressure lowering has also been linked to an active lifestyle.

In addition to physical and cardio-metabolic benefits, regular physical exercise can result in decreased depression and anxiety symptoms,⁹ which may influence cognitive

functions, probably by increased brain oxygenation due to a greater blood flow and perfusion, which directly influences cognition, regardless of age.¹⁰

¹Federal University of Viçosa, Viçosa, Brazil

²Kazan Federal University, Kazan, Russia

³Kazan State Medical University, Kazan, Russia

⁴Asklepios-Med (Private Medical Practice and Research Center), Szeged, Hungary

Corresponding author:

András Palotás, Asklepios-Med, Kossuth Lajos sgt. 23, H-6722 Szeged, Hungary.

Email: palotas@asklepios-med.eu

Apêndice C

Artigo 3

Uso de drogas entre estudantes universitários: uma perspectiva nacional

Uso de drogas entre estudantes universitários: uma perspectiva nacional*Drug use among undergraduate students: a national perspective*Bianca Pereira de Assis Trindade¹, Alessandra Vieira Diniz¹, Antonio Reis Sá- Júnior²**Resumo**

O objetivo do presente estudo foi avaliar fatores associados ao uso de álcool, tabaco e outras drogas entre universitários no Brasil. Os resultados mostraram taxas mais altas de uso ao longo do tempo para álcool, tabaco e maconha entre as substâncias disponíveis. As evidências encontradas sugerem ainda que o abuso de substâncias está associado com prejuízo no desempenho acadêmico. Esta revisão da literatura médica identificou que o problema do abuso de drogas entre os estudantes universitários no Brasil é um problema reconhecido. Com isso, os riscos do uso de drogas lícitas e ilícitas devem ser levados em consideração no desenvolvimento de estratégias de prevenção e tratamento para estudantes universitários.

Palavras-chave: Transtornos Relacionados ao Uso de substâncias; Prevalência; Cognição; Alunos de Graduação.

Abstract

The aim of the present study was to assess the factors associated with use of alcohol, tobacco and other drugs by undergraduate students. Results showed highest lifetime use rates for alcohol, tobacco and marijuana among available substances. The evidence suggests that substance abuse is associated with impairment in academic performance. This review of the medical literature identified that the problem of drug abuse among university students in Brazil is a recognized problem. The risks in licit and illicit drug use should be taken into account in the development of preventive and treatment strategies for undergraduate students.

Keyword: Substance-Related Disorders; Prevalence; Cognition; Undergraduate Students.

1. Acadêmicos do curso de Medicina da Universidade Federal de Viçosa

2. Mestre e Doutorando pela Faculdade de Medicina da Universidade de São Paulo (FMUSP). Professor na Universidade Federal de Santa Catarina (UFSC)

E-mail do primeiro autor: biancapereira@gmail.com

Recebido em 31/10/2017

Aceito, após revisão, em 05/03/2018

Apêndice D

Artigo 4

Aprendizagem Baseada em Equipes e Aprendizagem Baseada em Projetos no Curso de Fisioterapia: Estudo Piloto

APRENDIZAGEM BASEADA EM EQUIPES E APRENDIZAGEM BASEADA EM PROJETOS NO CURSO DE FISIOTERAPIA: ESTUDO PILOTO

TEAM-BASED LEARNING AND PROJECT-BASED LEARNING IN PHYSIOTHERAPY COURSE: PILOT STUDY

Melissa Negro-Delacqua¹, Franciele Vaneza Costa², Iana Francochet de Sousa³,
Antonio Reis de Sá Junior⁴, Graziela Lieber⁵

¹ Farmacêutica. Doutora em Farmacologia de Produtos Naturais e Sintéticos Bioativos pela UFPB.
Professora do Departamento de Ciências da Saúde da UFSC, Araranguá, SC, Brasil.

² Farmacêutica. Mestre em Ciências Ambientais pela UNESB.

Técnica de Laboratório/Biologia da UFSC, Araranguá, SC, Brasil.

³ Farmacêutica. Doutora em Saúde e Desenvolvimento pela UFMS.

Professora do Departamento de Ciências da Saúde da UFSC, Araranguá, SC, Brasil.

⁴ Médico. Mestre em Psiquiatria pela Faculdade de Medicina da USP.

Professor do Departamento de Ciências da Saúde da UFSC, Araranguá, SC, Brasil.

⁵ Fonoaudióloga. Mestre em Saúde Coletiva pela UFJF.

Doutoranda em Saúde Coletiva pela UFJF. Doutoranda em Saúde Coletiva pela UFJF.

Juiz de Fora, MG, Brasil.

Contato: Melissa Negro-Delacqua, Universidade Federal de Santa Catarina, Rod. Gov. Jorge Lacerda,
3201, Unidade Jardim das Avenidas, Araranguá, Brasil. E-mail: melissanegrodelacqua@gmail.com

RESUMO

Introdução: A Aprendizagem Baseada em Equipes (ABE) e a Aprendizagem Baseada em Projetos (ABP) são métodos dinâmicos, que trabalham a autonomia do estudante e incentivam o trabalho em grupo. **Objetivos:** Analisar algumas percepções e a autoavaliação dos estudantes no emprego de metodologias ativas de ensino-aprendizagem em um curso tradicional de Fisioterapia. **Metodologia:** a pesquisa foi realizada a partir da aplicação aos dois primeiros níveis do modelo de Kirkpatrick, utilizando-se ferramentas como questionários de

autoavaliação. **Resultados:** 83 alunos participaram da pesquisa. 44,6% participaram da ABE e 55,4% da ABP. Foram encontrados percentuais elevados de aprovação em relação à satisfação, à importância das metodologias para a compreensão dos conteúdos, ao aumento do conhecimento sobre as metodologias ativas e ao aumento da capacidade de trabalho em grupos. **Conclusão:** Os alunos aprovam a utilização dos métodos ativos, mostrando que é possível a utilização dos mesmos em cursos estruturados de forma disciplinar.