Psychological Assessment

Beyond frequency: Evaluating the validity of assessing the context, duration, ability and botherment of depression and anxiety symptoms in south Brazil

--Manuscript Draft--

Manuscript Number:	PAS-2023-1767R1
Full Title:	Beyond frequency: Evaluating the validity of assessing the context, duration, ability and botherment of depression and anxiety symptoms in south Brazil
Abstract:	Assessment tools for depression and anxiety usually inquire about frequency of symptoms. However, evidence suggests that different question framings might trigger different responses. Our aim is to test if asking about symptom's context, ability, duration and botherement adds validity to PHQ-9, GAD-7 and PROMIS depression and anxiety. Participants came from two cross-sectional convenience-sampled surveys (N=1,871) of adults (66% females, aged 33.4 ± 13.2), weighted to approximate with the state-level population. We examined measurement invariance across the different question frames, estimated whether framing affected mean scores, and tested their independent validity using covariate-adjusted and sample-weighted structural equation models. Validity was tested using tools assessing general disability, alcohol use, loneliness, well-being, grit, and frequency-based questions from depression and anxiety questionnaires. A bifactor model was applied to test the internal consistency of the question-frames under the presence of a general factor (i.e., depression or anxiety). Measurement invariance was supported across the different frames. Framing questions as ability (i.e., "how easily…") produced a higher score, compared with framing by context (i.e., "in which daily situations…"). Construct and criterion validity analysis demonstrate that variance explained using multiple question frames were similar to using only one. We detected a strong overarching factor for each instrument, with little variances left to be explained by the question frame. Therefore, it is unlikely that using different adverbial phrasings can help clinicians and researchers to improve their ability to detect depression or anxiety.
Article Type:	Article
Keywords:	PHQ-9; GAD-7; PROMIS; mental health questionnaire; question frame
Manuscript Classifications:	adults; anxiety; depression; psychometrics; Latinos
Funding Information:	Universidade Federal de Santa Maria (FIPE/UFSM 2022) — Prof. Mauricio Scopel Hoffmann Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (22/2551-0000764-3) — Prof. Mauricio Scopel Hoffmann Conselho Nacional de Desenvolvimento Científico e Tecnológico (21/2022) — Prof. Mauricio Scopel Hoffmann
Corresponding Author:	Mauricio Scopel Hoffmann, Ph.D, MD Universidade Federal de Santa Maria Santa Maria, RS BRAZIL
Corresponding Author E-Mail:	mauricio.hoffmann@ufsm.br
Corresponding Author's Institution:	Universidade Federal de Santa Maria
Corresponding Author's Secondary Institution:	
Order of Authors (with Contributor Roles):	Reza Brümmer (Conceptualization: Supporting; Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Project administration: Equal; Visualization: Lead; Writing – original draft: Equal; Writing – review & editing: Supporting) Karolin Rose Krause, PhD (Conceptualization: Equal; Writing – review & editing: Equal) Giovanni Abrahão Salum, MD, PhD (Conceptualization: Supporting; Methodology: Supporting; Writing – review & editing: Equal) Marcelo Fleck, MD, PhD (Conceptualization: Supporting; Methodology: Supporting)

Beyond frequency: Evaluating the validity of assessing the context, duration, ability and botherement of depression and anxiety symptoms in south Brazil

Reza de Souza Brümmer[1], Karolin Rose Krause[2], Giovanni Abrahão Salum[3,4,5], Marcelo Pio de Almeida Fleck[3,4], Ighor Miron Porto[1,6], João Villanova do Amaral[1,3,6], João Pedro Gonçalves Pacheco[1,3,6], Bettina Moltrecht[7,8], Eoin McElroy[8,9], Mauricio Scopel Hoffmann[1,3,6]

[1] Department of Neuropsychiatry, Universidade Federal de Santa Maria, Avenida Roraima 1000, building 26, office 1446, Santa Maria, 97105-900, Brazil (UFSM), phone +55-55-3220-8427.

[2] Cundill Centre for Child and Youth Depression, Centre for Addiction and Mental Health, Toronto, ON, Canada.

[3] Graduation program in Psychiatry and Behavioral Sciences, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

[4] Department of Psychiatry and Legal Medicine, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos 2350, Porto Alegre, 90035-003, Brazil

[5] Child Mind Institute, New York, NY 10022, USA

[6] Mental Health Epidemiology Group (MHEG), Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil.

[7] Evidence-Based Practice Unit, Anna Freud National Centre for Children and Families, 4–8 Rodney Street, London, N1 9JH, UK

[8] Centre for Longitudinal Studies, University College London, 55-59 Gordon Square, London, WC1H 0NU, UK

[9] School of Psychology, Ulster University, Derry, Northern Ireland, UK

Corresponding author: Dr. Mauricio Scopel Hoffmann. Universidade Federal de Santa Maria, Avenida Roraima 1000, building 26, office 1446, Santa Maria, 97105-900, Brazil (UFSM), phone +55-55-3220-8427; e-mail: mauricio.hoffmann@ufsm.br.

Conflict of Interest Statement

KRK received fees from the International Consortium for Health Outcomes Measurement (ICHOM) for work as a research fellow on a project that aimed to develop a standard set of outcomes for child and youth anxiety and depression (October 2018 through March 2020). She is a member of the International Alliance of Mental Health Research Funders' (IAMHRF) Common Measures Advisory Board and provides advice on the establishment of common measures in mental health research, including the PHQ-9 and GAD-7.

RSB was supported by FIPE/UFSM 2022 Júnior scholarship, granted to MSH. IMP was supported by PROBIC/FAPERGS 2022 scholarship (22/2551-0000764-3) and is supported by PIBIC/CNPq 2023 scholarship (application call 21/2022 CNPq) at Universidade Federal de Santa Maria, granted to MSH.

MSH is supported by the United States National Institutes of Health grant R01MH120482 under his post-doctoral fellowship at UFRGS and by the Wellcome Mental Health Data Prize, granted by the Wellcome Trust (grant number 84494R). During the development of this project, he was supported by the Newton International Fellowship (Ref: NIF\R1\181942), awarded by the Academy of Medical Sciences through the UK Government's Newton Fund Programme.

Contribution Statement

RSB: Conceptualisation, formal analysis, investigation, project administration, visualisation, data collection, writing – original draft, and writing – review & editing. KRK, GAS, BM and EM: Conceptualization, writing – review & editing. MPAF: Conceptualisation and Methodology. IMP, JVA and JPGP: Project administration, writing – review & editing. MSH: Conceptualisation, funding acquisition, data curation, formal analysis, investigation, methodology, project administration, validation, visualisation, supervision, writing – original draft, and writing – review & editing.

RSB and MSH directly accessed and verified the underlying data reported in the manuscript.

All authors read and approved the final version of the manuscript and had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Abstract

Assessment tools for depression and anxiety usually inquire about frequency of symptoms. However, evidence suggests that different question framings might trigger different responses. Our aim is to test if asking about symptom's context, ability, duration and botherement adds validity to PHQ-9, GAD-7 and PROMIS depression and anxiety. Participants came from two cross-sectional convenience-sampled surveys (N=1,871) of adults (66% females, aged 33.4 ± 13.2), weighted to approximate with the state-level population. We examined measurement invariance across the different question frames, estimated whether framing affected mean scores, and tested their independent validity using covariate-adjusted and sample-weighted structural equation models. Validity was tested using tools assessing general disability, alcohol use, loneliness, well-being, grit, and frequency-based questions from depression and anxiety questionnaires. A bifactor model was applied to test the internal consistency of the question-frames under the presence of a general factor (i.e., depression or anxiety). Measurement invariance was supported across the different frames. Framing questions as ability (i.e., "how easily…") produced a higher score, compared with framing by context (i.e., "in which daily situations…"). Construct and criterion validity analysis demonstrate that variance explained using multiple question frames were similar to using only one. We detected a strong overarching factor for each instrument, with little variances left to be explained by the question frame. Therefore, it is unlikely that using different adverbial phrasings can help clinicians and researchers to improve their ability to detect depression or anxiety.

Key words: PHQ-9, GAD-7, PROMIS, mental health questionnaire, question frame.

Public significance statement

Levels of depression and anxiety symptoms in patients are often assessed in terms of their frequency in past weeks ("How often…"). We found that asking participants about their depression and anxiety symptoms in relation to context, ability, duration and botherement,

instead of frequency alone, do not add information while assessing these symptoms in online assessments.

Introduction

Questionnaires are the most common tool used for mental health assessments in clinical and research settings, with hundreds of options developed over the past century (Santor et al., 2006). Even considering a single construct, such as depression or anxiety, these questionnaires are diverse in their content, structure, scoring and adverbial framing (Fried, 2017; Newson et al., 2020; Wall & Lee, 2021) and factors such as the number of items, response options, and item order can impact the construct's assessment (Schwarz, 1999). This can result in reproducibility problems as measures may not be assessing the same construct (Fried, 2017). Furthermore, researchers seek to pool data sets using different questionnaires and not knowing the impact of certain characteristics, such as question adverbial framing, can lead to uncertainty on how to harmonize different questionnaires (Curran & Hussong, 2009; Fortier et al., 2017; McElroy et al., 2021).

In an attempt to minimize the lack of standardization on mental health measures, two large funders – the US National Institute of Mental Health (NIMH) and the Wellcome Trust – established the use of specific questionnaires as pre-requisites to obtain funding. These include the Patient Health Questionnaire (PHQ-9) for the assessment of depressive symptoms and the General Anxiety Disorder (GAD-7) for the assessment of anxiety symptoms in adults (Wolpert, 2020). Both are brief self-reports that have been extensively validated and inquiry about symptom's frequency over the past two weeks (Kroenke et al., 2001; Spitzer et al., 2006). In turn, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR) recommends the Patient-Related Outcome Measurement Information Systems (PROMIS) for depression and anxiety, which inquiry about symptom's frequency over the past week. Critiques have been raised regarding these recommendations because there is a lack of empirical testing to establish

which is the best assessment for depression and anxiety, which vary in terms of included symptoms and also adverbial frame (Patalay & Fried, 2021).

In a typical mental health questionnaire, questions about symptoms (e.g. "feeling down, depressed, or hopeless") are preceded by an adverbial question framing (e.g. "How often…") that asks respondents to think about their symptoms in specific terms. For example, the question framing may lead respondents to consider symptom *presence* or absence, symptom *frequency*, the specific *contexts* in which symptoms emerge, how *easy* the symptoms occur or is felt (e.g., as small things can quickly trigger an anger outburst or make one cry), symptom *duration*, or how *much they have been bothered* by the symptom. Depression and anxiety questionnaires usually inquiry about symptom presence, intensity or frequency (Newson et al., 2020), in line with the diagnostic criteria of the DSM and ICD manuals, while alternatives have been rarely explored (Krabbe & Forkmann, 2012).

The framing effect was first described by Tversky and Kahneman (1981), who demonstrated that participants' responses can be systematically influenced by the way options are framed. They described a public health problem and framed options as positively (e.g., "how many people can be saved") and negatively (e.g., "how many people can be lost"). Their results showed that, despite the fact that the outcome was essentially the same (e.g., the same number of people survived), people tended to change their responses when options were presented in a matter of gain or loss. This cognitive bias reveals how the mere linguistic formulation of options can sway choices and perceptions.

Thus, the type of question framing may change the ways in which respondents report their mental health symptoms. For example, inquiry about frequency can trigger a memory of when someone experienced a given symptom, whereas questions about how bothered they were by the symptom may trigger affective memories about the intensity of the experience (Mordeno et al., 2021; Schwarz, 1999). Furthermore, as different perceptions about symptomatology

might emerge depending on how the symptom is asked, they can be complementary to inform the target construct. As an example, person "A" can be bothered by feeling mildly depressed but feel it every day and person "B" can be severely bothered, to a point being in bed all day, but feel it occasionally. If a questionnaire inquires only about frequency, it might miss incapacitating depressive symptoms of person B. However, the hypothesis that framing mental health questions with different adverbs would lead to meaningful differences in response patterns has rarely been tested.

Previous qualitative studies of question framing found that symptom intensity (i.e., how strong the symptom is), as opposed to frequency and duration, is poorly understood by participants when assessing pediatric symptoms using the Child and Adolescent Psychiatric Assessment (Angold et al., 1995). Empirically, this can imply that the scores of a construct measured with intensity-framed questions would have most of its variation explained by measurement error instead of the intended construct, when compared with frequency and duration frames. In clinical sample, self-reports of depressive symptom frequency (e.g. "Sometimes I am sad") showed higher stability over time compared to symptom intensity (e.g. "My sadness is strong") (Krabbe & Forkmann, 2012, 2014). Moreover, frequency and intensity presented low score congruency and both frames rendered scores that were more influenced by personal factors when compared to unframed questions (e.g. "I am sad") (Krabbe & Forkmann, 2014). This implies that some of the information from depression and anxiety scores is being carried out by other factors (e.g., time, gender, social class) and less by the target construct when questions are framed as frequency or intensity. Furthermore, if there are score differences depending on how the question is framed, it is not clear if this is mere score inflation/deflation, or these differences impact on construct validity. In that sense, studies found equivalent correlations between frequency- and intensity-framed depression and anxiety questionnaires

with external validators, suggesting that they present equal predictive validity (Niileksela & Jones, 2022; Zimmerman & Kerr, 2019).

To examine the impact of question frame for measuring mental health, few studies have explicitly tested for measurement invariance across different framings. Moreover, the few studies that exist report contradictory findings while testing frequency- and intensity-framed questions for depression and anxiety (Mordeno et al., 2021; Niileksela & Jones, 2022; Zimmerman & Kerr, 2019). This contradiction might be explained by study design (e.g., frames not presented to the same participant, clinical vs community samples, online vs in-person) or "intensity" operationalization, which was found to be invariant if described as a "problem" (Niileksela & Jones, 2022) or non-invariant if described as "botherement" caused by the symptom (Mordeno et al., 2021). Furthermore, in the study conducted by Niileksela and Jones (2022), several relevant issues were not examined while testing measurement invariance of question frames, such as the priming effect of the adverbial frames and the additive validity of inquiring multiple question frames.

Priming effects might pose a particular methodological issue in studies that aim to examine the impact of adverbial frames in capturing a given construct (Gries, 2005; Schwarz, 1999; Strack et al., 1988). As an example, the PHQ-9 is primed with frequency-framed questions, meaning that all questions are framed the same and only variable part of the questionnaire is the symptom. Thus, participants are usually asked about different symptoms within the same adverbial frame (frame-primed). In this presentation, attention might be given to symptoms as they are the varying part of the questionnaire. To investigate this effect, questionnaires should also be structured to ask about the same symptom (symptom-primed), inquiring about it in many frames (i.e., "how often do you feel depressed", "in which situations do you feel depressed", etc).

Krishnakumar and colleagues (2021) examined five adverbial frames (frequency, context, duration, ability and botherement) for the Extended Strengths and Weaknesses Assessment of Normal Behavior (E-SWAN), a dimensional measure of attention and hyperactivity symptoms. They found that when they asked about the context of symptom onset (i.e., "When does your child…") and a patient's ability to control it (i.e., "How well does your child…") more individual E-SWAN items were associated with levels of disability (as measured by the WHO Disability Assessment Schedule - WHODAS) than items framed as frequency. These findings encourage to expand the investigations into question framing in other mental health assessment, beyond intensity and frequency. However, these findings were related to assessing attention and hyperactivity, in which context is relevant even for the diagnostic operationalization of attention-deficit and hyperactivity disorder. For suicide ideation, which is related with depression as one of its symptoms, previous findings on adolescents revealed that frequency of suicidal thoughts better predict future suicide attempts than asking about for how long adolescents had been contemplating suicide (Miranda et al., 2014). The reason for these results remains elusive. It is possible that some frames, such as recalling frequency of events, might be based on less affective memories when compared to intensity or botherement, which are affective evaluations of the experience and more prone to recall bias (Redelmeier & Kahneman, 1996). Nonetheless, alongside measurement invariance and score differences, question frames need to have and to add validity for measuring psychological constructs (VanderWeele, 2022) such as depression and anxiety.

Although the PHQ-9, GAD-7 and PROMIS are widely used and recommended measures of depression and anxiety, little evidence supports the assumption that asking only about the frequency of symptoms is the most reliable and valid approach to capturing these constructs. Context, ability, and duration of symptoms may be similarly relevant for assessing information about depression and anxiety. For example, certain symptoms might be frequent

but occur only in a specific situation, such as while at work or home. Moreover, sadness or irritability can be experienced more easily than usual, regardless of frequency or context, which could be informative in terms of episodic mood change. In terms of duration, depression and anxiety symptoms may occur a few days a week but experienced for hours instead of a few minutes. Therefore, in the present study, we aimed to answer the following research questions:

1) Do question frames focused on symptom context, ability, duration and botherement capture the same underlying construct as frequency-based frames?
2) Is there mean score differences among questionnaires framed as context, ability, duration and botherement and the originals frequency-based questionnaires, independently of how symptoms are presented to the respondent (priming with adverbial frames or symptoms first)?
3) If adverbial frames present different mean score levels, are these differences meaningful in terms of convergent, divergent and criterion validity (disability, alcohol use and loneliness – convergent validity, well-being and grit – divergent validity, and frequency-based questions from equivalent depression and anxiety questionnaires – criterion validity)? Furthermore, do multiple adverbial frames add construct's convergent, divergent and criterion validity in comparison with single frame (e.g., asking about frequency only)?
4) Are the question frames reliable when analyzed under a model that considers the general construct that the questionnaire aims to capture?

Based on previous studies, we hypothesize that question frames will invariantly capture the targeted constructs, there will be mean differences between frames regardless of priming effects (i.e., which is constant presented, the adverbial frame or the symptom), and they will have independent associations with the validators, adding on the total explained variance. We expect these results will improve the utility of well-established instruments by informing if and

how different question frames are useful to measure depression and anxiety and, thus, get implemented in routinely psychopathology assessment.

Methodology

Transparency and Openness

All data, study design, analysis plan and analysis code for this study have been made publicly available at the OSF and can be accessed at https://osf.io/p8j2v. Preregistration with study design and hypotheses are available at https://osf.io/gtm2j (Brümmer & Hoffmann, 2021).

Sample

We conducted two cross-sectional studies, using convenience samples from the Federal University of Santa Maria, a Southern-Brazilian state-funded university. Data was gathered in a web-based survey using GoogleForms. Participants were recruited through the university's e-mail system to anonymously complete the survey only once, between September and December 2021 (study design 1 – SD1) or April 2022 (study design 2 – SD2). SD1 and SD2 are described below and were filled out by different participants. Inclusion criteria were to be a Brazilian citizen or live in Brazil, being aged 18 years or above, and having read and accepted the informed consent form. We used the last census data for the state of Rio Grande do Sul to compute survey weights, used in the analysis (IBGE, 2012), as described below.

Questionnaire construction

We constructed our questionnaire based on the PHQ-9 (9 items) and GAD-7 (7 items). In addition, PROMIS depression (8 items) and anxiety (7 items) were administered to test

criterion validity. We also tested different frames on the PROMIS measures for sensitivity analysis in study 1 (PROMIS license nº 1357-1085). All items of the original questionnaire (PHQ-9, GAD-7, PROMIS depression and anxiety) assesses symptom frequency, and are framed by relatively similar prompts, that consist of "Over the last two weeks/seven days, how often have they been bothered by [symptom]". In our adapted survey, we decomposed these prompts into a pre-prompt describing the period of assessment ("Over the last two weeks"), the adverbial framing (e.g. "*How often*…?") and a verb (e.g. "have you been *bothered by*…"). Our goal was to change the adverbial framing from frequency to context, ability, duration and botherement intensity without changing the probed symptom or the phrase structure, hence introducing the least source of non-intended variance as possible.

Figure 1 depicts an example of the process. First, we extracted the symptom assessed in each original item, in each questionnaire. For instance, from the PHQ-9 item "Over the last two weeks, how often have you been bothered by [little interest or pleasure in doing]", we considered "little interest or pleasure in doing" as the symptom. Second, we created an adverbial framing template for the set of items (originally, in Portuguese): "In which daily situations" for context; "For how long" for duration; "How easily" for ability; and "How much have you been bothered" for botherement intensity; frequency framing remained the same. Of note, in the Brazilian Portuguese PHQ-9 and GAD-7, the "have you been bothered by" expression is absent in the original frequency-framed questions. This process produced 54 PHQ-9 and 42 GAD-7 items, including the original frequency items. Finally, we developed the response options for each new framing. We maintained the Likert-type four options for PHQ-9 and GAD-7 (and a five-point response scale for the newly reframed PROMIS items as part of the sensitivity analysis). Response options are depicted in Figure 1 and were devised to be coherent with the adverbial framing and to capture the whole spectrum of symptom presentation. Frequency questions' original response options were maintained.

- Figure 1 here -

PHQ-9 and GAD-7 are constructed in a way that the items (which each refer to specific symptoms) are primed by the frequency-framed prompt. By applying other adverbial frames to the same structure, it is possible that the adverbial frame priming effect could fade out while repeating the inquiry about the same symptom (Gries, 2005; Schwarz, 1999; Strack et al., 1988). Furthermore, the order of the questions can have an impact on item correlations (McFarland, 1981; Tourangeau & Smith, 1996). We mitigated these risks using two strategies.

First, because participants could simply ignore the given framing and directly ponder about the symptoms, we designed two studies (i.e. SD1 and SD2) with different presentations of the framed questions. For that, we applied different versions of the questionnaire to the same population. In SD1 the questions were presented to the participants as multiple-choice grids containing all symptoms grouped by framing (as originally performed by PHQ-9 and GAD-7), displaying each framing on separate pages, to assure the independence of each set of questions (i.e., frame priming). In SD2, we grouped each symptom by the five framings. That is, a given symptom was asked in five different ways depending on the five adverbial frames, aiming to direct respondent's attention to the framing itself (i.e. symptom primed). SD1 and SD2 are depicted in Figure 2. If results from both studies are equivalent, we may conclude that priming the question either by adverbial frame or symptom does not have a relevant effect while capturing depression and anxiety. Of note, sensitivity analysis with PROMIS questionnaire used the SD1 design only.

- Figure 2 here -

Second, we altered the framing presentation sequence every time the questionnaire collected around 250 new answers in SD1 and around 100 in SD2, allowing the randomization of this priming effect. Additionally, we deactivated the possibility to edit given answers and the "back" option in the online questionnaire.

Measures for validity tests

Participants responded to our modified versions of the PHQ-9, GAD-7, PROMIS depression and anxiety along with sociodemographic questions and six other measures, described below. All of these instruments had previously been subject to translation, cross-cultural adaptation, and validation in Brazil (Barroso et al., 2016; Castro et al., 2014; Moreira et al., 2015; Moreno et al., 2016; Moretti-Pires & Corradi-Webster, 2011; Noronha & Almeida, 2022; dos Santos et al., 2015; Santos et al., 2013). The complete questionnaire, consisting of 221 questions, is available in the supplemental material (page 9 – Brazilian Portuguese – and 41 – English translation) and online at https://osf.io/p8j2v, in English and Portuguese.

Participants first answered a set of socio-demographic questions comprising date of birth, state, sex, racial/ethnic group, religion, sexual orientation, marital status, number of children, family income, level of education and employment status. Family income response options corresponded to the five social classes established by the Brazilian Institute of Geography and Statistics' socioeconomic stratification based on multiples of the minimum wage in 2020 – R$1,045.00 – and coded as E (≤2 minimum wages), D (2-4), C (4-10); B (10 to 20) and A (≥ 20 minimum wages) (IBGE, 2012).

Convergent validity. Depression and anxiety are frequently associated with functional disability (Edlund et al., 2018), alcohol use problems (Merikangas et al., 1998) and loneliness (Palgi et al., 2020). Therefore, we assessed participants' level of impairment through the

Portuguese translation of World Health Organization Disability Assessment Schedule (WHODAS 2.0). The 12-item WHODAS evaluates disability assessing respondents' difficulty doing day-to-day activities such as taking care of household responsibilities, learning new tasks and maintaining friendships. The WHODAS is a five-point Likert-type scale with response options ranging from “None” to “Extreme or cannot do” (Moreira et al., 2015).

The brief version of Alcohol Use Disorders Identification Test (AUDIT-C), preconized by the WHO, consists of three items assessing frequency of alcohol use, number of doses and binge drinking. It was chosen because of its objectivity and satisfactory psychometric properties (Moreira et al., 2015).

The UCLA-Loneliness scale 3-item version consists of three frequency-probing items with three response options that evaluate one's feelings of isolation and perception of social support, such as “How often do you feel isolated from others?”. In previous studies, this instrument presented good convergent validity - greater scores were associated with worsening of physical and mental health status -, internal consistency (Cronbach’s alpha = 0,72) and a 0.82 correlation with the 20-item version (Barroso et al., 2016; Hughes et al., 2004)

Divergent validity. Positive socioemotional skills, measured by consistency of interest, perseverance of effort (Almlund et al., 2011; Duckworth & Quinn, 2009) and well-being (Keyes, 2002; Patalay & Fitzsimons, 2016) are orthogonal constructs in relation to psychopathology. Therefore, we used the short version of the Grit Scale (GRIT-S), a two-factor scale for assessing consistency of interest and perseverance of effort to assess these positive skills (Duckworth et al., 2021; Duckworth & Quinn, 2009). GRIT-s is scaled to a five-point Likert-type response options ranging from “Very much like me” to “Not like me at all”. The GRIT-S Portuguese version demonstrated good internal consistency and high correlation (r=0.90) with contentiousness personality trait (Primi et al., 2019).

Well-being was assessed using the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS). We chose the 7-item instrument, which assesses the frequency at which the respondent has experienced feelings and thoughts presented in statements such as "I've been feeling useful" with five Likert-type response options ranging from "None of the time" to "All of the time". This version presents good internal consistency (Cronbach's alpha = 0.89) and a factorial structure that explains 65% of the data variance (dos Santos et al., 2015).

Criterion validity. We evaluated if depression and anxiety scores yielded by different adverbial frames on the PHQ-9 and GAD-7 correlated with depression and anxiety scores yielded by frequency-based frames on the original PROMIS depression and PROMIS anxiety questionnaires. For questionnaire sensitivity analysis (i.e., reframed PROMIS), frequency-framed PHQ-9 and GAD-7 were used as criterion validity.

Data analysis

Survey weights. All confirmatory factor analysis (CFA), regression models and structural equation models (SEM) were weighted using survey weights to maximize sample representativeness. We weighted our sample using population margins regarding sex at birth, age groups and race/ethnicity, according to the Brazilian last census on the state of Rio Grande do Sul (IBGE, 2012). This procedure applies iterative post-stratification to match population margins to the survey sample proportions. Survey weight was trimmed to a minimum of 0.25 and up to 4.

Confirmatory factor analysis and measurement invariance. We performed CFA for each original depression and anxiety questionnaire and the newly framed ones in SD1 and SD2, to analyze their global fit. We applied the delta parameterization and weighted least square with diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test

statistics (WLSMV) estimators to account for the ordinal nature of the scales. Model fit parameters were Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker Lewis Index (TLI) and standardized root mean-square residual (SRMR). Values of RMSEA near or below 0.08 represent acceptable model fit, and values lower than 0.06 represent good-to-excellent model fit (Hu & Bentler, 1999; Kline, 2015). CFI and TLI values near or above 0.90 represent acceptable model fit, while values higher than 0.95 represent a good-to-excellent model fit (Hu & Bentler, 1999; Kline, 2015). SRMR lower or equal than 0.10 indicate adequate fit, and lower than 0.06 in combination with previous indices indicate good fit (Hu & Bentler, 1999; Kline, 2015).

Question 1) Are frames capturing the same construct? Measurement invariance. After estimating global fit and factor structures, we performed CFA to test measurement invariance across the question frames within SD1 and SD2. This would ensure that differences between depression and anxiety scores were truly attributable to depression and anxiety differences and not that a different phenomenon is being captured due to different adverbial frames. We compared a series of nested models with increasing levels of restrictions. We examined if model structure (configural invariance), structure and factor loadings (metric invariance) and structure, loadings and thresholds (scalar invariance) maintained a similar model fit when constrained to be the same across question frames. We used the method described by Liu et al (2017), developed for testing measurement invariance for repeated ordered-categorical variables. As X^2 is highly sensitive to sample size, we used three alternative measures of fit index derived from differences between more and less restricted models. Model fit parameters were RMSEA, CFI, and SRMR. ΔCFI <0,01 and ΔRMSEA<0,015 or SRMR <0.010 imply that added restrictions don't impact model fit and, hence, framings could be considered invariant (Chen, 2007; Meredith, 1993; Svetina et al., 2020).

Question 2) Do question frames have mean differences? Mean difference between frequency-framed total score and other question frames. We estimated mean differences between the total item sum score for each framing type to understand whether changing adverbial framing could change reported depression/anxiety score levels. We used linear regression models with total score as the dependent variable and question frames as the independent variable (frequency frame as reference category). PHQ-9 and GAD-7 regression models used interaction terms between question frame and study design to account and test for design effects (i.e., SD1 and SD2, describe above). All regression models were adjusted for age, sex and family income.

Question 3) Are differences meaningful? Structural equation modeling to test convergent, divergent and criterion validity. The main objective was to test if question frames were independent sources of information by testing their associations with relevant validators. To test this hypothesis, we first evaluated model fit of each measure - WHODAS, AUDIT-C, UCLA-Loneliness, GRIT-S (interest and effort), WEMWBS and frequency-framed depression and anxiety questionnaires (Kline, 2015). Subsequently, we performed multivariate regression analyses through structural equation modelling (SEM), using each framed construct regressing all the latent validators (i.e., one SEM for each frame, five SEM in total). Afterwards, we constructed a model containing all framings together to test their independent associations with the validators (multiple predictors and multiple outcomes). This allowed the comparison between regression coefficients and R^2 of each framing separately and the additive association of inquiring about symptoms in multiple ways.

Question 4) Are frames relevant sources of variance? Bifactor modelling. Bifactor model analysis was applied to evaluate the internal consistency and model-based reliability of each framing. Bifactor modelling is a CFA-based technique in which each item loads both on a general factor (construct) and on a specific factor (e.g., question framing). Here, the bifactor

model comprises two distinct sources of variance, namely: the overarching target construct (e.g. PHQ-9 depression) and the specific framing factors, as depicted in Figure 3. Items assessing the same symptom were allowed to correlate. Internal consistency for latent factors were assessed using omega (ω) and omega hierarchical (ωH). ω is a model-based reliability estimate, analogous to the alpha coefficient, but appropriate for tests that have varying factor loadings (Lucke, 2005; Raykov, 2001). ωH represents the proportion of total variance attributed to the general or specific factors and varies between 0 and 1, where higher scores indicate greater internal consistency. Specifically, for the bifactor model, we further estimated the factor determinacy (FD - the correlation between the factor scores and the estimated factor and that factor scores can be used) and H index (a measure of construct replicability) (Dueber, 2017; Rodriguez et al., 2016). Good internal consistency is estimated if a factor has ωH>=0.8. A well-defined and replicable factor is considered when H index >= 0.8 and FD >= 0.9 (Dueber, 2017; Rodriguez et al., 2016). If a framing factor presents good internal reliability, it can help to explain if findings from the SEM analysis are truly due to question frames or the overarching construct regardless of the question frames.

- Figure 3 here -

All analyses were conducted in R software version 4.2.1. Survey weights were calculated using the *rake* function from the *survey* package (Lumley, 2021). CFA and SEM were conducted using the *lavaan* package (Rosseel et al., 2018). Reliability and model-based indices were calculated using the *BifactorIndicesCalculator* package (Dueber, 2017).

Ethics

The study was approved by the Ethics Committee (CAAE: 49106221.5.0000.5346, approval nº 4.844.393). Participants were encouraged to contact researchers if they felt in need of mental health assistance, in which case they were referred to the state-funded health system. All subjects were informed about the study's procedures and purposes and had to sign a consent form before responding to the questionnaire. No respondent had to give any personal information or identify themselves during participation. All answers were stored under a random assigned number and were managed by the senior author.

Results

From the universities' population, 1,311 and 595 agreed to participate in SD1 and SD2 respectively. After exclusion of duplicates and filtering for age ≥18, the final sample was 1256 (mean age = 33.5 ± 13.2; 66.8% females) and 572 (mean age = 33.0 ± 12.8; 69.8% females) for study 1 and 2 respectively. Table 1 shows the demographics, total scores for frequency-framed PHQ-9, GAD-7 and PROMIS questionnaires, as well as the validators. 1252 (99.7%) participants in SD1 and 565 (98.8%) in SD2 presented complete PHQ-9, GAD-7 and PROMIS data. Census data used to calculate survey weights (M=1.0; SD=0.71; IQR=0.52 – 1.24) are described in Table S1.

- Table 1 here -

Depression and anxiety factor models and frame invariance testing

Fit indices for each question frame factor model were all excellent in both SD1 and SD2, except for PROMIS depression (Table S2). Factor structure for each model/frame are described in Table S3. Question frames were all frame-invariant (Table 2) and, therefore, differences between new and frequency-framed questionnaires were estimated.

- Table 2 here -

Question frame mean difference and study design effects

Table 3 describe the mean total score difference by question frame. Context frame ("in which situations") resulted in lower scores when compared with frequency frame in all questionnaires aside PHQ-9. Asking questions as ability ("how easy") resulted in higher scores for all questionnaires. Duration ("for how long") resulted in higher scores for PHQ-9 and GAD-7 but lower in PROMIS anxiety, showing some sensitivity depending on the questionnaire. Framed as botherement ("how much have you been bothered by") did not present overall difference when compared with frequency-framed questionnaires.

Table 3 also shows that there is no evidence that the study design (i.e., either if the participants were primed with question frames or symptoms) have an association with mean score of PHQ-9 and GAD-7, nor that they modify a specific question frame (i.e., no interaction).

- Table 3 here -

Convergent, divergent and criterion validity – independent associations of question frames

CFAs of the questionnaires used for convergent (WHODAS, AUDIT-C and UCLA-Loneliness), divergent (GRIT-S and WEMWBS) and criterion validity (frequency-framed questions for depression and anxiety measures) demonstrate that the measurement models, as suggested by the developers, fitted the data well (see Table S4 for model fit and Table S5 for factor structure).

Fit indices were all excellent for the SEM using multivariate (all validators regressed on one question frame at a time) and the multiple SEM (all validators regressed on depression/anxiety questionnaires framed in five ways simultaneously) (Table S6). Figure 4A

demonstrate that PHQ-9 (Table S7), GAD-7 (Table S8) and the PROMIS questionnaires (depression in Table S9 and anxiety in Table S10) framed in five ways are equivalently associated with the validators. In the multiple SEMs, association with the validators are parsed out, showing some independent associations but with smaller regression coefficients (Figure 4B). The R^2 (explained variance) of the multiple SEMs have not increased substantially when compared with the R^2 from the multivariate SEMs (Figure 4C). This demonstrates that there is no advantage of inquiring about symptoms of depression and anxiety using different adverbial phrasings.

- Figure 4 here -

Testing question frame common variance – bifactor models

The bifactor models fitted the data well for the PHQ-9 (RMSEA = 0.042, 90%CI = 0.041-0.044; CFI=0.9999, TLI=0.9999, SRMR=0.038), GAD-7 (RMSEA = 0.024, 90%CI = 0.022-0.027; CFI=0.999, TLI=0.999, SRMR=0.024), PROMIS depression (RMSEA = 0.078, 90%CI = 0.076-0.080; CFI=0.998, TLI=0.998, SRMR=0.044) and anxiety (RMSEA = 0.045, 90%CI = 0.043-0.047; CFI=0.999, TLI=0.999, SRMR=0.024). Factor structure of the four bifactor models are described in tables S11 to S14. The model shows that there is a strong source of variance coming from the general factor (i.e., targeted construct) and the question frame factors do not contribute as a major source of information, given low ωH, H index and FD (Table 4). This provides evidence on why the adverbial frames do not provide independent information and presented almost equivalent level in terms of correlation coefficient and variance explained in Figure 4.

- Table 4 here -

Discussion

This study demonstrates that different adverbial frames applied to commonly used depression and anxiety questionnaires invariantly capture the targeted constructs. Furthermore, we found that PHQ-9, GAD-7 and PROMIS questionnaires presented lower, higher, and inconsistently different mean scores when framed as context, ability, and duration, respectively, when compared to frequency. This was independent of how the questions were presented (study design) and covariates. However, these did not translate into meaningful substantive differences as the validators do not relate with the question frames but only with the targeted construct (i.e., depression or anxiety general factor). This was evident by observing that questionnaires framed with frequency, context, ability, duration and botherement have similar convergent, divergent and criterion latent validity. While testing independent associations, different question frames parsed out their associations and did not present additive information for the validators. This is probably due to low internal consistency of frame factors for depression and anxiety and the association might be given due to the construct itself rather than the way the questions were framed.

First and foremost, this study aims to test if changing how the symptom of depression and anxiety are asked would add on the validity of depression and anxiety constructs. Considering depression and anxiety assessment, this study adds to the existing literature that has raised questions regarding potential differences when asking subjects about different symptom aspects (Newson et al., 2020). Previous literature suggested that differences could exist when multiple frames were analyzed (Angold et al., 1995; Krishnakumar et al., 2021; Levin et al., 1998; Miranda et al., 2014), but not between frequency and intensity (Krabbe & Forkmann, 2014; Niileksela & Jones, 2022; Pattanaik et al., 2022). As different from externally observing symptom characteristics, inquiring them online as we did present some validity to capture disability, well-being and other constructs, but they are not different ways to triangulate symptomatology. Rather, they can be used interchangeably and produce a true general factor

of psychopathology with very little information left for framing factors (Van Bork et al., 2017). In fact, previous studies even suggest that simply inquiring unframed questions, such as a “yes or no” answer to questions like “I have being feeling depressed over the past two weeks” would capture information just as framed questions (Krabbe & Forkmann, 2012, 2014). Our study has not directly tested this hypothesis, but the bifactor analysis support this notion.

However, the terminology used in this field of research might have impact on differences among study findings. The “question frame” term used in this study refers to different adverbs in questions about symptoms. However, other studies have used different terms to it, such as "symptom aspect" (Newson et al., 2020), “item response option” (Niileksela & Jones, 2022) or "rating scale anchors" (Krabble 2012-2014, Newson 2020). We understand that “aspect” refers to different ways that the phenomenon can be observed but perhaps not in online assessments. Depressive symptoms might be frequent and mild, or unfrequented and very strong. In fact, previous studies have examined that socioeconomic status might be positively associated with frequency of depressive symptoms, but negatively associated with their intensity (Schneider & Stone, 2014), meaning that the wealthier you are, the less intense might be your depressive symptom, despite frequent. This may not be captured by online self-reports, which is a common way to inquiry about how people feel in modern research settings. “Item response option” is confusing because it might refer to the number of options in a Likert-type scale, if a scale is unipolar or bipolar in their response options, and many other features in a scale. “Rating scale anchors” might not be precise enough because anchors are response options that do not serve to be chosen but to make more likely a subject to endorse a given response option when choosing among two or more uncertain options. We rather prefer to use the term “frame” as like valence frame (Levin et al., 1998).

This study has practical implications for assign symptoms in the general and clinical population with validated measures, to attend different purposes, from screening to follow-ups.

Online self-reports have the potential to reach people on a wider level than individual clinician evaluations, especially in the context of online care (Moreno et al., 2020). Considering the present results, online self-reports capture the same constructs (confirming our first hypothesis) regardless of how the question is being farmed and they do not add validity if symptoms are asked in multiple ways (contrary to our hypothesis). One explanation is that the question can be understood in some form of heuristic process in which questions are "*perceived as belonging to the same conversational context*" (Strack et al., 1988). This is reinforced in our study because results were not sensitive to priming effects of presenting the symptom or adverbial frame first. Furthermore, this study is also valuable to inform data harmonization efforts while integrating different data sets. Considering online self-reports on depression and anxiety, it is very unlikely that adverbial framing have an impact on data harmonization and items can be harmonized as long as the content is similar (Fortier et al., 2017; Hoffmann et al., 2023; McElroy et al., 2021).

Such investigation is not free from limitations. First, although we examined depression and anxiety, corroborative results could support investigating multi-dimensional framing in other mental disorders. However, extending two-frame studies, our study demonstrates that despite mean differences, they were not meaningful while investigating five adverbial frames for inquiring depression and anxiety symptoms in different questionnaires. Second, we used a convenience sample from a specific university in south Brazil and results may not be generalizable. We used sampling weights in all analysis to minimize the impact of missing population representativeness. Third, the results may not apply in different settings, such as in-person or clinician-rated interviews. Despite evidence on equivalence between self-reports and structured clinician-rated interviews (Boyle et al., 2022), this study might be present higher external validity on online self-rated questionnaires to university personal.

Conclusions

This study provides evidence that different adverbs (frequency, context, ability, duration and botherement) when applied to questions about depression and anxiety symptoms measure the same underlying construct. Furthermore, we demonstrate that despite significant mean differences, these are not meaningful in validity tests. PHQ-9, GAD-7 and PROMIS depression and anxiety present equivalent convergent, divergent and criterion validity and adding different frames when asking about symptoms does not add in terms of overall construct validity (i.e., aspects do not add information). This indicates that people with depression and anxiety understand different adverbs in the same way in an online setting and the important information relies on the overarching construct of interest. Studies need to be mindful while using online self-reports to investigate different aspects from depression and anxiety, which might be better investigated by external observation or interviews. Furthermore, for online assessments, adverbial frame might have little impact to measure depression and anxiety. Future studies need to replicate these findings for different constructs and clinical samples.

Constraints on Generality

The study sample consists of students and employees from a major state-funded university in Brazil. We have used post-stratification weighting procedure to match population margins to the survey sample proportions so the differences of the sample and the state population was minimized. As it's possible to observe in the data description, age and income were similar to census data for the state of Rio Grande do Sul and the sample were over representative for female sex. Therefore, caution must be taken when applying these findings to other Brazilian regions and countries, especially due to translation characteristics of the questionnaires. Nonetheless, we hypothesize these differences could be minimal once measurement invariances hold true for the questionnaire translations.

REFER ENCES

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality Psychology and Economics* (Working Paper 16822). National Bureau of Economic Research. http://www.nber.org/papers/w16822

Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., & Rutter, M. (1995). The Child and Adolescent Psychiatric Assessment (CAPA). *Psychological Medicine*, *25*(4), 739–753. https://doi.org/10.1017/S003329170003498X

Barroso, S. M., Andrade, V. S. de, Midgett, A. H., & Carvalho, R. G. N. de. (2016). Evidências de validade da Escala Brasileira de Solidão UCLA. *Jornal Brasileiro de Psiquiatria*, *65*, 68–75. https://doi.org/10.1590/0047-2085000000105

Boyle, M. H., Duncan, L., Wang, L., & Georgiades, K. (2022). Problem checklists and standardized diagnostic interviews: Evidence of psychometric equivalence for classifying psychiatric disorder among children and youth in epidemiological studies. *Journal of Child Psychology and Psychiatry*, *n/a*(n/a). https://doi.org/10.1111/jcpp.13735

Brümmer, R. de S., & Hoffmann, M. S. (2021). *Depression and anxiety question framing: The impact of asking about symptom's frequency, context, duration, ability and botherement.* https://doi.org/10.17605/OSF.IO/P8J2V

Castro, N. F. C. de, Rezende, C. H. A. de, Mendonça, T. M. da S., Silva, C. H. M. da, & Pinto, R. de M. C. (2014). Adaptação transcultural dos Bancos de Itens de Ansiedade e Depressão do *Patient-Reported Outcomes Measurement Information System* (PROMIS) para língua portuguesa. *Cadernos de Saúde Pública*, *30*, 879–884. https://doi.org/10.1590/0102-311X00096113

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Curran, P. J., & Hussong, A. M. (2009). Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2021). Revisiting the Factor Structure of Grit: A Commentary on Duckworth and Quinn (2009). *Journal of Personality Assessment*, *103*(5), 573–575. https://doi.org/10.1080/00223891.2021.1942022

Dueber, D. (2017). Bifactor Indices Calculator: A Microsoft Excel-Based Tool to Calculate Various Indices Relevant to Bifactor CFA Models. *Educational, School, and Counseling Psychology Research Tools*. https://doi.org/10.13023/edp.tool.01

Edlund, M. J., Wang, J., Brown, K. G., Forman-Hoffman, V. L., Calvin, S. L., Hedden, S. L., & Bose, J. (2018). Which mental disorders are associated with the greatest impairment in functioning? *Social Psychiatry and Psychiatric Epidemiology*, *53*(11), 1265–1276. https://doi.org/10.1007/s00127-018-1554-6

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., Doiron, D., Stolk, R. P., Knoppers, B. M., Ferretti, V., Granda, P., & Burton, P. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, *46*(1), 103–105. https://doi.org/10.1093/ije/dyw075

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Gries, S. Th. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399. https://doi.org/10.1007/s10936-005-6139-3
Hoffmann, M. S., Moore, T. M., Axelrud, L. K., Tottenham, N., Pan, P. M., Miguel, E. C., Rohde, L. A., Milham, M. P., Satterthwaite, T. D., & Salum, G. A. (2023). An Evaluation of Item Harmonization Strategies Between Assessment Tools of Psychopathology in Children and Adolescents. *Assessment*, 10731911231163136. https://doi.org/10.1177/10731911231163136
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A Short Scale for Measuring Loneliness in Large Surveys. *Research on Aging*, *26*(6), 655–672. https://doi.org/10.1177/0164027504268574
IBGE. (2012). *Censo demográfico: 2010: Educação e deslocamento: Resultados da amostra*. Instituto Brasileiro de Geografia e Estatística (IBGE). https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7545
Keyes, C. L. M. (2002). The Mental Health Continuum: From Languishing to Flourishing in Life. *Journal of Health and Social Behavior*, *43*(2), 207–222. JSTOR. https://doi.org/10.2307/3090197
Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.
Krabbe, J., & Forkmann, T. (2012). Frequency vs. intensity: Which should be used as anchors for self-report instruments? *Health and Quality of Life Outcomes*, *10*(1), 107. https://doi.org/10.1186/1477-7525-10-107
Krabbe, J., & Forkmann, T. (2014). Frequency vs. intensity: Framing effects on patients' use of verbal rating scale anchors. *Comprehensive Psychiatry*, *55*(8), 1928–1936. https://doi.org/10.1016/j.comppsych.2014.06.010
Krishnakumar, A., Scopel Hoffmann, M., Schoeller, F., Clucas, J. C., Son, J., Müller-Naendrup, L., Salum, G., Milham, M., Lindner, A., & Klein, A. (2021). *Framing mental health questions using context better predicts disability than does frequency* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jm9kc
Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x
Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188. https://doi.org/10.1006/obhd.1998.2804
Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. https://doi.org/10.1037/met0000075
Lucke, J. F. (2005). The alpha and the omega of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29*(1). https://works.bepress.com/joseph_lucke/19/
Lumley, T. (2021). *survey: Analysis of Complex Survey Samples* (4.1-1) [Computer software]. https://CRAN.R-project.org/package=survey
McElroy, E., Villadsen, A., Patalay, P., Goodman, A., Richards, M., Northstone, K., Fearon, P., Tibber, M., Gondek, D., & Ploubidis, G. B. (2021). *Harmonisation and Measurement Properties of Mental Health Measures in Six British Cohorts*. CLOSER. https://www.closer.ac.uk/wp-content/uploads/210715-Harmonisation-measurement-properties-mental-health-measures-british-cohorts.pdf
McFarland, S. G. (1981). Effects of Question Order on Survey Responses. *The Public*

Opinion Quarterly, *45*(2), 208–215.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825
Merikangas, K. R., Mehta, R. L., Molnar, B. E., Walters, E. E., Swendsen, J. D., Aguilar-Gaziola, S., Bijl, R., Borges, G., Caraveo-Anduaga, J. J., DeWit, D. J., Kolody, B., Vega, W. A., Wittchen, H. U., & Kessler, R. C. (1998). Comorbidity of substance use disorders with mood and anxiety disorders: Results of the International Consortium in Psychiatric Epidemiology. *Addictive Behaviors*, *23*(6), 893–907. https://doi.org/10.1016/s0306-4603(98)00076-8
Miranda, R., Ortin, A., Scott, M., & Shaffer, D. (2014). Characteristics of suicidal ideation that predict the transition to future suicide attempts in adolescents. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *55*(11), 1288–1296. https://doi.org/10.1111/jcpp.12245
Mordeno, I. G., Nalipay, Ma. J. N., Luzano, J. G. C., Galela, D. S., & Ferolino, M. A. L. (2021). Development and validation of a DSM-5-based generalized anxiety disorder self-report Scale: Investigating frequency and intensity rating differences. *Current Psychology*, *40*(11), 5247–5255. https://doi.org/10.1007/s12144-019-00475-8
Moreira, A., Alvarelhão, J., Silva, A. G., Costa, R., & Queirós, A. (2015). Tradução e validação para português do WHODAS 2.0 - 12 itens em pessoas com 55 ou mais anos. *Revista Portuguesa de Saúde Pública*, *33*(2), 179–182. https://doi.org/10.1016/j.rpsp.2015.06.003
Moreno, A. L., DeSousa, D. A., Souza, A. M. F. L. P. de, Manfro, G. G., Salum, G. A., Koller, S. H., Osório, F. de L., & Crippa, J. A. de S. (2016). Factor structure, reliability, and item parameters of the brazilian-portuguese version of the GAD-7 questionnaire. *Temas Em Psicologia*, *24*(1), 367–376. https://doi.org/10.9788/TP2016.1-25
Moreno, C., Wykes, T., Galderisi, S., Nordentoft, M., Crossley, N., Jones, N., Cannon, M., Correll, C. U., Byrne, L., Carr, S., Chen, E. Y. H., Gorwood, P., Johnson, S., Kärkkäinen, H., Krystal, J. H., Lee, J., Lieberman, J., López-Jaramillo, C., Männikkö, M., … Arango, C. (2020). How mental health care should change as a consequence of the COVID-19 pandemic. *The Lancet Psychiatry*, *7*(9), 813–824. https://doi.org/10.1016/S2215-0366(20)30307-2
Moretti-Pires, R. O., & Corradi-Webster, C. M. (2011). Adaptação e validação do Alcohol Use Disorder Identification Test (AUDIT) para população ribeirinha do interior da Amazônia, Brasil. *Cadernos de Saúde Pública*, *27*, 497–509. https://doi.org/10.1590/S0102-311X2011000300010
Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The Heterogeneity of Mental Health Assessment. *Frontiers in Psychiatry*, *11*. https://www.frontiersin.org/article/10.3389/fpsyt.2020.00076
Niileksela, C. R., & Jones, N. B. (2022). Measurement Equality of Frequency and Severity Item Response Options on Depression and Generalized Anxiety Scales. *Assessment*, 10731911221134599. https://doi.org/10.1177/10731911221134599
Noronha, A. P., & Almeida, L. S. (2022). A construção e estudos psicométricos da escala de avaliação da garra: Versão internacional em língua portuguesa. *Revista de Psicologia, Educação e Cultura*, *26*(1), 8–23.
Palgi, Y., Shrira, A., Ring, L., Bodner, E., Avidor, S., Bergman, Y., Cohen-Fridel, S., Keisari, S., & Hoffman, Y. (2020). The loneliness pandemic: Loneliness and other concomitants of depression, anxiety and their comorbidity during the COVID-19 outbreak. *Journal of Affective Disorders*, *275*, 109–111. https://doi.org/10.1016/j.jad.2020.06.036
Patalay, P., & Fitzsimons, E. (2016). Correlates of Mental Illness and Wellbeing in Children: Are They the Same? Results From the UK Millennium Cohort Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(9), 771–783.

https://doi.org/10.1016/j.jaac.2016.05.019
Patalay, P., & Fried, E. I. (2021). Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *62*(8), 1032–1036. https://doi.org/10.1111/jcpp.13333
Pattanaik, S., John, M. T., Chung, S., & Keller, S. (2022). Should the frequency, severity, or both response scales be used for multi-item dental patient-reported outcome measures? *PeerJ*, *10*, e12717. https://doi.org/10.7717/peerj.12717
Primi, R., Santos, D. D. dos, Hauck, N., Fruyt, F. D., & John, O. P. (2019). Mapping self-report questionnaires for socio-emotional characteristics: What do they measure? *Estudos de Psicologia (Campinas)*, *36*. https://doi.org/10.1590/1982-0275201936e180138
Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 315–323. https://doi.org/10.1348/000711001159582
Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, *66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6
Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045
Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Chow, M., & Jorgensen, T. D. (2018). *lavaan: Latent Variable Analysis* (0.6-3) [Computer software]. https://CRAN.R-project.org/package=lavaan
Santor, D. A., Gregus, M., & Welch, A. (2006). FOCUS ARTICLE: Eight Decades of Measurement in Depression. *Measurement: Interdisciplinary Research and Perspectives*, *4*(3), 135–155. https://doi.org/10.1207/s15366359mea0403_1
Santos, J. J. A. dos, Costa, T. A. da, Guilherme, J. H., Silva, W. C. da, Abentroth, L. R. L., Krebs, J. A., Sotoriva, P., Santos, J. J. A. dos, Costa, T. A. da, Guilherme, J. H., Silva, W. C. da, Abentroth, L. R. L., Krebs, J. A., & Sotoriva, P. (2015). Adaptation and cross-cultural validation of the Brazilian version of the Warwick-Edinburgh mental well-being scale. *Revista Da Associação Médica Brasileira*, *61*(3), 209–214. https://doi.org/10.1590/1806-9282.61.03.209
Santos, I. S., Tavares, B. F., Munhoz, T. N., Almeida, L. S. P. de, Silva, N. T. B. da, Tams, B. D., Patella, A. M., & Matijasevich, A. (2013). Sensitivity and specificity of the Patient Health Questionnaire-9 (PHQ-9) among adults from the general population. *Cadernos de Saúde Pública*, *29*(8), 1533–1543. https://doi.org/10.1590/0102-311X00144612
Schneider, S., & Stone, A. A. (2014). Distinguishing between frequency and intensity of health-related symptoms from diary assessments. *Journal of Psychosomatic Research*, *77*(3), 205–212. https://doi.org/10.1016/j.jpsychores.2014.07.006
Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93
Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092
Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*(5), 429–442. https://doi.org/10.1002/ejsp.2420180505
Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the

lavaan/semTools Packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776
Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *The Public Opinion Quarterly*, *60*(2), 275–304.
Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683
Van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a Causal Interpretation of the Common Factor Model. *Disputatio*, *9*(47), 581–601. https://doi.org/10.1515/disp-2017-0019
VanderWeele, T. J. (2022). Constructed Measures and Causal Inference: Towards a New Model of Measurement for Psychosocial Constructs. *Epidemiology*, *33*(1), 141–151. https://doi.org/10.1097/EDE.0000000000001434
Wall, A., & Lee, E. (2021). *What do Anxiety Scales Really Measure? An Item Content Analysis of Self-Report Measures of Anxiety* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/t7gpx
Wolpert, M. (2020). *Funders agree first common metrics for mental health science*. https://www.linkedin.com/pulse/funders-agree-first-common-metrics-mental-health-science-wolpert
Zimmerman, M., & Kerr, S. (2019). How should the severity of depression be rated on self-report depression scales? *Psychiatry Research*, *280*, 112512. https://doi.org/10.1016/j.psychres.2019.112512

Table 1 - Sample description

	Sample	
	Study design 1	Study design 2
Sample size	1256	572
Age (in years, mean and SD)	33.5 (13.2)	33.0 (12.8)
Sex (female)	829 (66.0%)	399 (69.8%)
Race/ethnicity		
White, Japanese or Chinese	1045 (83.2%)	495 (86.5%)
Black, Indigenous or *Pardo* (Brown)	193 (15.3%)	77 (18.0%)
Other or I rather not answer	18 (1.4%)	0 (0%)
Sexual orientation		
Heterosexual	982 (78.2%)	422 (73.78%)
Homosexual	70 (5.6%)	37 (6.47%)
Bisexual	141 (11.2%)	87 (15.21%)
Other	27 (2.1%)	12 (2.10%)
I rather not answer	36 (2.9%)	14 (2.45%)
Marital status		
Married	402 (32.0%)	191 (33.39%)
Divorced	52 (4.1%)	21 (4.72%)
Dating	274 (21.8%)	106 (18.53%)
Widow	6 (0.5%)	1 (0.17%)
Single	522 (41.6%)	247 (43.18%)
Religion		
Afrobrazilian	27 (2.1%)	15 (2.62%)
Agnostic	104 (8.3%)	57 (9.97%)
Atheist	121 (9.6%)	66 (11.54%)
Budhist	8 (0.6%)	2 (0.35%)
Catholic	444 (35.4%)	194 (33.92%)
Spiritualized but non-religious	223 (17.8%)	92 (16.08%)
Evangelical	95 (7.6%)	39 (6.82%)
Islamist	2 (0.2%)	0 (0%)
Lutheran	21 (1.7%)	10 (1.75%)
Protestant	22 (1.8%)	11 (1.92%)
Espírita (christian)	119 (9.5%)	52 (9.09%)
Other	70 (5.6%)	34 (5.94%)
Have children	389 (31.0%)	179 (31.29%)
Family income		
E (≤2 minimum wages)	280 (22.3%)	128 (22.38%)
D (2+R$1.00 - 4 miminimum wages)	346 (27.5%)	167 (29.20%)
C (4+R$1.00 - 10 miminimum wages)	373 (29.7%)	176 (30.77%)
B (10+R$1.00 - 20 miminimum wages)	193 (15.4%)	82 (14.34%)
A (> 20+R$1.00 minimum wages)	64 (5.1%)	19 (3.32%)
Highest educational attainement		

Secondary school (complete or incomplete	533 (41.9%)	244 (42.13%)
Undergraduate student	122 (9.7%)	51 (8.92%)
Graduate student (current or finished)	601 (47.9%)	276 (48.25%)
Working status		
Public sector worker	405 (32.2%)	176 (30.77%)
Formal work	165 (13.1%)	65 (11.36%)
Informal work	122 (9.7%)	52 (9.09%)
Student	487 (38.8%)	241 (42.13%)
Unemployed	66 (5.3%)	32 (5.59%)
Retired	11 (0.9%)	6 (1.05%)
Hospitalization for mental problems		
Yes	79 (6.2%)	37 (93.53%)
No	1189 (93.8%)	535 (93.53%)
Current treatment for mental problems		
Yes	365 (28.8%)	174 (30.42%)
No	903 (71.2%)	398 (69.58%)
PHQ-9 (frequency frame)		
Mean [Min, Max]	10.90 [0, 27]	11.5 [0, 27]
SD	± 7.5	± 6.8
GAD-7 (frequency frame)		
Mean [Min, Max]	9.70 [0, 21]	9.88 [0, 21]
SD	± 6.10	± 5.66
PROMIS depression (frequency frame)		
Mean [Min, Max]	19.30 [8, 40]	20.50 [8, 40]
SD	± 8.53	± 9.17
PROMIS anxiety (frequency frame)		
Mean [Min, Max]	22.80 [8, 40]	22.70 [8, 40]
SD	± 8.59	± 8.59
WHODAS		
Mean [Min, Max]	23.45 [12, 56]	23.67 [12, 56]
SD	± 8.86	± 8.83
AUDIT-C		
Mean [Min, Max]	2.67 [0, 12]	2.64 [0, 12]
SD	± 2.41	± 2.46
UCLA		
Mean [Min, Max]	5.42 [3, 9]	5.48 [3, 9]
SD	± 2.04	± 2.02
GRIT-S		
Mean [Min, Max]	26.46 [8, 40]	25.94 [8, 40]
SD	± 6.08	± 6.12
WEMWBS		
Mean [Min, Max]	22.75 [7, 35]	22.78 [7, 35]
SD	± 5.96	± 6.25

Note: PHQ-9, Patient Health Questionnaire-9; GAD-7, Generalized Anxiety Disorder-7; PROMIS, Patient-Related Outcome Measurement Information Systems; WHODAS, World Health Organization Disability Assessment Schedule; AUDIT-C, Alcohol Use Disorders Identification Test; UCLA-Loneliness, University of California loneliness 3-item scale; GRIT-S, short version of the grit scale; WEMWBS, Warwick-Edinburgh Mental Well-Being Scale; Study design 1, the questions were presented to the participants containing all symptoms grouped by framing (i.e., frame priming); Study design 2, each symptom was asked within the five frames (i.e., symptom primed)

Table 2 - Framing invariance testing for each questionnaire

Study	Sample in each group	Model	Invariance	CFI	TLI	RMSEA	SRMR	Δ CFI	Δ RMSEA	Δ SRMR	Decision
1	1252	PHQ-9	Configural	0.999	0.999	0.045	0.041				
			Loadings	0.999	0.999	0.044	0.041	< 0.001	< 0.001	< 0.001	Invariant
			Thresholds	0.999	0.999	0.044	0.041	< 0.001	< 0.001	< 0.001	
		GAD-7	Configural	1.000	1.000	0.022	0.026				
			Loadings	1.000	1.000	0.023	0.026	< 0.001	0.001	< 0.001	Invariant
			Thresholds	1.000	1.000	0.026	0.026	< 0.001	0.003	< 0.001	
		PROMIS Depression	Configural	0.998	0.997	0.083	0.046				
			Loadings	0.998	0.997	0.081	0.046	< 0.001	0.002	< 0.001	Invariant
			Thresholds	0.998	0.998	0.078	0.046	< 0.001	0.003	< 0.001	
		PROMIS Anxeity	Configural	0.999	0.999	0.046	0.024				
			Loadings	0.999	0.999	0.045	0.024	< 0.001	0.001	< 0.001	Invariant
			Thresholds	0.999	0.999	0.044	0.024	< 0.001	0.001	< 0.001	
2	565	PHQ-9	Configural	1.000	1.000	0.033	0.046				
			Loadings	1.000	1.000	0.034	0.046	< 0.001	0.002	< 0.001	Invariant
			Thresholds	0.999	0.999	0.038	0.046	< 0.001	0.004	< 0.001	
		GAD-7	Configural	1.000	1.000	0.009	0.032				
			Loadings	1.000	1.000	0.037	0.041	< 0.001	0.029	0.008	Invariant
			Thresholds	1.000	1.000	0.044	0.040	< 0.001	0.007	< 0.001	

Note: Invariance decision is based on ΔCFI < 0.010 supplemented by ΔRMSEA < 0.015 or ΔSRMR < 0.010. RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; SRMR, Standardized Root Mean-square Residual; Δ, differences between fit index.

Table 3 - Mean total score difference by question frame (frequency as reference) calculated using covariate- and study design-adjusted and weighted regression models

	PHQ-9		GAD-7		PROMIS Depression		PROMIS Anxiety	
	Mean difference	95% CI	Mean difference	95% CI	Mean difference	95% CI	Mean difference	95% CI
Question frame (ref. Frequency)								
Context	-0.41	[-0.93, 0.11]	-0.67 **	[-1.09, -0.24]	-1.48 ***	[-2.07, -0.88]	-1.26 ***	[-1.88, -0.65]
Ability	1.16 ***	[0.64, 1.68]	1.01 ***	[0.58, 1.44]	1.39 ***	[0.79, 1.98]	1.85 ***	[1.24, 2.46]
Duration	1.26 ***	[0.74, 1.78]	0.54 *	[0.11, 0.97]	-0.56	[-1.15, 0.04]	-0.88 **	[-1.49, -0.27]
Bothered	0.56 *	[0.04, 1.09]	0.40	[-0.03, 0.83]	0.37	[-0.22, 0.97]	0.03	[-0.59, 0.64]
Study design and frame by design interaction								
Study desing (ref. SD1)	0.56	[-0.11, 1.24]	0.12	[-0.43, 0.68]				
Context/Frequency by Study	-0.32	[-1.27, 0.64]	-0.12	[-0.91, 0.66]				
Ability/Frequency by Study	0.05	[-0.91, 1.01]	-0.32	[-1.11, 0.46]				
Duration/Frequency by Study	0.46	[-0.49, 1.42]	0.17	[-0.61, 0.96]				
Bothered/Frequency by Study	-0.12	[-1.07, 0.84]	-0.17	[-0.96, 0.61]				
Covariates								
Age (continuous)	-0.10 ***	[-0.11, -0.09]	-0.08 ***	[-0.09, -0.07]	-0.12 ***	[-0.13, -0.10]	-0.12 ***	[-0.14, -0.10]
Sex (ref. Female)	-2.85 ***	[-3.13, -2.57]	-2.32 ***	[-2.55, -2.09]	-1.77 ***	[-2.16, -1.37]	-3.01 ***	[-3.43, -2.60]
Family income (D/E)	-1.58 ***	[-2.03, -1.13]	-1.10 ***	[-1.47, -0.73]	-1.60 ***	[-2.24, -0.97]	-1.19 ***	[-1.84, -0.53]
Family income (C/E)	-2.27 ***	[-2.71, -1.84]	-1.43 ***	[-1.79, -1.07]	-1.83 ***	[-2.45, -1.20]	-1.62 ***	[-2.26, -0.98]
Family income (B/E)	-3.34 ***	[-3.83, -2.86]	-2.25 ***	[-2.65, -1.84]	-3.34 ***	[-4.04, -2.65]	-3.40 ***	[-4.12, -2.69]
Family income (A/E)	-4.45 ***	[-5.09, -3.80]	-3.09 ***	[-3.62, -2.56]	-4.86 ***	[-5.75, -3.96]	-4.23 ***	[-5.15, -3.30]

Note: PHQ-9, Patient Health Questionnaire-9; GAD-7, Generalized Anxiety Disorder-7; SD1, study design 1, the questions were presented to the participants containing all symptoms grouped by framing (i.e., frame priming); SD2, study design 2, each symptom was asked within the five frames (i.e., symptom primed); *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 4 - Bifactor model-based reliability indices

Scale	Factor (general or framing)	Omega	OmegaH	H	FD
PHQ-9	General factor	0.984	0.975	0.989	0.994
	Frequency	0.925	0.014	0.264	0.782
	Context	0.927	0.006	0.289	0.748
	Ability	0.928	0.037	0.307	0.753
	Duration	0.925	0.021	0.297	0.746
	Bothered	0.925	0.078	0.404	0.829
GAD-7	General factor	0.984	0.972	0.989	0.990
	Frequency	0.932	0.034	0.201	0.690
	Context	0.931	0.037	0.191	0.657
	Ability	0.929	0.037	0.213	0.694
	Duration	0.930	0.052	0.246	0.690
	Bothered	0.931	0.077	0.345	0.802
PROMIS Depression	General factor	0.991	0.975	0.992	0.996
	Frequency	0.960	0.007	0.309	0.861
	Context	0.960	0.002	0.315	0.862
	Ability	0.955	0.012	0.388	0.889
	Duration	0.961	0.105	0.476	0.887
	Bothered	0.958	0.120	0.514	0.899
PROMIS Anxiety	General factor	0.992	0.973	0.992	0.992
	Frequency	0.960	0.029	0.246	0.733
	Context	0.963	0.047	0.289	0.758
	Ability	0.960	0.063	0.355	0.797
	Duration	0.962	0.112	0.477	0.864
	Bothered	0.966	0.096	0.442	0.855

Note: PHQ-9, Patient Health Questionnaire-9; GAD-7, Generalized Anxiety Disorder-7; OmegaH, omega-hierarchical; H, index of construct replicability (> 0.8 suggests a well-defined latent variable); FD, factor determinacy (> 0.9 indicate that the factor score can be used).

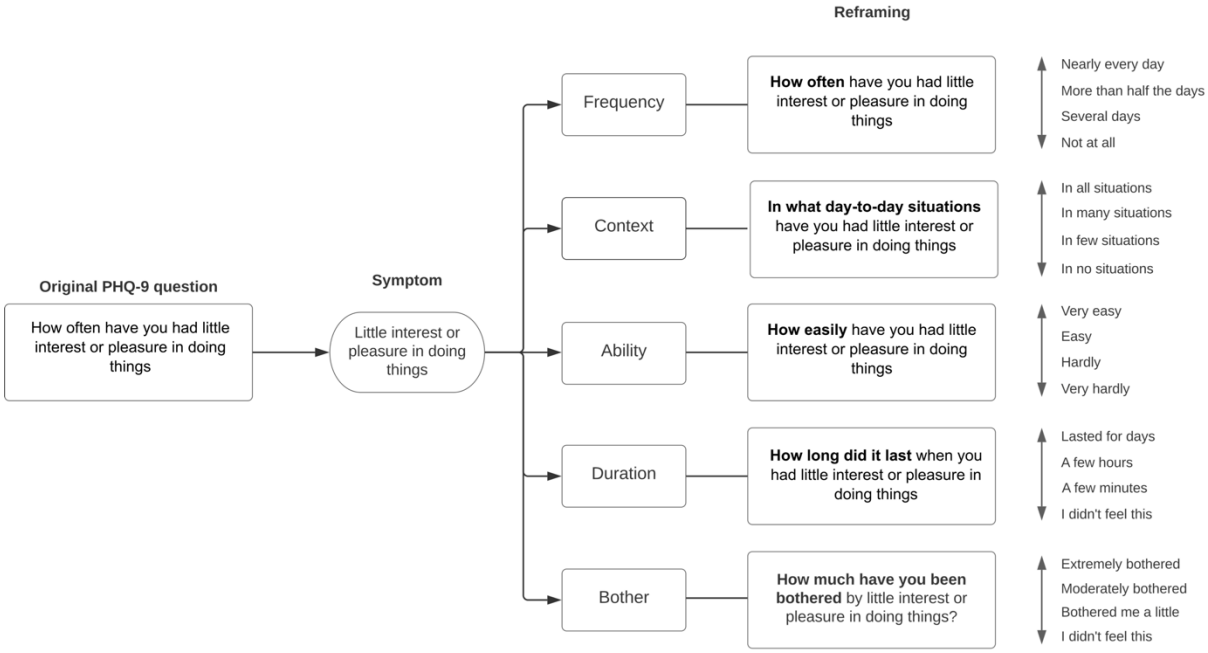

Figure 1. Question framing process. Example is given with a PHQ-9 question regarding "little interest or pleasure in doing things". Original questions were in Brazilian Portuguese and PHQ-9 and GAD-7 do not contain the expression "have you been bothered by" in the original frequency-framed questions.

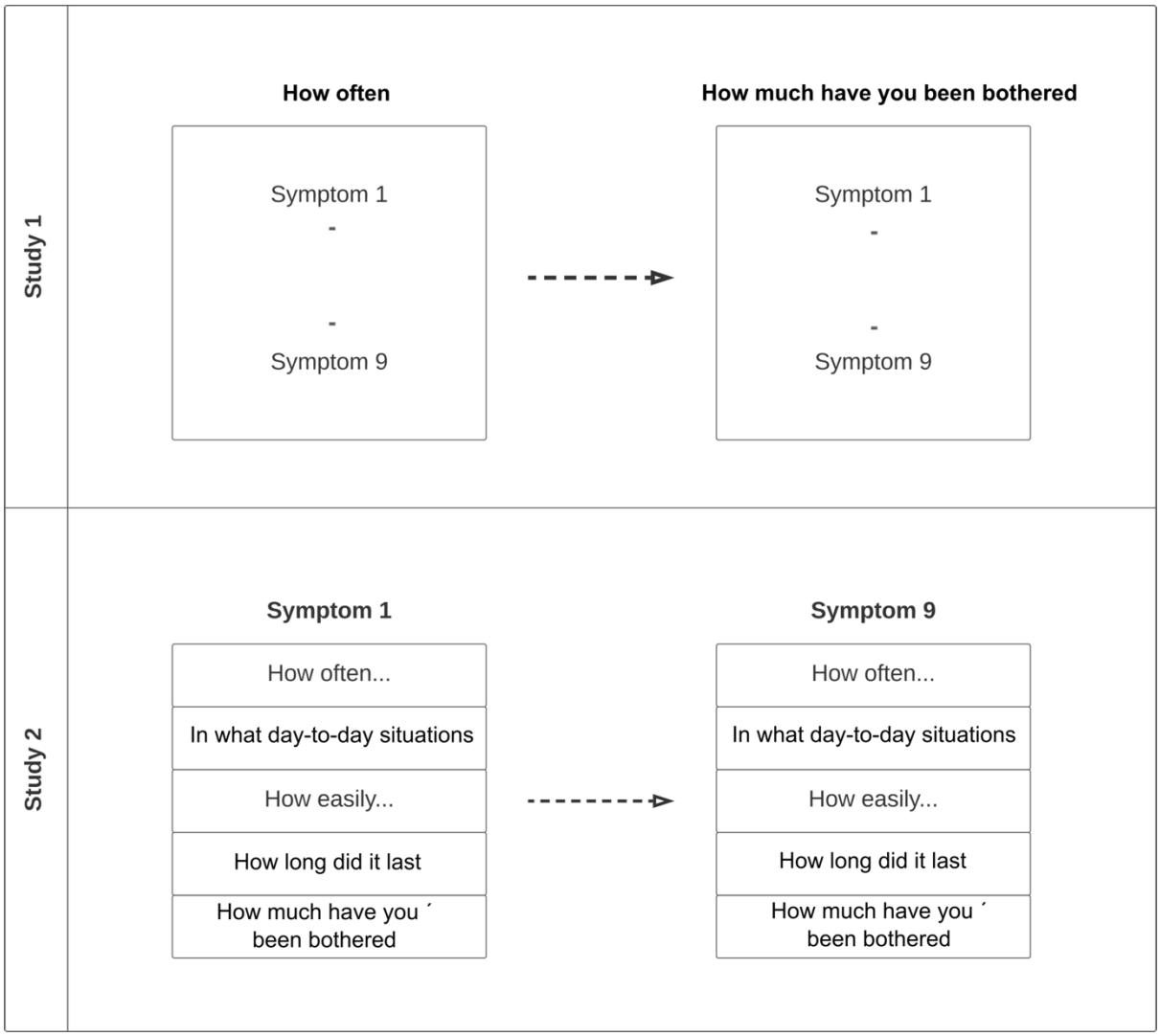

Figure 2. Question presentation for two academic subsamples. Study design 1 (SD1): The questions were presented to the participants containing all symptoms grouped by framing (i.e. frame priming). Study design 2 (SD2): Each symptom was asked within the five frames (i.e. symptom primed).

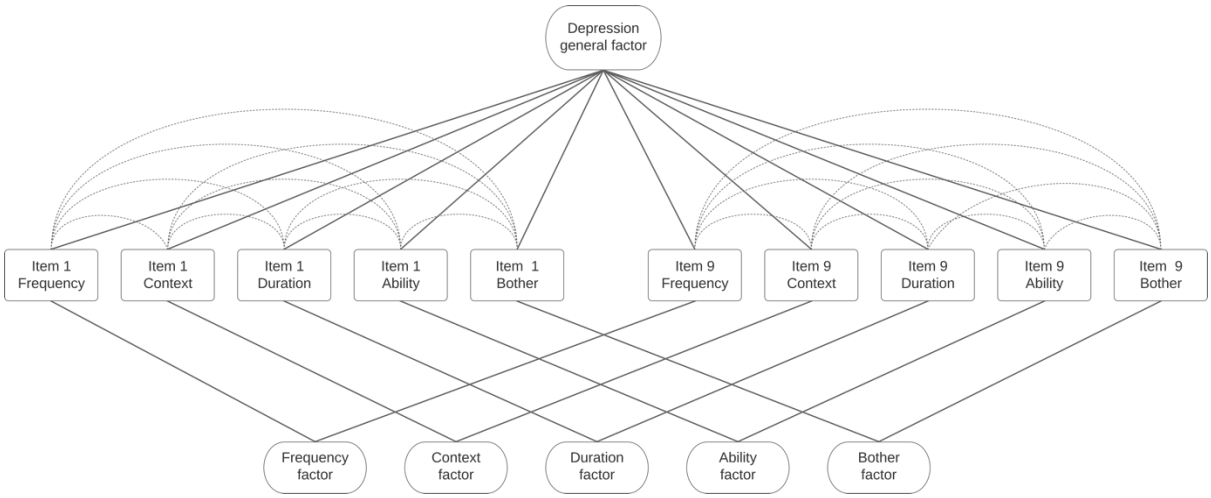

Figure 3. Bifactor model for PHQ-9.

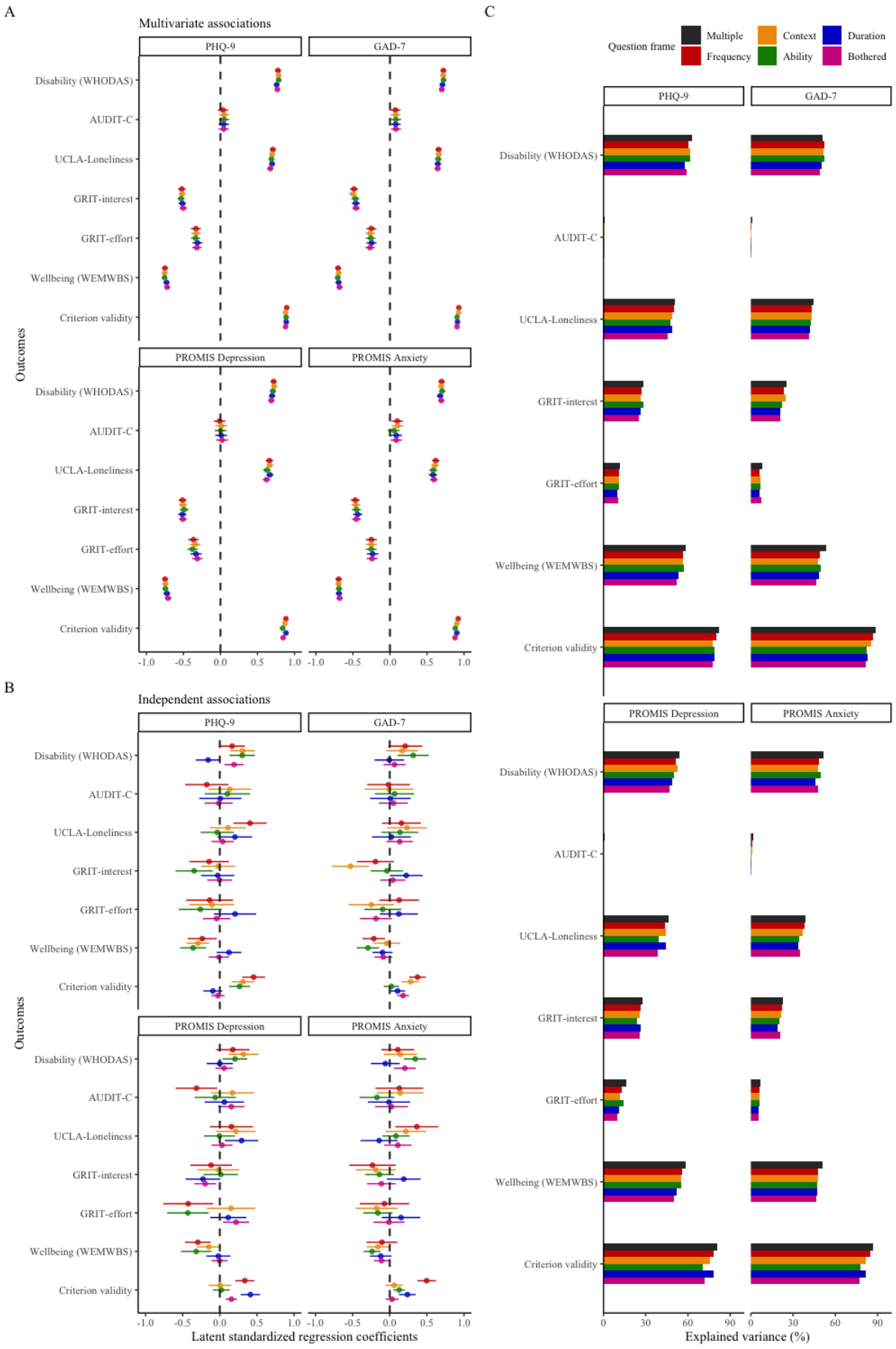
A
Multivariate associations
PHQ-9
GAD-7
PROMIS Depression
PROMIS Anxiety
Outcomes
Disability (WHODAS)
AUDIT-C
UCLA-Loneliness
GRIT-interest
GRIT-effort
Wellbeing (WEMWBS)
Criterion validity
-1.0
-0.5
0.0
0.5
1.0
B
Independent associations
Latent standardized regression coefficients
C
Question frame
Multiple
Frequency
Context
Ability
Duration
Bothered
0
30
60
90
Explained variance (%)

Figure 4. A: Multivariate structural equation models (SEM), regressing all validators (y-axis) on separated predictors (questionnaires framed as frequency, context, ability, duration and botherement) for PHQ-9, GAD-7, PROMIS depression and anxiety. B: Multiple SEM egressing all validators on all predictors for each questionnaire. C: Explained variance for multivariate and multiple regression models.