UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Thalmayer, A. G., Marshall, J., & Scalise, K. (2023). The International Mental Health Assessment: Validation of an Efficient Screening Inventory. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.74546

Clinical Psychology

The International Mental Health Assessment: Validation of an Efficient Screening Inventory

Amber Gayle Thalmayer[1][a], Julie Marshall[2][b], Kathleen Scalise[3][c]

[1] Department of Psychology, University of Zürich, Zürich, Switzerland, [2] Canopy Wellbeing, Portland, Oregon , [3] Department of Methodology, Policy, and Leadership, College of Education, University of Oregon, Eugene, Oregon

Keywords: inventory construction, reference group effects, internalizing and externalizing, hierarchical taxonomy, psychological disorders, item response theory, HiTOP, Partial credit model, employee assessment, screening for prevention

https://doi.org/10.1525/collabra.74546

Collabra: Psychology

Vol. 9, Issue 1, 2023

The International Mental Health Assessment (IMHA) was developed to provide efficient screening to facilitate prevention and early intervention among employees or community adults at three levels of analysis: a P-factor of general functioning and tendency toward disorder; broad spectra of internalizing and externalizing tendencies and for life difficulties; and nine subscales for common, familiar psychological and behavioral health categories. This study describes the development, refinement, and validation of the inventory using item response theory (IRT), specifically the partial credit model (PCM). Explicit, behavior-focused items drew on commonalities among domain-specific inventories, the *DSM-V* and empirical literature. A response scale based on concrete frequency of occurrence over the last month was developed to avoid the reference-group effects that plague cross-group survey research, facilitating cross-group comparison at both scale and item levels. In Study 1, a preliminary 69-item version was administered to 5,307 employees, family members, and counseling clients. PCM calibration was used to remove items with overlapping discrimination or unclear scale correspondence. In Study 2, the refined 59-item IMHA was administered to 4,048 employees. In Study 3, the subscales were compared to relevant established inventories to assess and confirm their convergent/divergent validity in a third sample (N = 500). The final 54-item IMHA, intended both for screening for psychological problems among community adults and to facilitate research including cross-cultural and cross-group comparisons, is made available freely for educational, non-profit or research purposes. The three-level measurement strategy draws on recent evidence for the continuous nature of psychopathology and on the well-established co-morbidity of traditional disorder categories, making use of them for communication purposes without unnecessarily reifying them in the model.

Prevention and early intervention of psychological difficulties can spare individuals prolonged suffering and can maintain an effective workforce. Depression, for example, is the single largest contributor to disability worldwide, and anxiety disorders are in sixth place (World Health Organization, 2017). The current project describes the creation and validation of an efficient inventory to assess for common psychological and behavioral health problems among community adults, resulting from an effort by a North American employee assistance program (EAP)[1] to develop an online assessment to direct at-risk employees to appro-

a Correspondence should be addressed to: Amber Gayle Thalmayer, Department of Psychology, University of Zurich, Binzmühlestrasse 14/19, 8050 Zürich, Switzerland, ambergayle@gmail.com. https://psychologie.uzh.ch/de/bereiche/sob/psyges.html

b https://canopywell.com/

c https://education.uoregon.edu/directory/faculty/all/kscalise

1 Access to an employee assistance program (EAP) is provided by some employers as a benefit alongside workplace-based health insurance. In the United States, like health-insurance providers, EAPs are typically for-profit companies, which offer a range of services to members, for example including workplace wellness programs, mental health counseling, and financial, legal, and conflict resolution consultation services.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

priate services, and to help employers tailor their wellness and prevention efforts to the current needs of their staff.

The desired inventory for this purpose would be targeted to a normal adult population, meaning that it should be as brief as possible to encourage completion from most, not only from those actively curious about mental health, and to minimize the emotional stress that answering questions about psychological disorders can invoke (e.g. Labott et al., 2013). It would include subscales for multiple common problems, to allow for feedback related to familiar domains (e.g. anxiety, depression, substance abuse) and to facilitate communication with counselors. However, it would also be shaped by the overwhelming evidence for the co-morbidity of common disorders and allow for higher-level, dimensional scoring, as defined by the Hierarchical Taxonomy of Psychopathology (HiTOP) consortium (Conway et al., 2019; Kotov et al., 2017), i.e. providing a meaningful global score and scores for broad spectra. As much as possible, items would refer to specific behaviors and experiences of distress and impairment in order to reduce response biases and to access maladaptive problems rather than self-concepts or personality traits (Hopwood et al., 2022). Finally, the desired inventory would gather concrete information about typical experiences of different symptoms to facilitate comparison and research, by avoiding vague response options, i.e., subjective levels of relative frequency, such as "often" or "sometimes", which conflate moral judgements and subjective impressions with functional impairment (e.g. Schaeffer, 1991), and make inventories highly subject to reference group effects (Heine et al., 2002; Van de Gaer et al., 2012), inhibiting comparison across groups, or even indivudals.

Because no such inventory could be identified, we took up the challenge of creating it, drawing on a broad base of empirical literature and the strengths of many existing domain-specific inventories. This private-sector and academic partnership led to the creation of both a proprietary product used by Canopy Wellbeing (The WholeLife Scale), and to the International Mental Health Assessment (IMHA), an inventory that is freely available for non-commercial research use and allows for efficient assessment at multiple levels of analysis.

A Dimensional and Integrated Approach to Assessment

From the outset, based on the needs of EAP counselors and for feedback to respondents, we expected to include common domains of disorders and difficulties, including depression, anxiety, post-traumatic stress, substance abuse, anger, sleep problems, and interpersonal conflict. A number of well-validated inventories exist to measure these specific problems, an empirical literature that forms the backbone of the development of the IMHA, as described below. But these narrower inventories are not intended for screening in a non-clinical population, and deploying several as a group would require many items, often overlapping in content, and with a cacophony of response options.

Such an approach would also ignore the evidence for the co-morbidity of common disorders and over-emphasize categorial distinctions (Conway et al., 2019; Kotov et al., 2017). Symptoms across dozens of diagnoses have been shown to aggregate into overarching spectra (externalizing, internalizing, psychotic experience; Caspi & Moffitt, 2018; Conway et al., 2019; Kotov et al., 2017) which in turn have been shown to aggregate into a general dimension of psychopathology (Caspi & Moffitt, 2018; Conway et al., 2019). High scores on general liability associate strongly with a family history of psychiatric illness, brain function, developmental history, and life impairment (Caspi & Moffitt, 2018), and with underlying genetic vulnerability for psychopathology (Pettersson et al., 2016; Selzam et al., 2018). Thus, for efficiency and to reflect a growing scientific consensus, a dimensional approach was preferred, with familiar categories nested into broader spectra and a meaningful overall score.

Current Broad Inventories for Psychological Problems

Existing broad inventories, for example the Symptom Checklist and Brief Symptom Inventory (BSI) platform (e.g. Derogatis & Derogatis, 2001) and the Behavior and Symptom Identification Scale (BASIS-32; Eisen et al., 1999), are designed for use in medical settings and clinical intake, not for screening in the workplace. Furthermore, both inventories assess perceptions of difficulty (BASIS) or being bothered by symptoms (SCL), which conflate presence of a symptom with social and cultural norms and expectations, including ideas about how frequent something ought to be. While this approach may be justifiable in a clinical setting, where distress is a key factor, it inhibits the comparison of experiences across individuals or groups on any factor other than 'perceived distress', which may stem from specific mental problems, from a lack of social support in the context of normal challenges, or from cultural or role expectations. This is discussed in more detail below.

Inventories aimed at a general population include the Outcome Questionnaire-45.2 (OQ-45; Lambert et al., 2004), a measure of symptoms designed to assess functioning and change during clinical treatment, which has been translated into many languages and is used at clinics around the world. It aims to measure three broad categories (Symptom Distress, Interpersonal Relations, and Social Role Functioning), but no later studies have replicated the intended structure (e.g. Thalmayer, 2015) and the length of 45 items is unnecessarily long for a survey primarily validated for use as a single, total score. The broad Adult Behavior checklist (ABCL; Achenbach & Rescorla, 2003) provides scores on 24 scales within four domains, and has also been translated into many languages, but at 126 items it is problematically long. People vary in their willingness to complete surveys, and shorter measures can have advantages in terms of validity, for example by reducing boredom and fatigue (Burisch, 1984; Goring et al., 2004), increasing response rates and reducing costs (Edwards et al., 2004), potentailly without reducing predictive validity (Kemper et al., 2019; Thalmayer et al., 2011). Both inventories use vague, relative response options. For example, in the ABCL respondents are asked to rate their family relations as worse than

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

average, average, or better than average. But expectations will vary around 'average' family relations, and few will be well-informed about typical levels of family conflict or closeness across their society.

Similar to the current project's goals are two proprietary instruments, the 120-item, 10-subscale, Employee Assistance Program Inventory (EAPI; Anton & Reed, 1994), and the 96-item Spectra Indices of Psychopathology, which provides hierarchical-dimensional assessment (Blais & Sinclair, 2019). Both are intended for counseling intake and are longer than ideal for community and employee samples, and their proprietary nature inhibits collaborative scientific uses. The first was only available on paper, making it unsuitable for online and app-based testing. A shorter candidate is the 38-item Mental Health Inventory (MHI; Veit & Ware, 1983), which provides overall scores on distress and well-being, and five subscales. This inventory has the advantage of brevity, but excludes most content related to 'externalizing' problems (e.g. substance abuse, anger, conflict; Conway et al., 2019), and all three inventories use vague response options (e.g. not at all, a little bit, moderately, quite a bit, or completely true).

Response Scales and Reference Group Effects

A common problem across almost all inventories for psychological disorder symptoms is vague response scales. When not focusing narrowly on perceived difficulty, inventories typically ask for subjective, relative frequency: Did you experience this symptom "rarely", "sometimes", "frequently", or "almost always" (OQ-45)? Substantial psychometric work during the rise of survey research in the 20th century compared relative versus absolute frequency choices (reviewed by Friedman & Amoo, 1999 and Schaeffer, 1991). While many problems with relative frequencies have been demonstrated (described below), assessing absolute occurrence was argued to be more demanding on participants, especially in the case of vague, subjective events (Bradburn & Miles, 1979). It is also probably relevant that survey data is analyzed solely using the numbers assigned to options, and may have come to put less weight on the wording than is warranted. While the numbers on a Likert scale suggest interval scaling, which is generally how analysis proceeds, descriptive terms are not necessarily perceived as equal steps (Brown, 2011; Friedman & Amoo, 1999; Loevinger, 1957; Schriesheim & Novelli, 1989). Despite these issues, the norm for clinical scales to use relative qualifiers has become so strong that consideration of this choice is now generally absent from reports on new inventories.

We find this norm problematic for many reasons. First, there is a lack of precision as to what is being measured – the occurrence of a symptom, or a feeling or judgement about it? Ratings of relative frequency that use vague qualifiers or intensifiers focus comparisons on an internal, implicit standard and are thus shaped by many factors beyond frequency (Schaeffer, 1991). Consider the difference between answering whether you lost your temper last month or whether you lose your temper "often". The former is a reasonably factual question, referring to an experience that is visible to others as well as oneself, while the latter may feel like a moral judgement: "Often" implies that you lose your temper more than is "normal" or "average", but compared to whom? Relative-frequency ratings have also been shown to be influenced by expected frequency, with different standards for rare events (earthquakes) versus frequent ones (rain), as well as by valence: Something unpleasant that occurs three times a week will be described with a qualifier indicating more frequency than something pleasant occurring at the same rate (Schaeffer, 1991).

Secondly, there is the difficulty of comparing scores between groups. Relative frequency is shaped by social norms and expectations, leading, for example, to differences among subgroups in how many days of the month are associated with "very often", "pretty often" and "not too often" (Bradburn & Miles, 1979). Vague options are also vulnerable to response styles, such as acquiesce bias (a general tendency to agree to survey items) or extremeness in responding (sticking to the low and high ends), both of which have been shown to covary with cultural traits (e.g. power distance, individualistic values, uncertainty avoidance; Johnson et al., 2005), which vary between nations, as well as between individuals.

Further complicating cross-group comparisons is the fact that vague response scales are subject to reference group effects (e.g. Heine et al., 2002; Van de Gaer et al., 2012). For visible traits, such as Conscientiousness, people can only self-report tendencies assessed with vague qualifiers by comparing themselves to those they know: classmates, colleagues, friends, family, and acquaintances. They are limited by social circles and self-evaluations are strongly shaped by local expectations. This leads to counterintuitive findings, for example, national mean scores for self-reported Conscientiousness correlate poorly with objective society-level indicators (e.g. accuracy of public clocks, efficiency of postal workers, public-sector corruption; Heine et al., 2008; Oishi & Roth, 2009), and students from the world's best schools in terms of academic achievement rate themselves lower on academic talent than those from low-performing schools (Shen & Tam, 2008).

This measurement challenge is further complicated in the case of psychological disorder symptoms, which are less visible. How depressed or anxious someone feels, or how well they sleep, is not obvious to an outside observer. Most people will know about such experiences only for intimate friends and family. Thus, what does it mean when an individual answers that they worry "a lot" or "seldom"? Only that this is the true in comparison to a few close others, or in relation to local assumptions, for example media depictions of 'normal' functioning.

The problems of vague response scales are avoided in the International Mental Health Assessment (IMHA) by using a response scale referring to specific frequencies within the last month. While testing the cross-cultural applicability of this inventory is beyond the scope of this report on the development of the inventory, this value was taken into consideration from the outset, to maximize validity in the diverse society of the United States, and for comparing across groups, be they cultural, subcultural, regional, or by gen-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

der, age, social class, education level, etc. The impact of response biases is also addressed by using items that are as concrete and behavioral as possible (e.g. "I had difficulty falling asleep", "I argued with friends or family", "I drank enough to pass out") to insure state rather than trait assessment (Hopwood et al., 2022). Specific items also minimize the key advantage of relative-frequency ratings, that they reduce respondent effort in the case of hard-to-quantify experiences. Note that concrete specificity does not preclude assessing for distress or impairment, as in: "Not getting enough sleep interfered with my daily activities", "I had difficulty making decisions" and "My worrying got in the way of doing something I intended to do". Over the long-term, concrete items and a specific, absolute response scale are intended to facilitate the establishment of cross-cultural measurement invariance to allow for cross-group, cross-cultural comparisons (e.g. Fischer & Karl, 2019), and it creates the potential for meaningful item-level comparisons across individuals or groups even where such measurement invariance cannot be established.

The Current Study

The aim of this project was to follow a multi-step, iterative process to create and validate a comprehensive but maximally efficient inventory of common psychological problems, integrating the strengths of existing domain-specific measures (in terms of item content) and the values of hierarchical (spectra) assessment, and using a concrete response format less vulnerable to reference group effects and thus suited for assessment and comparison across groups. Allowing common cross-syndrome symptoms to serve as overall screeners was a means of keeping the scale as brief as possible, to reduce the stress of completing it, and to increase completion rate (e.g. Deutskens et al., 2004) and thus utility as a screening tool.

Study 1 includes the creation of a preliminary 69-item version and its administration to a large group of employees and a sample of clients presenting for services at the EAP's counseling office, in order to increase the base rate of complaints and to allow for comparison as an initial assessment of validity. IRT analyses using the Rasch Partial Credit Model (PCM) emphasized improving the inventory to meet the goals of reliable and internally-valid subscales and Total Score, good distinction between domains (reducing subscale intercorrelations), and minimizing length. Study 2 includes the administration of the refined 48- to 59-item version to a second large sample of employees, which was split into two random halves. Analyses in the first split-half focused on refinement to meet the same criteria, with exploration at p-factor, spectra, and subscale levels. The split-half procedure allowed for testing a refined version in the second half; calibration of the final version of the IMHA and its psychometric properties are reported for the full sample. In Study 3, the IMHA subscales were compared to 21 scales for relevant subdomains, belonging to 15 established inventories, to assess convergent and divergent validity.

Study 1: Create, Test and Refine a Preliminary Version of the CHMA

Methods

Only anonymous survey data were made available by Canopy to the study authors. The Committee for the Protection of Human Subjects of the Research Compliance Services of the University of Oregon found this use of existing data, which was collected following strict HIPAA (Health Insurance Portability and Accountability Act) guidelines, to be exempt from the need for full institutional review (2019).

The raw data and scripts for running the analyses are available on the Open Science Framework (https://osf.io/vysj4/). Sufficient information is provided for an independent researcher to reproduce all reported results (Open Data) and all reported methodology (Open Materials).

Materials. The preliminary IMHA was created by identifying common psychological and behavioral health problems among community adult and employee populations: alcohol and drug abuse, anxiety, depression, post-traumatic stress, sleep problems, life stress, work disengagement, anger, interpersonal conflict, and protective factors. Prevalence rates, drawn primarily from the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–V*; American Psychiatric Association, 2013), were an important criteria in choosing domains for inclusion. Other domains (e.g. eating disorders, psychotic experiences, the identity disturbances of borderline personality disorder) were considered but excluded, because these less prevalent conditions can be assumed to share symptoms or be comorbid with more common problems such as anxiety and depression (e.g. Conway et al., 2019; Eaton et al., 2010; Kotov et al., 2017). Allowing common cross-syndrome symptoms to serve as screeners was a way to keep the inventory brief. Although we aimed to make reliable distinctions between domains, and retained them to facilitate psycho-education and communication with clinicians, the subscales were expected from the outset to correlate with each other due to established patterns of covariance among disorder categories (Conway et al., 2019). Spectra for internalizing and externalizing tendencies and a total score were planned to provide assessment of broader domains of psychological difficulties, allowing for three levels of analysis. This hierarchical approach allows clinicians to use familiar terms while raising consciousness about their overlap.

To make best use of the large body of empirical literature on clinical assessment, the strategy for creation of the preliminary IMHA was to seek content convergence among validated inventories for each domain of interest. For each domain, three to five inventories were identified (detailed in Supplemental Table S1). Public domain inventories were favored and viewed at the item level. In some cases, proprietary instruments were viewed, in this case with a focus only on content areas. Inventories were compared to each other, to clinical criteria in the the *DSM–V* (American Psychiatric Association, 2013), and to empirical studies of the constructs in order to define regularities in key content. For example, for post-traumatic stress five inventories were viewed. At least three included items about 13 symptoms:

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

intrusive memories, reliving, sleep disruptions, irritability, emotional suppression, numbness, dissociation, avoidance, physiological arousal, flooding, concentration difficulties, nightmares, and vigilance. Other symptoms, like memory loss, loss of interest, and pessimism were included on fewer inventories and are less specific to post-traumatic stress. Of the core 13, sleep problems, irritability, and concentration items were already included for other IMHA domains. IMHA items were thus developed for the remaining 10 symptoms, focusing on specific behaviors and feelings and fit to the response scale, following best practices in item writing (e.g. Clark & Watson, 1995). The initial version was intentionally over-comprehensive, including more items than expected to be necessary for the final version (Clark & Watson, 1995).

The response scale was developed to refer to specific frequency in order to minimize reference group effects. Respondents are asked to recall how often a behavior or feeling occurred in the last month, and to answer on a 7-point scale of frequency: daily, half the days, about twice a week, about once a week, about twice a month, monthly, not in the last month. (The choice of these options was also guided by their potential to correspond to an interval scale with a natural log transformation if quantified as days per month, though this was not done in the analyses). The Protective Factors items were reverse-scored for the unidimensional model.

An initial draft inventory was reviewed by clinical psychologists and piloted with individuals who described their perceptions and reactions while or after completing it. While some noted appreciation in seeing a symptom on the scale (e.g. intrusive thoughts) made them feel less alone, in general they reported that the questions were 'heavy' and reminded them of the 'dark' side of life and they found it important to be as brief as possible. An *a priori* construct map (in online supplemental materials) details the expected meaning of scores on the Total Score and subscales based on clinical experience and existing surveys in each domain. For the pilot sample, feedback text was desired, so estimated cut-off scores for low, medium, and high risk were rationally developed by the first two authors and EAP psychologists, based on a face-valid understanding of the distress and impairment implied by different frequencies. The initial cut off scores and observed percentage of participants who fell into each category, relevant primarily for the feedback system used by Canopy Wellbeing, are presented in Supplemental Table S2.

Participants. A total of 5,307 individuals completed the preliminary survey, including 5,170 public employees of either a medium-sized west-coast city or a west-coast state, and 137 clients who presented for services at the counseling office of the EAP. Multiple completions by the same person and responses from anyone under age 18 were excluded. Age, gender, and ethnicity information for the two Study 1 samples are reported in the left-hand side of Table 1. For this study, we used all cases available in the existing data, assuming the sample size was more than sufficient for the planned PCM analyses, following De Ayala (2009), who suggests a loose rule of thumb of about 250 for Partial Credit models, though understanding that this is not precise and that formal power analysis is not often done for IRT modeling (Zimmer et al., 2022).

Procedure. The EAP made the pilot version of the survey available free-of-charge to two employers, with the understanding that anonymous information from their employees would be used to test and develop the IMHA and the full product in which it is embedded, which includes a proprietary interface and feedback protocols. Emails from the EAP to employees of these organizations notified them of the opportunity to complete the survey and to receive individualized feedback. A report that appeared after survey completion categorized each respondent in terms of low-, medium-, and high-risk categories for each of 10 domains, and included links to articles about self-care and the phone number for EAP counseling services. Confirmatory answers to three 'red flag' items that indicated potentially violent feelings or being a victim of violence led to immediate feedback urging a call to the counseling center. Aggregate, anonymized summaries were delivered to employers, advising them about general areas of concern for their employee population.

Analyses. IRT estimation relied on Rasch family partial credit models (PCM2; Andrich, 1978; Masters, 1982; Rasch 1960/1980). The unidimensional model was tested in R package Test Analysis Modules (TAM; Kiefer et al., 2017) using marginal maximum likelihood estimation; multidimensional models in R package supplementary item response theory (sirt; Robitzsch & Robitzsch, 2020) also with marginal maximum likelihood estimation. For all models we report the number of 'steps' (six per item, given the seven-point response scale) with mean square weighted fit outside a ¾ - 4/3 tolerance (Adams et al., 1997; Adams & Khoo, 1996). Following those authors, our *a priori* standard was to avoid more than 5% misfit, as this would indicate that too many items correspond poorly to others in the set. Also assessed were PCM and Cronbach alpha reliability, and scale intercorrelations. Raw mean scores and PCM scaled scores (thetas) were compared between the employee and clinical samples. Models were compared using difference in deviance (χ^2), AIC and BIC (Schwarz, 1978). 'Wright Map' graphical representations (Wilson, 2005) were used to view the score distribution on the latent trait for individual items, to help identify redundant items and to aid scale refinement.

Results

Descriptive statistics. In the full sample, all options (0 - 6) were used for all items with two exceptions. The item "My partner physically hurt me" had no responses beyond "monthly" (1). The item "I got so angry I had a physical altercation with someone" was not responded to above "about twice a week" (4). The six protective factors items had the highest mean scores, from 3.3 to 4.7. Raw means for the other items ranged from 0.01 to 3.08 (reported in Supplemental Table S2.)

PCM Estimation and Exploration for Refinement. The unidimensional Rasch model had high reliability (.95; personal separation .93; Cronbach alpha .96). However, 11% of

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 1. Sample Characteristics for All Studies

	Study 1			Study 2	Study 3
	Employee	Clinical	Total	Employee	Prolific
Sample size	5,170	137	5,307	4,048	500
Age range	18 - 76	18 - 75	18 -76	18-79	18-70
Mean	48.6	42.2	48.4	41.7	34.5
Standard Deviation	11.1	11.7	11.1	11.9	12.1
Gender = Woman	72%	66.4%	71.8%	68.8%	49.0%
Transgender or gender variant	.2%	0	.2%	.7%	0
Primary Ethnic Identity, %:					
Caucasian	71.5	80.9[2]	70.7	88.6[3]	70.6
Hispanic	-[1]	-[1]	-[1]	5.0[3]	-[1]
Asian or Asian American	1.8	5.9[2]	1.9	.2[3]	9.8
American Indian/Alaska Native	.6	1.5[2]	2.1	.5[3]	-[1]
African American		2.6[2]	1.2	2.6[3]	11.8
Native Hawaiian/Pacific Islander	1.2	3[2]	.6	.5[3]	-[1]
Other	2.1	8.8[2]	6.0	2.7[3]	1.4
Mixed	-[1]	-[1]	-[1]	-[1]	5.4
Missing/prefer not to answer	16.6	50.4	17.5	67.2	1

[1] Not a response option for this sample.
[2] This percentage is among the 49.6% of respondents (n = 68) who answered this item.
[3] This percentage is among the 32.8% of respondents (n = 1,328) who answered this item.

the 413 item-steps had mean square weighted fit outside the recommend tolerance, indicating lack of correspondence with other items in the inventory. The problematic steps were for three substance-use items: one ("I had 3 or more alcoholic beverages in a day") may not relate closely to psychological disorder symptoms; two were very infrequently endorsed. There was also problematic fit for five of the six protective factor items.

Model fit was better for a 10-dimension Rasch model, χ^2 (63) = 27,244, p < .01, person separation reliability (details reported in Supplemental Table S3). Only five items (< 5%) had steps outside the acceptable range. Reliability and intercorrelations among the scales are reported in Table 2. Two scales had low IRT reliabilities, and three low Cronbach alpha. Not surprisingly, the internalizing subscales (Depression, Anxiety, Post-Traumatic Stress) correlated highly with each other (.90-.93). Mean thetas (PCM scaled scores) for the employee and clinical samples are graphically contrasted in Figure 1. The clinical sample had significantly higher scores on all scales, as expected.

Wright Map graphical representations of the score distribution on the latent trait were viewed for each subscale to aid scale refinement. These are included in Supplemental Figure S2 with detailed explanation and examples. While the histograms show normal distributions of the latent traits, the distinctions between steps are most reliable at the higher end. Given that this is a screener and that less than a tenth of the data came from a clinical population, this was seen as appropriate: most people from the community will have relatively low scores, and it is not efficient to make fine-grained distinctions between non-problematic score levels, e.g. between extremely low and very low depression scores. Instead, distinctions are most relevant at the higher end. These charts are also useful for identifying items that make the same distinctions and thus may be redundant. For example for Depression, four items, while covering different content, were seen to make the same distinctions in this sample, adding little incremental value.

Discussion

In Study 1, a preliminary 69-item version of the IMHA was developed to meet the goals of broad assessment of common psychological problems for use in a normal adult population, with concrete behavioral items and an objective response scale. It was administered to a large employee sample and a sample of counseling clients. PCM analyses were used to identify potential modifications to increase reliability and decrease scale intercorrelations and length.

The Total Score was seen to have good reliability: the items worked together to define overall psychological functioning, with the main exception of the Protective Factor items, which did not distinguish well between the trait levels defined by other items. This result was not surprising, given that this subscale included the inventory's only reverse-scored items, and that they were drawn from domains of literature (positive psychology, personality, values) outside clinical psychology. This subscale was included so that feedback could cover a positive domain and to lighten the experience of completing the survey. On balance, however, given the items' lack of correspondence with the IMHA's core content, keeping the assessment brief was deemed to be a higher priority, and this subscale was dropped.

For clinical reasons, we retained screener items (suicidal thoughts, partner violence) and the other nine subscales.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 2. Item Response Theory Scale Correlations and Reliabilities and Classical Test Theory Reliabilities for the Preliminary, 69-item Version of the IMHA

Scale (items)		Dep	2	3	4	5	6	7	8	9	10
2	Anxiety (10)	**.93**									
3	Post-Traumatic Stress (9)	**.91**	**.91**								
4	Sleep Problems (5)	.80	.83	.76							
5	Life Stress (5)	**.93**	**.92**	.89	.80						
6	Work-Disengagement (3)	.86	.84	.83	.71	**.90**					
7	Partner Conflict (4)	.67	.59	.62	.49	.62	.64				
8	Substance Abuse (8)	.37	.37	.39	.34	.34	.34	.37			
9	Anger (5)	.82	.76	.75	.63	.83	.81	.77	.24		
10	Protective Factors[1] (6)	.67	.60	.54	.51	.62	.60	.58	.27	.60	
	EAP reliability	.94	.92	.87	.89	.92	.84	.65	.55	.77	.81
	Cronbach α	.92	.87	.84	.84	.85	.80	.61	.70	.34	.78
Raw Mean, Employees (SD)		18.8 (16.6)	10.3 (10.0)	8.5 (9.7)	10.5 (7.7)	7.5 (6.9)	2.1 (3.6)	1.6 (3.3)	2.5 (5.7)	1.9 (2.9)	12.1 (8.2)
Raw Mean, Clinical (SD)		24.3 (19.0)	12.49 (10.4)	10.68 (10.6)	11.2 (7.4)	9.7 (8.1)	2.5 (3.6)	3.0 (4.9)	3.4 (7.1)	2.2 (3.2)	13.4 (8.8)
Mean Theta, Employees (SD)		.08 (.87)	.07 (.76)	.07 (.76)	.05 (.66)	.08 (.92)	.07 (.96)	.04 (.68)	-.01 (.48)	.06 (.69)	.03 (.46)
Mean Theta, Clinical (SD)		.33 (.87)	.28 (.72)	.28 (.77)	.17 (.63)	.36 (.92)	.35 (.94)	.28 (.70)	.19 (.58)	.25 (.70)	.17 (.44)

Note. Dep = Depression, which had 13 items. For raw scales, N = 3,854 for partner conflict scale. For other scales, N = 4,930 to 5,069. Correlations over .90 are bolded for emphasis.
[1] Items are reverse-scored to match overall content, and thus should be interpreted as "lack of".

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

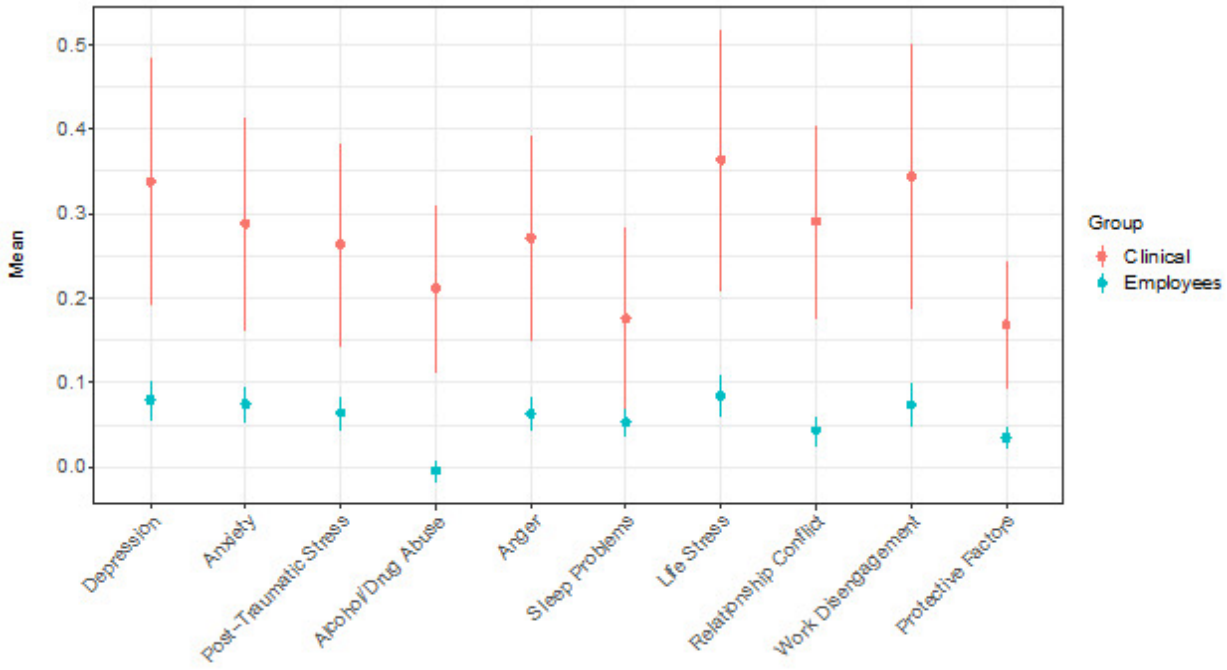

Figure 1. Study 1: Mean Subscale Thetas (Scaled Scores) for Employee and Clinical Samples

Note. 95% confidence intervals are shown. The protective factors scale is reverse scored, indicating lack of these resources. Between-group differences for all subscales in independent samples t-tests significant at $p < .01$, except Sleep Problems, $p = .034$.

Of these, the six subscales of the internalizing spectra had good reliabilities but relatively high intercorrelations. We sought to maximize distinctions, while accepting inevitable intercorrelation. For these scales, 23 items were removed based on (a) Wright Map indication of redundant coverage, and (b) in PCA, not loading highest on intended subscale, where removal did not reduce match to intended structure.

Three subscales, Substance Abuse, Anger, and Partner Conflict, had low reliabilities. Substance Abuse (EAP reliability .65; α =.61) was more categorical and less normally distributed than the other subscales, with relatively few people reporting problems. As the subscale appeared to conflate occasional social drinking with dependency, we separated it into use and abuse subcomponents, and added items about other drugs with the aim of increasing reliability. Likewise, Anger focused on quite disruptive behavior, which led to low endorsement and reliability (EAP reliability .77; α =.34). We addressed this by removing the most extreme item and adding an 'easier' item about feeling rather than expressing anger. For partner conflict (EAP reliability .65; α =.61), we added items to assess for interpersonal conflict beyond romantic relationships and added an item about being threatened by a partner to increase coverage and reliability, including the ability to screen for dangerous situations.

Study 2: Test and Refine a Final Version of the IMHA

Methods

Materials. The changes described above led to a refined 48-59 item version of the IMHA. The number of items varied by participant because for three subscales, Interpersonal conflict (3-8 items), Substance Use (4-6), Substance Abuse (0-6) some items were only shown based on responses to prior items. This approach reduced length to streamline responding, while maintaining coverage where it would be most useful. Additionally, three outcome variables were administered: Participants were asked if they were currently in treatment with a mental health professional; to rate their overall work performance on a scale of 1 to 10 for the last month; and to estimate the number hours of work missed due to personal concerns for the last month.

Participants & Procedure. A total of 4,048 individuals completed the survey on a proprietary smartphone app. These were employees and spouses from over two dozen medium-sized employers (with 50 to 500 employees) in the United States, predominantly located in the Pacific Northwest, contracting with Cascade Centers' for EAP services. Feedback was generated and delivered as described for Study 1 using updated protocols. Multiple completions by the same person and responses from anyone under age 18 were excluded from the data delivered to the authors of this report. Age, gender, and ethnicity are reported in Table 1. Because the app required answers to move through the survey and only completed surveys were delivered, there was no missing data. Eight substance abuse items were only asked if the participant endorsed any use, and empty answers were coded as zero. The 831 respondents who said that they were not in a romantic relationship were not asked five items specific to partner conflict.

Analyses. Analyses proceeded as for Study 1 in the first split half of the data, allowing for the possibility of dropping items and testing an interim version in the second

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

half. Here instead of a Total Score, spectra and super-spectra were used to better match scoring to theory. A three-spectra model was created based on typical patterns of comorbidity reported by Conway and colleagues (2019). Depression, Anxiety, and Post-Traumatic Stress were included on an Internalizing spectrum, Substance Use and Abuse and Anger on Externalizing, and the remaining scales (Sleep Problems, Life Stress, Interpersonal and Partner Conflict, Work Disengagement) on a Life Difficulties spectrum. P-factor included Internalizing and Externalizing but not Life Difficulties. The full sample was used for final calibration. Confirmatory factor analysis was performed on the final hierarchical model, with P-factor, spectra, and the associated five subscales. For the outcome questions, regression analyses were used to test associations with each scale, after taking into account the other subscales, age and gender. Logistic regression was used for the binary item about seeking mental health treatment. Poisson regression was used for work hours missed, as the skewed responses included many zeros.

Results

Descriptive statistics. To improve on the rationally derived cut-off scores used for feedback during piloting, we provide 'norms' in the form of percentiles in Table 3. Because of the skewed sample in favor of women and gender differences in scores (see below) these norms are separated by gender. The proportion of the sample who scored zero (no symptoms in the last month) on each scale and scores associated with the 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles are shown: i.e., the score that distinguishes the less symptomatic half of the sample from the half with more symptoms (50%), the score that indicates being in the lower three quarters of the sample (75%), and the lower 90% and 95%. Scores for an individual or a group from a similar population (e.g. a North American employee) can be compared to these norms. For example, a woman with a depression score of 27 could be understood to be in the top 10% for these symptoms, with a score higher than 90% of her peers.

In the full sample, raw item means ranged from 0.01 to 2.85. All options (0-6) were used for all items with one exception: "Family or friends suggested I should cut down on my drinking or drug use" had no responses in the "about twice a week" range. Six other items had one or more response options with five or fewer cases. For PCM analysis, responses were adjusted downward to create a minimum of five cases per cell, with the top response category left empty.

Partial Credit Model Estimation and Exploration for Refinement. In the first random half of the data, the P-factor had EAP reliability of .90 and 16 steps (7%) with fit outside 3/4 - 4/3 tolerances. All misfitting steps were from the five substance use items about using a specific intoxicant. Only the use item "I drank enough to feel intoxicated" had no misfitting steps. Indeed, the use of substances, while a pre-requisite for abuse, does not itself constitute a mental health problem, and we thus removed five items, retaining only that about intoxication. This 35-item P-factor in the second random half of the data had EAP reliability of .90 and only 4% of items with fit outside tolerances. Reliabilities for the three-spectra model using 54-items in the second random half were .93, .72, and .91, for Internalizing, Externalizing, and Life Difficulties, respectively, with no misfitting steps. The three spectra were highly intercorrelated (r_{IE} = .72, r_{IL} = .91, r_{EL} = .69), which is unsurprising given their shared fit to a total score.

Exploration of the multidimensional model started with all 59 items in the first half of the data with 11 maximally-disaggregated subscales, separating Interpersonal from Partner Conflict and Substance Use from Abuse. This model had less than 5% misfitting steps, but, not unexpectedly, the two conflict scales (EAP reliability .78 and .81) were correlated .90, and these were thus seen as better combined. Substance Use and Abuse (EAP reliabilities both .67) were correlated .88, though item correlations for use indicated only small associations between alcohol and drug items. Based on their lack of fit to the P-factor either empirically or conceptually, the five use items were dropped going forward.

Depression, Anxiety, and Post-Traumatic Stress correlated at or over .90 with each other. Principal components analysis and perusal of Wrights Maps (shown for the full data set in Supplemental Materials Figure S2) were used to search for items not loading on the intended scale or contributing additional coverage to the subscale, respectively. However, reliabilities under the IRT model were not improved by the removal of two candidates that were identified ("It was hard to control my worrying" and "I felt jumpy or easily startled") and intercorrelations were reduced only slightly. Given the acceptable length for practical purposes and the goal to adapt the inventory to other languages and contexts, where a larger item pool could have advantages, it was decided to retain the items. Anger was also explored, as its reliability was low (EAP .71). An exploration of interitem correlations suggested no candidates for removal that could improve Cronbach Alpha, and removing the two least-correlated items did not improve IRT reliability.

Fifty-four items were thus retained for the final version of the IMHA. The three-level hierarchical model is displayed graphically in Figure 2. Reliability, IRT correlations and mean thetas (PCM scaled scores) for men and women for the nine-subscale and three-spectra models in the full dataset are reported in Table 4. These indicate that women had higher scores on every domain except Substance Abuse. Sex differences are displayed graphically in Figure 3. Model fit in the full dataset for the three-spectra, and nine-subscale models, also compared to a unidimensional total score including all 54-items, are reported in Table 5. According to χ^2, AIC and BIC, the more elaborated models fit significantly better than less elaborated models, with best fit for nine-subscales. CFA fit for the formal hierarchical model, excluding Life Difficulties, was strong, $\chi^2(3) = 41.01$, CFI = .996, TLI = .987, RMSEA = .056, SRMR = .011; standardized loadings are shown in Figure 4.

Outcome variables. Higher self-rated work performance over the last month (range = 1 - 10; M = 8.1; SD = 1.3), was significantly predicted (p < .001) by being older (β = .08), by lower scores on Depression (β = -.32), Sleep Problems (β

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 3. Raw Subscale Mean Scores, Percent who Scored Zero, and Score at Key Percentiles, by Gender for 54-Item Final Version of IMHA

Domain (number of items)	Max. Score	*M*	*SD*	% 0	50%	75%	90%	95%	*M*	*SD*	% 0	50%	75%	90%	95%
				Women (*N* = 2,786)							Men (*N* = 1,148)				
Depression (8)	48	9.89	10.29	18.4%	6	15	25	31	7.73	9.71	26.7%	4	11	22	29
Anxiety (8)	48	8.12	8.57	17%	5	12	20	25	5.58	7.42	31.5%	3	8	15	20
Post-Traumatic Stress (6)	36	4.47	5.61	27.7%	2	6	12	16	3.08	4.98	39.8%	1	4	9	13
Substance Abuse (7)	42	1.23	3.17	72.4%	0	1	4	7	1.80	3.83	61.7%	0	1	12	16
Anger (6)	36	1.58	2.42	50.7%	0	2	5	6	1.82	3.16	51.5%	0	3	5	9
Sleep Problems (4)	24	8.80	5.87	8%	8	13	17	19	7.15	5.50	11.8%	6	10	5	8
Life Stress (5)	30	9.20	6.72	9%	8	13	19	21	7.60	6.19	12.1%	6	11	15	18
Interpersonal Conflict (3)	18	2.74	3.39	34.9%	1	4	7	9	2.36	3.05	36.8%	1	3	16	19
Partner Conflict[1] (5)	30	1.24	2.38	48.3%	0	1	4	6	1.14	2.43	53.7%	0	1	6	8
Work Disengagement (2)	12	0.89	1.81	70%	0	1	3	5	0.65	1.59	77.2%	0	0	4	5
Internalizing Spectrum (22)	132	22.48	22.52	8.7%	15	33	55	69	16.39	20.48	15.8%	8	23	2	4
Externalizing Spectrum (18)	108	4.56	6.40	30.8%	2	6	13	17	5.92	8.39	26.1%	3	8	43	60
Life Difficulties[2] (14)	84[2]	21.63	14.29	3.5%	19	30	41	48	17.76	12.98	5.6%	15	25	15	19
P-factor (54)	324	24.53	23.93	7.5%	17	36	58	72	19.57	23.59	11.4%	10	28	34	41

Note. This table excludes 94 participants who did not report their gender and 20 who described themselves as "gender variant". Transgender women (*n* = 3) are included with women, and transgender men (*n* = 6) are included with men. Max. score indicates the highest score possible, if all items on scale were responded to with 6 (there were seven response options, from 0 to 6). '%0' = the percent that had a score of 0 for the scale. Correct reading starting from the fourth column, first row: "on the Depression scale, 50% of women had a score of 6 or lower."

[1] Only answered by those in a relationship: women *n* = 2,188; men *n* = 946.

[2] Excludes five items answered only by those in a relationship (which are included in IRT PCM analyses for these scales).

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

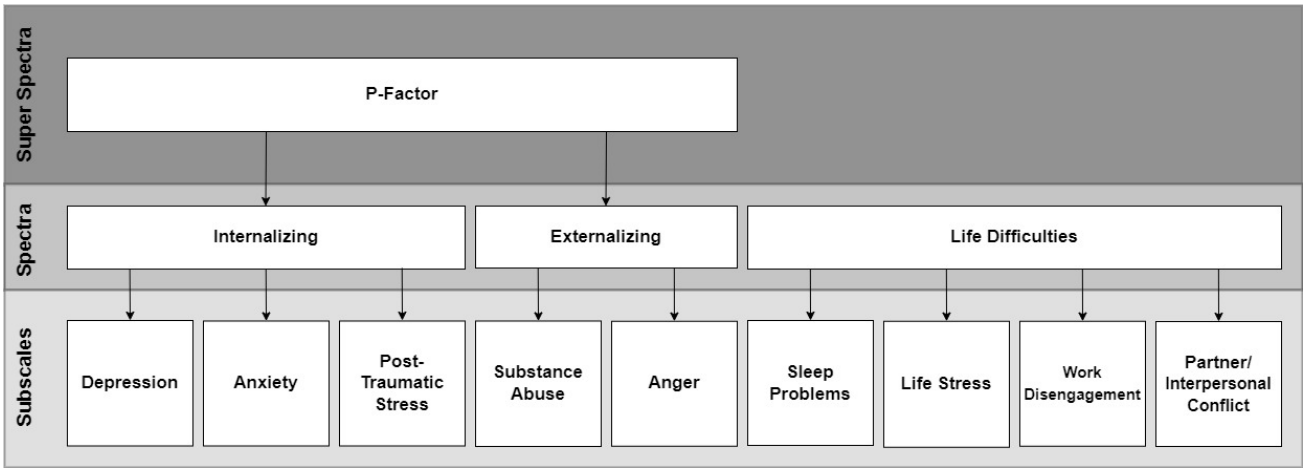

Figure 2. Graphical representation of the Final Three-Level International Mental Health Assessment

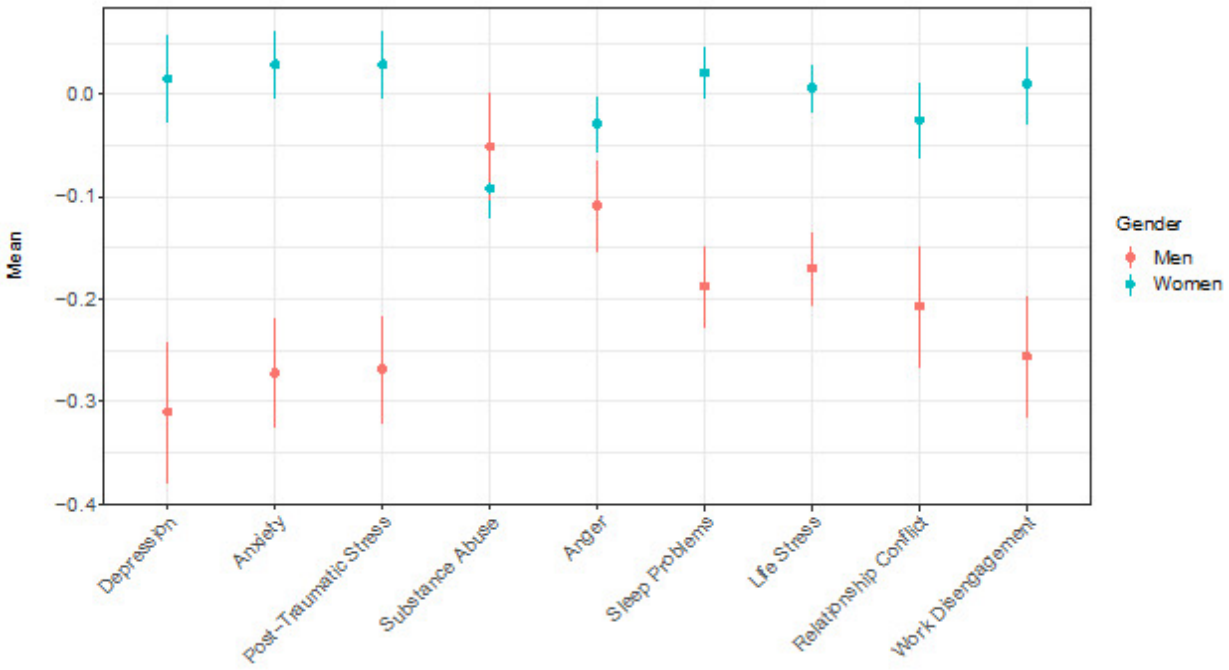

Figure 3. Mean Subscale Thetas (IRT Scaled Scores) by Gender, Final 54-Item IMHA

Note. 95% confidence intervals are shown. Women had significantly higher scores than men on all subscales except Substance Abuse at $p < .01$.

= -.12), and Work Disengagement (β = -.23), and, less intuitively, by higher scores on Post-Traumatic-Stress (β = .09). Together age, gender, and IMHA subscales accounted for 29% of variance in this rating. Hours of work missed (not a required response) was estimated by 57% of the sample (n = 2,322), with answers ranging from 0 (35% of respondents) to the maximum possible of 180 (M = 4; SD = 10.4). It was significantly predicted only by Life Stress and Work Disengagement ($p < .001$). Seventeen percent of the sample (n = 695) responded that they were in treatment with a mental health professional; this was significantly predicted by higher scores on Depression, Post-Traumatic Stress, and Work Disengagement ($p < .001$). The full model accounted for 15% of variance in treatment seeking.

Discussion

In this study, a 59-item version of the IMHA, refined based on the results of Study 1, was administered to a large employee sample. Good fit and reliability were found for a 54-item version, including a 35-item hierarchical model of P-factor, Internalizing and Externalizing spectra, and their associated five subscales: Anxiety, Depression, and Post-Traumatic Stress; and Substance Abuse and Anger. The full IMHA also includes a 19-item spectra of Life Difficulties, with subscales of Life Stress, Sleep Problems, Workplace Disengagement, and Interpersonal Conflict. These items and subscales are not conceptually solely related to mental health and thus are not included on p, although the scales and the spectrum are strongly associated with the other

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 4. Item Response Theory Scale Correlations and Reliabilities, Classical Test Theory Reliabilities, and Mean Thetas by Gender for the Final 54-item, Three-Spectra and Nine-Subscale versions of the IMHA

Scale (number of items)		1	2	3	4	5	6	7	8	9	10	11	12
2	Anxiety (8)	**.95**											
3	Post-Traumatic Stress (6)	**.92**	**.93**										
4	Substance Abuse (7)	.54	.50	.49									
5	Anger (6)	.73	.69	.66	.51								
6	Sleep Problems (4)	.81	.81	.78	.44	.58							
7	Life Stress (5)	.83	.83	.80	.38	.71	.80						
8	Interpersonal Conflict (3-8)	.79	.74	.74	.45	.66	.59	.77					
9	Work Disengagement (2)	.88	.89	.86	.54	.74	.75	.80	.71				
11	Externalizing Spectrum (18)	-	-	-	-	-	-	-	-	-	.71		
12	Life Difficulties (14)	-	-	-	-	-	-	-	-	-	**.90**	.66	-
	EAP reliability	.92	.90	.88	.60	.71	.85	.86	.81	.81	.92	.72	.91
	Cronbach α	.90	.86	.82	.79	.61	.79	.77	.79	.68	.94	.83	.89
Theta Women (*SD*)		.02 (1.16)	.03 (.91)	.03 (.90)	-.09 (.78)	-.03 (.72)	.02 (.68)	.01 (.64)	-.03 (1.01)	.01 (1.01)	.22 (.98)	.09 (.69)	.13 (.58)
Theta Men (*SD*)		-.31 (1.20)	-.27 (.93)	-.27 (.91)	-.05 (.91)	-.10 (.78)	-.19 (.68)	-.17 (.63)	-.21 (.03)	.26 (1.04)	-.08 (1.01)	.06 (.78)	-.04 (.58)

Note. N = 4,048. 1 = Depression, which has 8 items. 10 = Internalizing Spectrum with 22 items. Correlations over .90 are bolded for emphasis. Cronbach Alpha is on standardized items. Theta = mean scaled score. Women n = 2,783; men n = 1,142. In the 9-suscale model, t-tests indicate that women's scores are significantly higher than those of men for all subscales except Substance Abuse. For P-factor model, EAP reliability = .94, α = .95.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 5. Model fit indices for Unidimensional, Spectra, and Subscale Rasch Models

	Uni-dimensional	3 Spectra	9 Subscales
Estimated parameters	305	310	349
Item thresholds	304	304	304
Regression parameters	0	0	0
Covariance parameters	1	6	45
Deviance	383256.5	375044.6	369234.1
AIC	383867	375665	369932
BIC	384172	377619	372133
Compared to uni-dimensional			
Difference in deviance (χ^2)		8211.9	14022.4
Difference in parameters		5	44
Compared to 3-spectra model			
Difference in deviance (χ^2)			5810.5
Difference in parameters			39

Note: N = 4,048. For all models, constraint is on persons. Critical value for distribution with df 44 and p > .01 is 68.7.

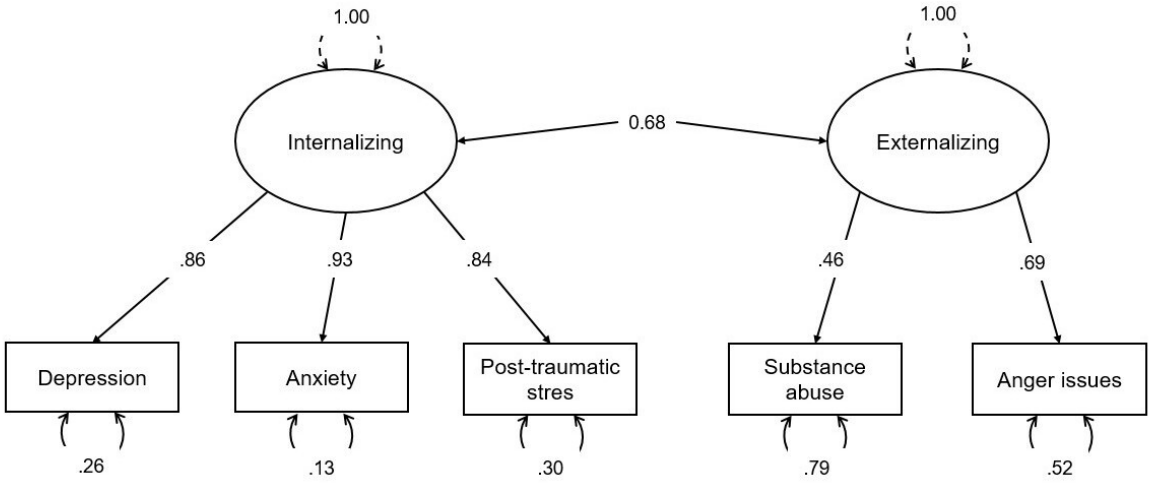

Figure 4. CFA Standardized Loadings for 35-items of Final IMHA forming Hierarchical Components: P, Spectra, and Five Subscales

spectra, especially Internalizing. These complaints about sleep length and quality, recent bad experiences and/or time deficits, conflict with close others (with most items focused on being a victim in the conflict), and missing work due to personal problems, could be consistent with internalizing or externalizing tendencies, but also with specific or temporary life circumstances such as the birth of a child, the death of a loved one, or social or economic hardship. Those with high scores may benefit from counseling or wellness interventions, without their scores indicating a tendency toward a disorder. We also find it advantageous to retain these subscales and spectra to help put mental health and functioning into a broader context of a person's life. It will also be useful to determine if the same strong associations are seen in future cross-cultural studies, for example in poorer countries where the incidence of Life Difficulties is higher.

The best fit was seen for the nine-subscale model. The three internalizing scales, Depression, Anxiety and Post-Traumatic Stress, remain highly intercorrelated (.90 to .94), which is a limitation of that model and indicates the advantage of the spectra model. However, intercorrelations are lower for the Externalizing scales (.51), and separate scores may be useful in some contexts, as long as the high intercorrelations among the internalizing scales are acknowledged. The three nested models otherwise have strong measurement properties, and each may be the best level of analysis depending on the intervention, tracking, or research goal.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 6. Inventories used for Assessment of Convergent and Divergent Validity of IMHA Subscales

IMHA Subscale	Measure for validation	Items
Depression	Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001); 4-point response scale, regarding frequency over the last two weeks	9
Anxiety	Generalized Anxiety Disorder (GAD-7; Spitzer et al., 2006); 4-point response scale, regarding frequency over the last two weeks	7
PTS	PTSD Checklist - Civilian Version (PCL-C; Lang et al., 2012; Weathers et al., 1993); 5-point scale from "not at all" to "extremely" regarding last month	17
Substance Use/Abuse	Alcohol Use Disorders Identification Test (AUDIT; Babor et al., 2001); three different 5-point ratings of frequency reporting on past year	10
	The Drug Abuse Screening Test (DAST-10; Skinner, 1982; Yudko et al., 2007); yes/no questions reporting on last year	10
	Negative urgency scale of UPPS (urgency-premeditation-perseverance-sensation seeking; Berg et al., 2015); 4 options, strongly agree to disagree [1]	4
Anger	Buss Perry Aggression Questionnaire (BPAQ; Buss & Perry, 1992); four subscales: physical aggression, verbal aggression, anger, hostility; 5 options, from extremely characteristic to uncharacteristic, two items reverse-scored	29
Sleep Problems	The Athens Insomnia Scale (AIS; Soldatos et al., 2000); three different 4-point response options for the various items	8
Life Stress	Stress Overload Scale (SOS; Amirkhan, 2012, 2018); subscales of personal vulnerability and event load; 5-point scale from "not at all" to "a lot"	10
Partner Conflict	Tool for Intimate Partner Violence Screening (HITS; Sherin et al., 1998); 5-point scale from "never" to "frequently"	4
	The Couples Satisfaction Index (CSI) Funk & Rogge, 2007 short version; 5- and 6-point scales combined in total	4
Work Disengagement	Workplace Outcome Suite (WOS; Lennox et al., 2010), 4 subscales: Absenteeism (write in), Presenteeism, Work Engagement, Work Distress; latter three answered on 5-point scale, "strongly disagree" to "strongly agree"	20
Other	General Self-Reported Health (GSRH; DeSalvo et al., 2006); Physical health rated on a 5-point scale from poor to excellent	2

[1] Results not reported. It was the only scale for which response level of agreement decreased from left to right, and this may have led to confusion. Correlations were low with all substance-related (.22-.26) and BPAQ scales (.30-.49), with which content overlapped. It is also less directly related to the construct.

Study 3: Assess Convergent and Divergent Validity of the IMHA Subscales

Methods

Materials. The 54-item version of the IMHA used in Study 2, as described above, was administered with 12 other inventories, detailed in Table 6.

Participants. Five hundred eight community adults recruited through Prolific completed the 13 surveys through a Lime Survey interface. Eight cases were rejected due to failure of more than half of eight attention checks, which was sometimes combined with very short completion times (under 5 minutes for 170 items), leading to a final sample of 500. This was estimated to be a larger sample than needed for the planned correlational analyses, with moderate to high anticipated correlations (Cohen, 1992). Participation was restricted to individuals currently residing in the United States who speak English fluently to maximize comparability with Study 1 and 2 samples. Age, gender, and ethnicity are reported in the far right-hand side of Table 1.

Procedure. Institutional ethical review of this study was not available at the lead authors' European institution for a survey study. However, survey responses were entirely anonymous and data collection complied with the ethical standards of the American Psychological Association. A debriefing page provided contact information for free phone and sms crises lines. A payment was made based on an hourly rate of US$10 per hour. The surveys were administered in five possible orders, and item presentation within most surveys was randomized. The format required answers to move through the survey, thus, there was virtually no missing data, with the following exceptions: Errors at the start of collection meant that some items were not presented to the first eight participants. The first WOS scale used a 'write in' format for number of hours, and responses for seven participants who wrote "not applicable" or described an issue that made them miss work were removed. Participants who reported that they were not in a relationship responded, in a few cases, to partner conflict questions, which were excluded from analysis.

Analyses. Pearson correlations were calculated for all subscales and inventories. A correlation of .70 or higher was seen as indicating that the two inventories measure highly overlapping content (convergent validity), as it indicates that half their variance is shared. It was hypothesized that the established inventories would have high correlations, and their highest correlations, with their related IMHA subscale. Thus, a correlation of this value with

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

an intended scale was interpreted as providing evidence of convergent validity. Otherwise, correlations are interpreted following Cohen's (1988) conventions: .10 or lower as weak; .30 as moderate; .50 or larger as strong.

Results

Scale scores and psychometric properties are reported in Table 7, along with correlations between the IMHA subscales with each other and with the 21 external scales, belonging to 11 inventories. Four IMHA subscales (Depression, Anxiety, Post-Traumatic Stress, Sleep) were each compared to a single other inventory. In these cases, all anticipated correlations were .70 or higher and the intended inventory had its highest correlation with the intended IMHA subscale.

The other five IMHA scales were compared to inventories with more than one subscale or to more than one inventory, which allows for further distinctions to be explored. IMHA Life Stress was compared to the 10-item SOS and its two subscales: Personal Vulnerability and Event Overload. While both the SOS Total and Event Overload had correlations over .70 with IMHA Stress, only Event Overload had its highest correlation with IMHA Life Stress. The 2-item IMHA Work Disengagement subscale was not associated above .62 with any WOS subscale; these were also not strongly associated with any other IMHA subscales. IMHA Partner Conflict (separated here from Interpersonal Confict which also includes three non-partner items) associated .70 with the Tool for Intimate Partner Violence Screening (HITS), but not above threshold with the Couple Satisfaction Index, which was not strongly associated with any IMHA subscale. IMHA Substance Abuse was more associated with AUDIT (alcohol abuse) than DAST (drug abuse). IMHA Anger was not strongly associated with any BPAQ subscales, which were also not strongly associated with any other IMHA subscales.

Discussion

In Study 3 the subscales of the IMHA were compared to 11 other inventories with 21 unique subscales to assess their convergent and divergent validity in an online sample of adults residing in the United States. Observed associations supported the convergent validity of Depression, Anxiety, Post-Traumatic Stress, and Sleep subscales, based on strong associations with a key inventory. The other five IMHA scales were compared to inventories with more than one subscale, or to more than one inventory, allowing for examination of specific distinctions.

In three cases, logical results confirm the focus and/or format of IMHA subscales. IMHA Life Stress primarily overlapped with SOS Event Overload, which was designed to assess 'impinging demands' as opposed to 'depleted resources' with the Personal Vulnerability scale. The latter includes internalizing content (underlined by associations with IMHA Depression and Anxiety) that was intentionally excluded in the IMHA subscale to maximize scale separation. IMHA Partner Conflict associated strongly with HITS (partner violence), but not with the Couple Satisfaction Index (CSI). This result was not surprising given the positive framing of CSI items, and the observation in Study 1 of low associations between IMHA symptoms and reverse-scored Protective Factors items. These do not appear to form the ends of a single pole, but to assess different aspects of the relationship experience.

IMHA Substance Abuse was more associated with AUDIT alcohol abuse (r = .78), than DAST drug abuse (.49). One explanation for this difference may be methodological - the AUDIT and IMHA use very similar response options, whereas DAST has a yes/no format. This supposition is supported by correlations below .50 for nearly synonymous items: e.g., IMHA "I felt guilt or remorse after drinking or using drugs" and DAST "Do you ever feel bad or guilty about your drug use?". Different dynamics between alcohol and drug use may also be at play. For efficiency, the IMHA scale integrates alcohol and drug problems (of seven items, two are specific to alcohol, and five refer to either/both). However, internal consistency is high (α = .92), and we find integration suitable for screening purposes.

Two IMHA scales did not have correlations .70 or higher with the intended inventories. The strongest association for IMHA Work Disengagement was with WOS Presenteeism (r = .58). This is logical given the two IMHA items are about distraction and missing small amounts of work. While not meeting our cut-off, we found this acceptable for this two-item scale, intended to provide a minimal overview assessment of workplace impact.

IMHA Anger, focused on feeling and/or expressing anger and/or arguing in a variety of contexts, had the expected convergent patterns with the associated BPAQ subscales, but the magnitudes were only moderate. There may be a methodological explanation: BPAQ items are answered on a scale of "extremely characteristic (versus uncharacteristic) of me" drawing on schemas about the self rather than on the concrete occurrence of a behavior. This hypothesis is supported by a correlation between nearly identical items: BPAQ "I have become so mad that I have broken things" and IMHA "I got so angry I broke something" of only .32. Future studies could further explore this subscale's associations. For the time being, given the lack of anger inventories in the public domain (we found no others) we stand by the value of the IMHA, as the length of the BPAQ and its wordy, complex items and response options make it less suitable for workplace screening or cross-cultural research.

A limitation here is the lack of divergent validity between the internalizing scales. PHQ Depression also correlated .70 or higher with IMHA Anxiety and Sleep Problems; GAD Anxiety with IMHA Depression and Post-Traumatic Stress; and PCLC Trauma, SOS Total and SOS Personal Vulnerability with IMHA Depression. To assess for which items impacted divergent validity, a post hoc exploration of correlations between IMHA Internalizing Spectra items plus Sleep Problems and the four external scales identified two IMHA items: "I felt hopeless about the future" (Depression) and "It was hard to control my worrying" (Anxiety), both of which had highest correlations as intended, but additional correlations > .70 with other external scales. These two items are valid for assessing the internalizing spec-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Table 7. Study 3: Scale Scores, Psychometric Properties and Correlations between IMHA and External Scales, and among Other Scales for Internalizing Conditions

				Correlations with IMHA Scales										Other Internalizing Scales				
Scale (items)	*M*	*SD*	α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Depression (8)	9.33	9.87	.91	-														
2. Anxiety (8)	6.17	7.81	.90	**.78**														
3. Post-Traumatic Stress (6)	3.99	5.60	.86	**.71**	**.83**													
4. Sleep Problems (4)	6.79	5.76	.87	.68	.64	.59												
5. Life Stress (5)	5.38	5.37	.77	.64	.69	.64	.66											
6. Work Disengagement (2)	1.15	2.08	.77	.61	.57	.58	.49	.62										
7. Partner Conflict (5)	1.14	3.49	.92	.37	.41	.46	.27	.39	.44									
8. Interpersonal Conflict (8)	3.06	5.72	.91	.54	.54	.60	.41	.53	.52	.91								
9. Substance Abuse (12)	1.50	4.05	.90	.31	.39	.42	.24	.29	.37	.50	.48							
10. Anger (6)	1.23	3.19	.89	.27	.40	.46	.19	.35	.42	**.73**	**.70**	.57						
11. PHQ Depression (9)	6.99	6.37	.90	**.85**	**.77**	.69	**.72**	.63	.58	.36	.50	.35	.33					
12. GAD Anxiety (7)	5.64	5.35	.91	**.75**	**.81**	**.72**	.68	.65	.49	.30	.45	.31	.29	**.84**				
13. PCLC Trauma (17)	14.79	12.88	.94	**.79**	**.80**	**.84**	.66	.65	.60	.47	.61	.42	.45	**.81**	**.78**			
14. Athens Insomnia Scale (7)	5.96	4.49	.87	.60	.58	.54	**.79**	.56	.44	.19	.33	.22	.18	**.72**	.67	.63		
15. SOS Total (10)	22.75	11.06	.95	**.75**	.68	.62	.65	**.72**	.60	.33	.48	.28	.30	**.78**	**.75**	**.74**	.65	
SOS Vulnerability (5)	10.94	6.00	.93	**.81**	.69	.63	.63	.65	.59	.35	.50	.30	.30	**.81**	**.76**	**.76**	.63	.94
SOS Event Load (5)	11.81	5.84	.92	.60	.59	.53	.60	**.70**	.52	.27	.41	.22	.27	.65	.65	.62	.58	.94
WOS Absenteeism (5)	6.68	15.74	-	.10	.22	.21	.14	.20	.25	-.01	.02	.14	.12	.20	.21	.22	.13	.18
WOS Presenteeism (5)	10.52	5.61	.95	.61	.56	.52	.45	.49	.58	.28	.39	.33	.30	.61	.55	.61	.49	.63
WOS Engagement (5)	15.23	4.94	.73	-.18	-.05	-.04	-.16	-.04	-.06	.03	.00	.04	.08	-.17	-.12	-.06	-.17	-.08
WOS Work Distress (5)	11.89	5.55	.91	.56	.48	.41	.50	.43	.40	.21	.30	.27	.25	.60	.56	.53	.51	.60
HITS Partner (4)	5.06	2.06	.83	.32	.36	.46	.31	.31	.33	**.70**	.68	.40	.55	.41	.35	.43	.32	.33
Couple Satisfaction Ind (4)	15.62	4.62	.96	-.41	-.32	-.32	-.45	-.28	-.18	-.24	-.34	-.13	-.03	-.42	-.41	-.37	-.42	-.38
AUDIT Alcohol (10)	4.34	5.57	.90	.24	.37	.39	.15	.25	.32	.55	.52	**.78**	.62	.29	.24	.36	.19	.23
DAST Drug Abuse (10)	0.89	1.67	.83	.23	.33	.37	.18	.26	.31	.25	.28	.49	.37	.31	.26	.34	.21	.29
BPAQ Total (29)	64.00	19.43	.93	.39	.39	.39	.28	.17	.32	.34	.36	.35	.40	.42	.42	.51	.07	.22
BPAQ Physical (9)	17.68	6.78	.85	.13	.17	.22	.06	.21	.23	.29	.35	.23	.27	.19	.17	.28	.12	.27
BPAQ Verbal (5)	12.30	4.35	.77	.20	.18	.19	.14	.29	.18	.37	.46	.34	.40	.23	.22	.26	.23	.40
BPAQ Anger (7)	14.40	5.52	.82	.35	.35	.35	.22	.40	.26	.29	.39	.26	.28	.36	.36	.46	.40	.58
BPAQ Hostility (8)	19.62	7.37	.86	.54	.49	.46	.43	.34	.32	.39	.48	.37	.42	.53	.56	.58	.27	.47
GSRH Health (2)	6.87	2.00	.93	-.40	-.31	-.27	-.41	-.31	-.21	.06	-.04	.04	.14	-.42	-.37	-.31	-.42	-.34

Note. N = 500 except: IMHA Partner Conflict and Interpersonal Conflict *n* = 389; PHQ, GAD, SOS, AUDIT, and DAST *n* = 492. Alpha based on standardized items. Correlations specific to convergent validity are underlined; correlations .70 and higher bolded for emphasis; correlations between scales using the same items (IMHA Conflict; SOS Total) are italicized and not bolded.

[1]Excludes 5 partner conflict items in order to use full sample, i.e. those in a relationship and not.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

trum, but appear to provide less subscale discrimination. Note that this problem is not unique to the IMHA. Intercorrelations among the external scales in Table 7 are generally even higher, for example .84 between GAD and PHQ. These results underline the relevance of spectra assessment.

Overall Discussion

This report describes the development of an efficient inventory for screening of psychological disorder symptoms in a normal adult population at three levels of analysis. It was intended to be as brief as possible to encourage completion and minimize distress, with subscales for multiple common problems to allow for feedback and communication with counselors, and a hierarchical structure including spectra and p-factor to account for the consistent overlap between traditional domains. Both specificity and the potential for intra- and inter-individual, as well as cross-group and cross-cultural comparisons were maximized by reference to specific behaviors and a response scale based on absolute frequencies of days in the last month. Item development drew on a large base of existing literature, using systematic comparison of the core content of common categories of disorders to create an integrated inventory. The intended uses of the IMHA are screening for prevention and early intervention, tracking individual change, and cross-group and cross-cultural research.

An initial 69-item version with 10 subscales was tested in a large employee sample and a sample of counseling clients using IRT (PCM). Refinements to increase subscale reliabilities and decrease item redundancy and subscale intercorrelations led to a 59-item version, which was administered to a second large employee sample. Tests in the first split half of the data allowed for testing refinements in the second half. At this stage, five substance use items were dropped, as they contributed neither conceptually nor empirically to P or to the Externalizing spectrum. Final calibration and calculation of percentiles for norms is reported using the full sample for three telescoping models: a unidimensional P-factor; three-spectra including Internalizing, Externalizing, and Life Difficulties; and nine-subscales. This dimensional approach situates the inventory in current empirical knowledge (e.g. Conway et al., 2019; Kotov et al., 2017) and increases the utility of the measure. For example, subscales provide familiar information for targeted interventions and tracking, to facilitate communication and access to psycho-education resources. P-factor and spectra can provide a useful overview for an individual or a counselor to track improvement or decline across a range of factors, including for those who do not meet criteria for a disorder but with personality traits that predilect them to internalizing or externalizing tendencies. Alongside this, high scores on Life Difficulties may put other elevated scores into context, or even in the absence of other high scores, indicate the need for support during a challenge or transition such as a new child, grief, unemployment, divorce, or conflict.

The brevity of the IMHA compared to other comprehensive inventories is central to its intended use for screening for early intervention among community adults, not for fine-grained diagnoses (e.g. Kemper et al., 2019). At 54 items, it requires half the time or less than the SCL, ABCL, EAPI, or Spectra. A period in an individual's life when psychological problems and/or life difficulties are increasing, may also be a time when it is difficult to proactively seek assistance. Drinking, anxious feelings, low moods, poor sleep and concentration, or partner conflict may have worsened over time, leading to habituation. Offering a screener to employees or community groups can help catch problems before a crisis, ideally, increasing awareness and ability to articulate current needs, and thus to access and benefit from resources.

Importantly, in contrast to many popular clinical measures, the IMHA is freely available for scientific and educational uses. The inventory is optimized for research given its brevity, making it practical to include with other variables. The specificity of items (e.g. as recommend by Hopwood et al., 2022) and a concrete rating scale in terms of how many times the symptom occurred, adds to its value as a research tool, serving to diminish reference group effects. These elements also allow for meaningful cross-group comparison of items: the average frequency of a symptom or experience can be compared by treating each item as its own variable (as in Saucier et al., 2015; Westen & Shedler, 2007), to assess for differences even without scale invariance.

Limitations and Future Directions

This report on the development of the IMHA was limited by not having concurrent data on the presence or absence of diagnoses of disorders. The percentiles from a large sample of indicate the prevalence of such symptoms in this population. Future work could relate scores on the IMHA to diagnoses by structured interview, or those estimated by longer surveys. The studies were also limited by gender imbalance: in Studies 1 and 2 the samples were around 70% women, potentially skewing results. However, this is mitigated by large samples that included over a thousand men, and reporting observed scores and norms separately.

A possible criticism of the IMHA is that it aims to 'have it both ways', using traditional disorder categories to suit clinicians, while also integrating evidence of higher-order spectra that are used to avoid reifying arbitrary distinctions. We believe, however, that the presence of traditional categories alongside the spectra serves to raise consciousness, allowing clinicians to integrate what is familiar into this broader, empirically-based context. The continuous measurement approach of the IMHA accurately reflects the continuous nature of psychopathology and its lack of discrete categorical entities (Kotov et al., 2017).

While the IMHA was designed with cross-cultural comparisons in mind, the current study relied solely on a North American sample and this potential has yet to be tested. Future studies should test the IMHA's utility and norms in other societies, and assess its measurement invariance across languages and groups (e.g. Fischer & Karl, 2019).

Future work should assess the effectiveness and the impact of the response option design The magnitude of difference between the response options, in terms of measurement and clinical practice, should be quantified. Qualitative

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

methods could be used to explore the experiences of participants with regard to the response scale. Estimating the specific frequency of a symptom over the last month is likely more cognitively demanding than making a less precise estimate of "sometimes" versus "often". However, it also provides for more comparable data, both within individuals over time and between individuals and groups. It would be useful to compare the time spent to complete an inventory in each of the two modes, and to elicit participant reactions to determine how these variations are perceived. It would also be useful to compare responses between the two types of inventories, to help elucidate a 'folk' understanding of how many days in the last month constitutes "a lot" or "often" and how this varies across societies, subcultural groups, or age cohorts. Comparisons to other-reports, to clinician assessment, and to diagnoses could be used to compare the validity of the two approaches.

Conclusions

The International Mental Health Assessment (IMHA), an efficient 54-item measure of psychological disorder symptoms is presented. It is made available freely for educational, non-profit or research purposes. This inventory was designed with two goals in mind: to effectively screen and assess degree of psychological problems among adults in the community, and to facilitate research with cross-cultural and cross-group comparisons. It uses concrete items and a response scale based on specific frequencies of symptoms in the last month, designed to minimize reference group effects by avoiding vague comparisons with others. Good reliability for the inventory was established at three levels: P-factor global assessment; three broad spectra of Internalizing, Externalizing, and Life Difficulties; and nine-subscales: Substance Abuse, Anxiety, Depression, Post-Traumatic Stress, Life Stress, Sleep Problems, Anger, Workplace Disengagement, and Interpersonal Conflict.

Author Contributions

- Contributed to conception and design: AGT, JM
- Contributed to acquisition of data: AGT, JM
- Contributed to analysis and interpretation of data: AGT, KS, JM
- Drafted and/or revised the article: AGT, KS, JM
- Approved the submitted version for publication: AGT, JM, KS

Competing Interests

Julie Marshall is the Chief Operating Officer at Canopy Wellbeing, which sells a product that incorporates the inventory presented, and as such she receives a salary and is part owner of the company. Amber Gayle Thalmayer received compensation from Canopy Wellbeing as an independent consultant while developing the inventory. She completed the manuscript while being supported by Swiss National Science Foundation fellowship PCEFP1_194552.

Author Note

Requests to use the inventory commercially can be directed to Julie Marshall jmarshall@cascadecenters.com or Anthony Brown abrown@cascadecenters.com. Requests to use the inventory free of charge for educational and research purposes can be made at: https://www.psychology.uzh.ch/en/areas/sob/psyges/research/imha.html.

Data Accessibility Statement

The raw data and scripts for running the analyses are available on the Open Science Framework (https://osf.io/vysj4/).

Submitted: July 20, 2022 PDT, Accepted: April 07, 2023 PDT

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

References

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for the ASEBA adult forms & profiles: an integrated system of multi-informant Assessment*. University of Vermont.

Adams, R. J., & Khoo, S. T. (1996). *Quest: The Interactive Test Analysis System*. http://eric.ed.gov/?id=ED362553

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics*, *22*(1), 47–76. https://doi.org/10.3102/10769986022001047

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Amirkhan, J. H. (2012). Stress Overload: A New Approach to the Assessment of Stress. *American Journal of Community Psychology*, *49*(1–2), 55–71. https://doi.org/10.1007/s10464-011-9438-x

Amirkhan, J. H. (2018). A brief stress diagnostic tool: The short stress overload scale. *Assessment*, *25*(8), 1001–1013. https://doi.org/10.1177/1073191116673173

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/bf02293814

Anton, W. D., & Reed, J. H. (1994). *Employee Assistance Program Inventory–Professional Manual*. PAR.

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *AUDIT the alcohol use disorders identification test: Guidelines for use in primary care* (WHO/MSD/MSB 01.6a). World Health Organization.

Berg, J. M., Latzman, R. D., Bliwise, N. G., & Lilienfeld, S. O. (2015). Parsing the heterogeneity of impulsivity: A meta-analytic review of the behavioral implications of the UPPS for psychopathology. *Psychological Assessment*, *27*(4), 1129–1146. https://doi.org/10.1037/pas0000111

Blais, M. A., & Sinclair, S. J. (2019). *Introduction to the SPECTRA: Indices of Psychopathology: An assessment inventory aligned with the hierarchical-dimensional model of psychopathology* [White paper]. PAR.

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly*, *43*(1), 92–101. https://doi.org/10.1086/268494

Brown, J. D. (2011). Likert items and scales of measurement. *Statistics*, *15*(1), 10–14.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*(3), 214–227. https://doi.org/10.1037/0003-066x.39.3.214

Buss, A. H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology*, *63*(3), 452–459. https://doi.org/10.1037/0022-3514.63.3.452

Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry*, *175*(9), 831–844. https://doi.org/10.1176/appi.ajp.2018.17121383

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Conway, C. C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., Mullins-Sweatt, S. N., Shackman, A. J., Skodol, A. E., South, S. C., Sunderland, M., Waszczuk, M. A., Zald, D. H., Afzali, M. H., Bornovalova, M. A., Carragher, N., Docherty, A. R., Jonas, K. G., Krueger, R. F., … Eaton, N. R. (2019). A Hierarchical taxonomy of psychopathology can transform mental health research. *Perspectives on Psychological Science*, *14*(3), 419–436. https://doi.org/10.1177/1745691618810696

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Derogatis, L., & Derogatis, L. (2001). *Brief Symptom Inventory (BSI). Administration, scoring, and procedures manual*. https://www.scienceopen.com/document?vid=ca91e009-0937-4008-b194-df872b5022cb

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question: a meta-analysis. *Journal of General Internal Medicine*, *21*(3), 267–275. https://doi.org/10.1111/j.1525-1497.2005.00291.x

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: an experimental study. *Marketing Letters*, *15*(1), 21–36. https://doi.org/10.1023/b:mark.0000021968.86465.00

Eaton, N. R., South, S. C., & Krueger, R. F. (2010). The meaning of comorbidity among common mental disorders. In *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (Vol. 2, pp. 223–241).

Edwards, P., Roberts, I., Sandercock, P., & Frost, C. (2004). Follow-up by mail in clinical trials: Does questionnaire length matter? *Controlled Clinical Trials*, *25*(1), 31–52. https://doi.org/10.1016/j.cct.2003.08.013

Eisen, S. V., Wilcox, M., Leff, H. S., Schaefer, E., & Culhane, M. A. (1999). Assessing behavioral health outcomes in outpatient programs: Reliability and validity of the BASIS-32. *Journal of Behavioral Health Services & Research*, *26*(1), 5–17. https://doi.org/10.1007/bf02287790

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Fischer, R., & Karl, J. A. (2019). A Primer to (Cross-Cultural) Multi-Group Invariance Testing Possibilities in R. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01507

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, 114–123.

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, *21*(4), 572–583. https://doi.org/10.1037/0893-3200.21.4.572

Goring, H., Baldwin, R., Marriott, A., Pratt, H., & Roberts, C. (2004). Validation of short screening tests for depression and cognitive impairment in older medically ill inpatients. *International Journal of Geriatric Psychiatry*, *19*(5), 465–471. https://doi.org/10.1002/gps.1115

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science*, *19*(4), 309–313. https://doi.org/10.1111/j.1467-9280.2008.02085.x

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

Hopwood, C. J., Wright, A. G. C., & Bleidorn, W. (2022). Person–environment transactions differentiate personality and psychopathology. *Nature Reviews Psychology*, *1*(1), 55–63. https://doi.org/10.1038/s44159-021-00004-0

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277. https://doi.org/10.1177/0022022104272905

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2019). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive–compulsive symptoms. *Assessment*, *26*(5), 767–782. https://doi.org/10.1177/1073191918810057

Kiefer, T., Robitzsch, A., & Wu, M. (2017). Package "TAM" Type Package Title Test Analysis Modules. In *mran.microsoft.com*. https://mran.microsoft.com/snapshot/2017-02-20/web/packages/TAM/TAM.pdf

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Labott, S. M., Johnson, T. P., Fendrich, M., & Feeny, N. C. (2013). Emotional risks to respondents in survey research: Some empirical evidence. *Journal of Empirical Research on Human Research Ethics*, *8*(4), 53–66. https://doi.org/10.1525/jer.2013.8.4.53

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. D., Shimokawa, K., Christopherson, C., & Burlingame, G. M. (2004). *Administraton and scoring manual for the OQ-45.2*. American Professional Credentialing Services.

Lang, A. J., Wilkins, K., Roy-Byrne, P. P., Golinelli, D., Chavira, D., Sherbourne, C., Rose, R. D., Bystritsky, A., Sullivan, G., Craske, M. G., & Stein, M. B. (2012). Abbreviated PTSD Checklist (PCL) as a guide to clinical response. *General Hospital Psychiatry*, *34*(4), 332–338. https://doi.org/10.1016/j.genhosppsych.2012.02.003

Lennox, R. D., Sharar, D., Schmitz, E., & Goehner, D. B. (2010). Development and Validation of the Chestnut Global Partners Workplace Outcome Suite. *Journal of Workplace Behavioral Health*, *25*(2), 107–131. https://doi.org/10.1080/15555241003760995

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/bf02296272

Oishi, S., & Roth, D. P. (2009). The role of self-reports in culture and personality research: It is too early to give up on self-reports. *Journal of Research in Personality*, *43*(1), 107–109. https://doi.org/10.1016/j.jrp.2008.11.002

Pettersson, E., Larsson, H., & Lichtenstein, P. (2016). Common psychiatric disorders share the same genetic origin: a multivariate sibling study of the Swedish population. *Molecular Psychiatry*, *21*(5), 717–721. https://doi.org/10.1038/mp.2015.116

Robitzsch, A., & Robitzsch, M. A. (2020). *sirt: Supplementary item response theory models*.

Saucier, G., Kenner, J., Iurino, K., Bou Malham, P., Chen, Z., Thalmayer, A. G., Kemmelmeier, M., Tov, W., Boutti, R., Metaferia, H., Çankaya, B., Mastor, K. A., Hsu, K.-Y., Wu, R., Maniruzzaman, M., Rugira, J., Tsaousis, I., Sosnyuk, O., Regmi Adhikary, J., … Altschul, C. (2015). Cross-Cultural Differences in a Global "Survey of World Views." *Journal of Cross-Cultural Psychology*, *46*(1), 53–70. https://doi.org/10.1177/0022022114551791

Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly*, *55*(3), 395–423. https://doi.org/10.1086/269270

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Schriesheim, C. A., & Novelli, L., Jr. (1989). A comparative test of the interval-scale properties of magnitude estimation and case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement*, *49*(1), 59–74. https://doi.org/10.1177/0013164489491007

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Selzam, S., Coleman, J. R. I., Caspi, A., Moffitt, T. E., & Plomin, R. (2018). A polygenic p factor for major psychiatric disorders. *Translational Psychiatry*, *8*(1). https://doi.org/10.1038/s41398-018-0217-4

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation*, *14*(1), 87–100. https://doi.org/10.1080/13803610801896653

Sherin, K. M., Sinacore, J. M., Li, X., & Zitter, R. E. (1998). "HITS" A domestic violence screening tool for use in the community. *Family Medicine*, *30*, 508–512.

Skinner, H. A. (1982). The drug abuse screening test. *Addictive Behaviors*, *7*(4), 363–371. https://doi.org/10.1016/0306-4603(82)90005-3

Soldatos, C. R., Dikeos, D. G., & Paparrigopoulos, T. J. (2000). Athens Insomnia Scale: Validation of an instrument based on ICD-10 criteria. *Journal of Psychosomatic Research*, *48*(6), 555–560. https://doi.org/10.1016/s0022-3999(00)00095-7

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Thalmayer, A. G. (2015). Alternative Models of the Outcome Questionnaire-45. *European Journal of Psychological Assessment*, *31*(2), 120–130. https://doi.org/10.1027/1015-5759/a000216

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*, *23*(4), 995–1009. https://doi.org/10.1037/a0024165

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, *43*(8), 1205–1228. https://doi.org/10.1177/0022022111428083

Veit, C. T., & Ware, J. E. (1983). The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*, *51*(5), 730–742. https://doi.org/10.1037/0022-006x.51.5.730

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993). The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility. *Annual Convention of the International Society for Traumatic Stress Studies, San Antonio, TX*, *462*.

Westen, D., & Shedler, J. (2007). Personality diagnosis with the Shedler-Westen Assessment Procedure (SWAP): Integrating clinical and statistical measurement and prediction. *Journal of Abnormal Psychology*, *116*(4), 810–822. https://doi.org/10.1037/0021-843x.116.4.810

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum.

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*. https://apps.who.int/iris/bitstream/handle/10665/254610/WHOMSD?sequence=1

Yudko, E., Lozhkina, O., & Fouts, A. (2007). A comprehensive review of the psychometric properties of the Drug Abuse Screening Test. *Journal of Substance Abuse Treatment*, *32*(2), 189–198. https://doi.org/10.1016/j.jsat.2006.08.002

Zimmer, F., Draxler, C., & Debelak, R. (2022). Power analysis for the Wald, LR, score, and gradient tests in a Marginal Maximum Likelihood framework: Applications in IRT. *Psychometrika*. https://doi.org/10.1007/s11336-022-09883-5

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/74546-the-international-mental-health-assessment-validation-of-an-efficient-screening-inventory/attachment/156915.docx?auth_token=aXTVFPsyEFFCjdi5d6mN

Supplemental Material

Download: https://collabra.scholasticahq.com/article/74546-the-international-mental-health-assessment-validation-of-an-efficient-screening-inventory/attachment/156916.docx?auth_token=aXTVFPsyEFFCjdi5d6mN

Downloaded from http://online.ucpress.edu/collabra/article-pdf/9/1/74546/829048/collabra_2023_9_1_74546.pdf by guest on 10 November 2024