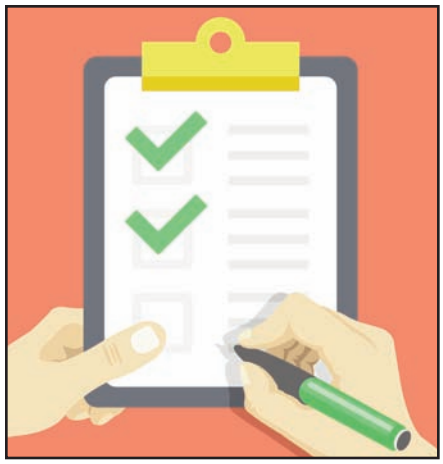


REVIEW

The Depression Inventory Development Workgroup: A Collaborative, Empirically Driven Initiative to Develop a New Assessment Tool for Major Depressive Disorder

by ANTHONY L. VACCARINO, PhD; KENNETH R. EVANS, PhD; AMIR H. KALALI, MD; SIDNEY H. KENNEDY, MD, FRCPC; NINA ENGELHARDT, PhD; BENICIO N. FREY, MD, MSc, PhD; JOHN H. GREIST, MD; KENNETH A. KOBAK, PhD; RAYMOND W. LAM, MD, FRCPC; GLENDA MACQUEEN, MD, PhD, FRCPC; ROUMEN MILEV, MD, PhD, FRCPC, FRCPsych; FRANCA M. PLACENZA, PhD; ARUN V. RAVINDRAN, MB, MSc, PhD, FRCPC, FRCPsych; DAVID V. SHEEHAN, MD, MBA; TERRENCE SILLS, PhD; and JANET B.W. WILLIAMS, PhD

Dr. Vaccarino is with Indoc Research, Toronto, Ontario, Canada; Dr. Evans is with Indoc Research, Toronto, Ontario, Canada, and Department of Pathology and Molecular Medicine, Queen's University, Kingston, Ontario, Canada; Dr. Kalali is with Quintiles Inc, San Diego, California, USA, and Department of Psychiatry, University of California, San Diego, California, USA; Drs. Kennedy and Placenza are with University Health Network, Toronto, Ontario, Canada; Dr. Engelhardt is with Cronos Clinical Consulting Services, Lambertville, New Jersey, USA; Dr. Frey is with Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada, and Mood Disorders Program and Women's Health Concerns Clinic, St. Joseph's Healthcare, Hamilton, Ontario, Canada; Dr. Greist is with Department of Psychiatry, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA; Dr. Kobak is with Center for Telepsychology, Madison, Wisconsin, USA; Dr. Lam is with Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada; Dr. MacQueen is with Mathison Centre for Mental Health Research and Education, Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada; Dr. Milev is with Department of Psychiatry, Queen's University, Kingston, Ontario, Canada; Dr. Ravindran is with Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada, and Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; Dr. Sheehan is with Department of Psychiatry and Behavioral Health, University of South Florida College of Medicine, Tampa, Florida, USA; Dr. Sills was with OCBN, Toronto, Ontario, Canada; Dr. Williams is with Columbia University, New York, New York, USA, and Medavante Inc., Trenton, New Jersey, USA.

Innov Clin Neurosci. 2016;13(9–10):20–31

FUNDING: CAN-BIND is an Integrated Discovery Program carried out in partnership with, and financial support from, the Ontario Brain Institute, an independent non-profit corporation, funded partially by the Ontario government. CAN-BIND also acknowledges support from the Canadian Institute for Health Research, Lundbeck A/S, Servier, Bristol Meyers Squibb, Lilly, Johnson and Johnson. Previous iterations of DID were supported by the ISCDD with funding from Eli Lilly and Company. The opinions, results and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred.

FINANCIAL DISCLOSURES: Provided at the end of the article, before the references.

ADDRESS CORRESPONDENCE TO:
Anthony L. Vaccarino, PhD, Indoc Research, 258, Adelaide Street East, Suite 200, Toronto, Ontario, Canada, M5A 1N1, Office: 416-703-0328, E-mail: avaccarino@indocresearch.org

KEY WORDS: Major depressive disorder, rating scales, item response theory, depressive symptoms

ABSTRACT

The Depression Inventory Development project is an initiative of the International Society for CNS Drug Development whose goal is to develop a comprehensive and psychometrically sound measurement tool to be utilized as a primary endpoint in clinical trials for major depressive disorder. Using an iterative process between field testing and psychometric analysis and drawing upon expertise of international researchers in depression, the Depression Inventory Development team has established an empirically driven and collaborative protocol for the creation of items to assess symptoms in major depressive disorder. Depression-relevant

symptom clusters were identified based on expert clinical and patient input. In addition, as an aid for symptom identification and item construction, the psychometric properties of existing clinical scales (assessing depression and related indications) were evaluated using blinded datasets from pharmaceutical antidepressant drug trials. A series of field tests in patients with major depressive disorder provided the team with data to inform the iterative process of scale development. We report here an overview of the Depression Inventory Development initiative, including results of the third iteration of items assessing symptoms related to anhedonia, cognition, fatigue, general malaise, motivation, anxiety, negative thinking, pain and appetite. The strategies adopted from the Depression Inventory Development program, as an empirically driven and collaborative process for scale development, have provided the foundation to develop and validate measurement tools in other therapeutic areas as well.

INTRODUCTION

The Hamilton Depression Rating Scale (HAMD)[1] and Montgomery–Åsberg Depression Rating Scale (MADRS)[2] remain the principal clinician-rated measures of depression severity used by clinical researchers and the pharmaceutical industry and accepted by regulatory bodies to demonstrate pivotal proof of efficacy.[3] Nevertheless, despite their widespread use and acceptance, a number of shortcomings have been identified. For example, studies have shown that many HAMD items discriminate poorly across different levels of depression severity, are not sensitive to change following antidepressant treatment and have poor reliability characteristics.[4–8] Another criticism of the HAMD is that vegetative and somatic symptoms disproportionately contribute to total HAMD score[4,9] and several distinct symptoms can be rated by a single items.[9] The MADRS was introduced as an alternative to the HAMD and generally shows measurement properties that are equal to or better than the HAMD.[5,9–11] However, the MADRS also includes items that discriminate poorly across levels of depressive severity,[5,12,13] and its lack of a structured interview guide and semi-detailed response options can affect the scale's reliability.[14,15] Furthermore, neither the HAMD nor MADRS capture all symptoms denoted by current diagnostic criteria (e.g., reversed neurovegetative symptoms).

To address some of these shortcomings, in 1999 a proposal was made at the National Institute of Mental Health-sponsored New Clinical Drug Evaluation Unit meeting to establish a common set of standards for scoring and administering the HAMD. This proposal led to the formation of the Depression Rating Scale Standardization Team (DRSST), a collaboration of individuals from academia, clinical practice, the pharmaceutical industry, and the United States government. The mission of the DRSST was to develop a standard approach to administering and scoring the HAMD that would remain acceptable to the United States Food and Drug Administration (FDA) and be used by pharmaceutical, academic, and clinical researchers.[4] This included a grid structure that operationalizes intensity and frequency of each item, and allows these to be rated simultaneously. Conventions for administering the scale, as well as a structured interview guide, were developed.[15,16] Consensus was achieved among the working group, and more than 200 worldwide experts in depression were consulted. The GRID-HAMD has been found to be user-friendly with acceptable agreement among independent raters.[16] None-the-less, although the GRID format and structured interview have improved inter-rater reliability, the constituent items are still based on the HAMD and therefore retain the inherent limitations of its psychometric issues and fail to capture many symptoms denoted by current diagnostic criteria or clinical opinion.[4]

In January 2003, a subcommittee of the International Society for CNS Drug Development (ISCDD) was formed to develop a new clinician-rated instrument to assess depression. The goal of the Depression Inventory Development project (DID) team was to develop a measure of depressive symptoms that could set a new standard for assessing major depressive disorder (MDD)—one that includes assessment of the symptoms presently believed to be important to the disease but are not included in existing instruments. As such, it was agreed that any new scale should do the following:

- Reflect diagnostic criteria and conceptualizations of depression
- Have sound psychometric properties
- Be straightforward to administer with similar or shorter length than existing scales
- Have clear symptom definitions, scoring anchors and scoring conventions.

In order to ensure broad input and acceptance, it was also decided that the process should be multidisciplinary and collaborative in nature, drawing its membership from academia, the US government, and the pharmaceutical industry. The availability of an empirically driven and comprehensive instrument would be invaluable in the assessment of depression, particularly in the evaluation of new medications, treatment strategies, and programs.

DID PROCESS AND CONCEPTUAL FRAMEWORK

Using a reciprocal, iterative process between field testing and psychometric analysis and drawing upon expertise of international researchers in depression, a protocol was developed for the creation of new items and vetting them in a clinical population of depressed patients (Figure 1).

One of the most important processes within the DID initiative was determining which symptom clusters are most relevant to MDD and, therefore, need to be measured. As outlined by the FDA Guidance for Patient Reported Outcome (PRO) measures,[17] constructing a new scale requires development of a conceptual framework for the condition under study that establishes the parameters to be measured, the purpose of the measurement tool, and how it will be administered. A key element of the conceptual framework therefore is an understanding of the prevalent, cardinal symptoms that need to be assessed in

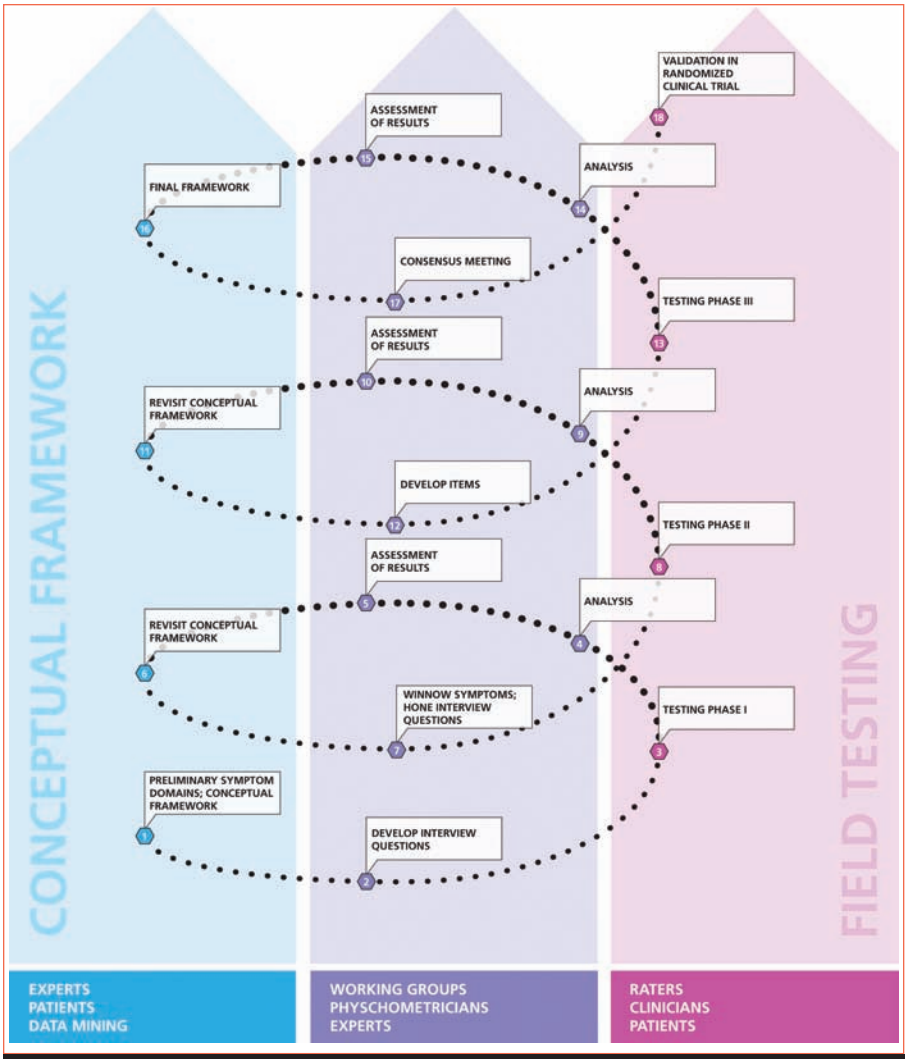


FIGURE 1. This schematic shows the clinical scale development process employed within DID. The process involves development of a conceptual framework for the scale (through input from experts, patients and the literature), development of interview questions to assess symptoms identified during this process, and field testing of items by expert research centers in the target population. Data from this field testing is analyzed synthesized by working group experts before being used to modify the conceptual framework and to guide further interview question development and modification. This process is repeated until all symptoms of interest have been assessed and optimized, a final framework can be assembled, items are finalized for inclusion in the instrument, and formal validation processes undertaken. Figure reproduced with permission: Vaccarino AL, Anderson K, Borowsky B, et al. The Functional Rating Scale Taskforce for Pre-Huntington Disease (FuRST-pHD): an initiative to develop a gold standard instrument to assess early manifestations of Huntington disease prior to formal diagnosis. Presented at International Congress of Parkinson's Disease and Movement Disorders. Toronto, Canada. June 2011.

order to detect differences in the severity of depression. The primary approaches employed in the DID program to identify symptoms that should be assessed are outlined as follows:

Expert clinical input. Input from the medical and scientific community is crucial for identifying symptoms that need to be included in the comprehensive assessment of any disease state. Working groups of experts/opinion leaders were formed to establish and define relevant symptom clusters. The DID Working Groups included Anger and Irritability, Anhedonia, Fatigue, Guilt, Memory and Cognition, Pain, Appetite, Anxiety, Delusions, Functional Impact, and Psychometrics.

The ISCDD provided an ideal platform to bring together leading experts with theoretical and clinical expertise, including expertise in the diagnosis and treatment of MDD and clinical scale development methodology, as well as experience with specific symptom domains. Multidisciplinary expertise throughout development ensured the team had access to important sources of information, but it also helped build consensus within the field regarding the symptoms to be assessed during development and the items that should be included in the final instrument. Furthermore, because the symptomatic profile of MDD is complex and multidimensional, it was recognized that these working groups should include experts with specialized knowledge of individual symptoms beyond the context of depression. The breadth of this expert base contributes to the collaborative structure of the DID project, and adds credibility to the argument that the item is measuring a core symptom of the disease. Working groups were responsible for providing specific input into development of new items (structured interview guides, scoring conventions, scoring anchors, and symptom definitions), evaluation of results, modification of items, and general input into the broader program regarding additional symptom clusters or studies that may be required by the project. Working groups operated through tele- and web conferences, emails, and face-to-face meetings organized in association with conferences widely attended by team members. Literature reviews were also conducted by the DID teams to establish the current thinking on MDD and the measurement of relevant symptoms.

Patient input. To ensure that the scale reflects concepts that are important from the patient's perspective (that can be lost when filtered through a clinical evaluation), input from patients was sought and considered wherever possible. Indeed, the FDA Guidance for PRO measure reflects many of the same principles adopted early on by the DID team in this regard.[17] Patient perspective on the relative importance of symptoms of their disorder is also critical to

decisions made regarding item weighting and symptom inclusion. To address this early on, patients were asked, during field testing, to rate items with respect to item comprehension, importance, and interference with daily activities, as well as have input regarding any additional symptoms they felt should be assessed. These results were used to help select symptom domains that required further testing and for which items to assess specific symptoms would need to be developed. Indeed, saturation of symptom assessment to assure that symptom domains important to the patients have been addressed is a component of the FDA evaluation of new questionnaires.[17]

Targeted data mining. In developing items to assess defined symptom clusters, we have benefited from the use of blinded datasets in which symptoms have been assessed using pre-existing scales. These datasets were acquired from pharmaceutical companies, research centers, and universities and used to identify items or symptoms that display good psychometric properties versus those that showed poor psychometric properties. These datasets provided invaluable information that has implications far beyond the drugs they were designed to evaluate, and in our experience, pharmaceutical companies are increasingly willing to share their data, particularly when they are to be used to increase our understanding of a given disease state. Indeed, the DID team has had tremendous success in gaining access to industry datasets for the purposes of examining symptom profiles in MDD, including the HAMD,[8] Hamilton Anxiety Rating Scale,[18] and Somatic Symptoms Inventory.[19,20]

Item development. Once the symptom domains believed to be most relevant to the disorder were identified, new items to assess specific symptoms within each domain were developed by the working groups. Some HAMD items are recognized as problematic because they assess more than one distinct symptom construct.[9] The HAMD Somatic Symptom General item, for example, simultaneously assesses heaviness in limbs, loss of energy, and headaches, which are distinct symptoms that should be assessed separately because they can be differentially experienced by depressed patients and may have different underlying mechanisms.[9,19] In the DID process, the working groups "deconstruct" each of the general symptom definitions into as many unique, constituent symptom definitions as possible and then develop items to assess each of these. These long lists of items were then tested in the target population in order to determine which aspect of each symptom cluster is most relevant so that the list can ultimately be winnowed into a usable scale. Figure 2 illustrates the DID process for breaking down symptom clusters into testable constituent symptoms.

Format for symptom assessment and scoring. It is also important to consider that an inherent source of variability in measurement is the manner in which symptom information is solicited from an interviewee and translated into an item score. Indeed, altering the manner in which a question is phrased or presented can alter the responses one receives. Since numerous interviewers (and study sites) are often used for data collection, it is important that the wording and presentation of the questions be consistent across sites. Issues related to the order of questions, the wording of the questions, and the need for a structured interview must be standardized if the scale is to be optimized. In addition, the experience of raters and the degree to which they have been trained in the use of a given scale can influence the quality of the data acquired through an interview,[21] and minimizing assessment variability can influence the success of a study.[22] Finally, item scoring strategies can also be a source of variability. If raters are not provided with an unambiguous scoring convention, they will develop idiosyncratic schemas for scoring that can affect the variability across raters and across centers. Clear scoring conventions are required along with clear anchors for each of the scoring options.

To address these issues, we adopted the "GRID" format for item scoring,[16] in concert with a semi-structured interview based on the SIGH-D (Figure 3).[15]

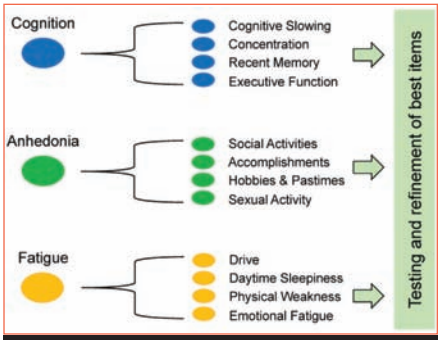
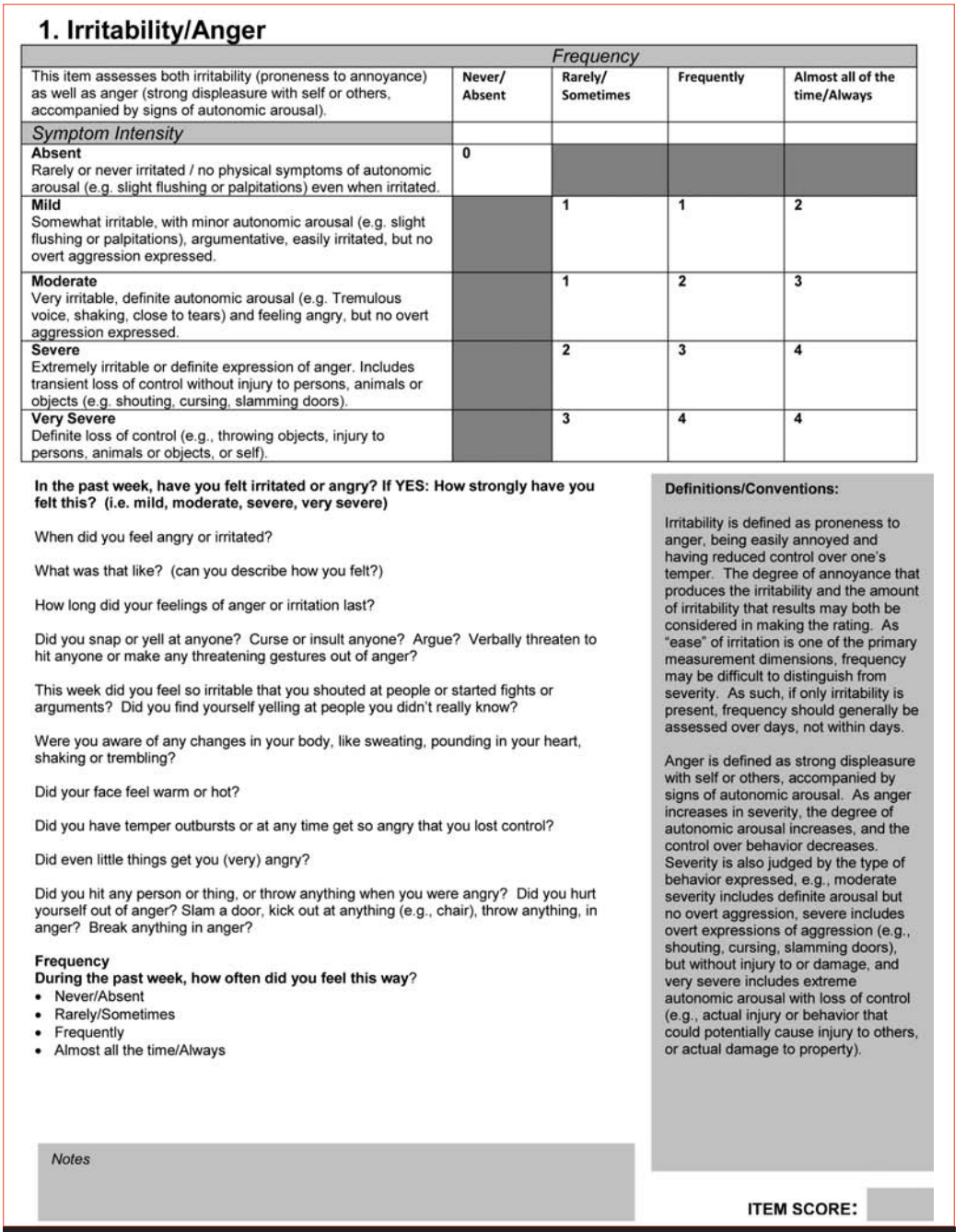


FIGURE 2. Based on input from experts, patients and the literature, symptom domains are identified that are thought to be important to major depressive disorder. The Working Groups identify the constituent symptoms based on conventional definitions and clinical and scientific opinion. Interview questions are then developed for field testing within each of these narrowly defined symptoms.

Originally developed as part of the DRSST project of the ISCDD as a method to ensure a more structured scoring approach to the HAMD, the GRID-HAMD format permits the rater to consider the dimensions of intensity and frequency independently for each relevant item in the scale, while giving them clear scoring anchors, a semi-structured interview guide, and overall definitions and conventions.[16] This method has been employed successfully in other disease states, and has been found to be user-friendly, with acceptable agreement among independent raters.[16,23–30] The GRID-HAMD has been used in clinical trials as a primary outcome measure.[31]

Field testing. Once new items were developed, they were distributed to a team of experienced clinical researchers and raters for field testing, with the results circulated to the broader team for interpretation, discussion, and development of next steps. Multiple sites were used throughout the process, and the depressed populations that were studied reflected those who would be included in clinical trials. During this process the relevance of each symptom, symptom cluster, and dimension in relation to the disorder was determined based on the degree to which each item is able to discriminate individual differences in disease severity using Item Response Theory (IRT).[32,33] IRT is



1. Irritability/Anger

This item assesses both irritability (proneness to annoyance) as well as anger (strong displeasure with self or others, accompanied by signs of autonomic arousal).	Frequency: Never/ Absent	Rarely/ Sometimes	Frequently	Almost all of the time/Always
Symptom Intensity				
Absent Rarely or never irritated / no physical symptoms of autonomic arousal (e.g. slight flushing or palpitations) even when irritated.	0			
Mild Somewhat irritable, with minor autonomic arousal (e.g. slight flushing or palpitations), argumentative, easily irritated, but no overt aggression expressed.		1	1	2
Moderate Very irritable, definite autonomic arousal (e.g. Tremulous voice, shaking, close to tears) and feeling angry, but no overt aggression expressed.		1	2	3
Severe Extremely irritable or definite expression of anger. Includes transient loss of control without injury to persons, animals or objects (e.g. shouting, cursing, slamming doors).		2	3	4
Very Severe Definite loss of control (e.g., throwing objects, injury to persons, animals or objects, or self).		3	4	4

In the past week, have you felt irritated or angry? If YES: How strongly have you felt this? (i.e. mild, moderate, severe, very severe)

When did you feel angry or irritated?

What was that like? (can you describe how you felt?)

How long did your feelings of anger or irritation last?

Did you snap or yell at anyone? Curse or insult anyone? Argue? Verbally threaten to hit anyone or make any threatening gestures out of anger?

This week did you feel so irritable that you shouted at people or started fights or arguments? Did you find yourself yelling at people you didn't really know?

Were you aware of any changes in your body, like sweating, pounding in your heart, shaking or trembling?

Did your face feel warm or hot?

Did you have temper outbursts or at any time get so angry that you lost control?

Did even little things get you (very) angry?

Did you hit any person or thing, or throw anything when you were angry? Did you hurt yourself out of anger? Slam a door, kick out at anything (e.g., chair), throw anything, in anger? Break anything in anger?

Frequency
During the past week, how often did you feel this way?
- Never/Absent
- Rarely/Sometimes
- Frequently
- Almost all the time/Always

Definitions/Conventions:

Irritability is defined as proneness to anger, being easily annoyed and having reduced control over one's temper. The degree of annoyance that produces the irritability and the amount of irritability that results may both be considered in making the rating. As "ease" of irritation is one of the primary measurement dimensions, frequency may be difficult to distinguish from severity. As such, if only irritability is present, frequency should generally be assessed over days, not within days.

Anger is defined as strong displeasure with self or others, accompanied by signs of autonomic arousal. As anger increases in severity, the degree of autonomic arousal increases, and the control over behavior decreases. Severity is also judged by the type of behavior expressed, e.g., moderate severity includes definite arousal but no overt aggression, severe includes overt expressions of aggression (e.g., shouting, cursing, slamming doors), but without injury to or damage, and very severe includes extreme autonomic arousal with loss of control (e.g., actual injury or behavior that could potentially cause injury to others, or actual damage to property).

Notes

ITEM SCORE:

FIGURE 3. Example of Depression Inventory Development project item showing the "GRID" format used by the DID initiative for interview conduct. This includes a GRID structure that operationalizes intensity and frequency of each item, and allows these to be rated simultaneously. In addition, clear symptom definitions and a structured interview guide are provided. Conventions for administering the scale have been previously developed, and the GRID format has been found to be user-friendly and have acceptable agreement among independent raters.[16]

based on the concept that scores on individual items in a scale or standardized test should have a direct relationship with the underlying construct being measured and has been successfully used in the evaluation of various rating scales, including the HAMD,[5,7,8] Hamilton Anxiety Rating Scale,[18] MADRS,[5] Beck Depression Inventory,[34] Somatic Symptoms Inventory,[19] Sexual Interest and Desire Inventory,[24] Unified Huntington's Disease Rating Scale,[35] and Functional Rating Scale for Prodromal Huntington Disease.[26–30] IRT methods can be extremely useful throughout the scale development process, though most IRT methods are traditionally applied in later stages of the process to determine the sensitivity and unidimensionality and for selection of optimal items for inclusion in the final instrument.[32,33] The purpose of applying IRT at an earlier stage of item development is to examine patterns of response of individual DID items which, together with information provided by other statistical analyses, would provide insight into which items should be moved forward, modified, combined, or excluded from further testing. The performance of new items was also evaluated with respect to items from other scales that assess a similar content domain (if such an item exists). All data were circulated back to the working groups and broader team for interpretation, discussion, and development of next steps (Figure 1).

Item modification and iterations. Once items had undergone testing in the target populations, the results were analyzed in a number of ways to determine the degree to which the score on each item is related to the severity of the underlying construct. Items showing poor discriminative properties were removed from further testing, while those with good properties continued in development. When the psychometric properties of a given item were established,

TABLE 1. Reasons for item deletion or modification and supporting analyses.*

ITEM PROPERTY	REASON FOR CHANGE OR DELETION	SUPPORTING ANALYSIS AND CRITERIA
Clarity or relevance	• Reported as not relevant by a large segment of the target population	Percentage of missing data; <10% missing data points considered within acceptable criteria
	• Generates an unacceptably large amount of missing data points	Rasch modelling provides a measure of how difficult or easy an item is relative to the subject and used to assess the relevance of an item to that population
	• Generates an unacceptably large amount of missing data points	n/a
	• Patients interpret items and responses in a way that is inconsistent with the PRO instrument's conceptual framework	n/a
Response range	• A high percent of patients respond at the floor (response scale's worst end) or ceiling (response scale's optimal end)	Percentage scoring zeros or maximum scores; <80% floor (option score of zero endorsed) or ceiling effects (maximum option score endorsed) considered within acceptable criteria
	• Patients note that none of the response choices applies to them	Aggregate frequency of pairs of adjacent options; aggregate frequency of >10% considered within acceptable criteria
	• Distribution of item responses is highly skewed	n/a
Variability	• All patients give the same answer (i.e., no variance)	Frequency of option endorsement; <50% endorsement of one option considered within acceptable criteria
	• Most patients choose only one response choice	
	• Differences among patients not detected when important differences are known	
Reproducibility	• Unstable scores over time when there is no logical reason for variation from one assessment to the next	Present data does not allow assessment of reliability; estimated by correlation of DID items with MADRS items that are assumed to assess the same construct; r>0.75 considered within acceptable criteria
Inter-item correlation	• Item highly correlated (redundant) with other items in the same concept of Interest	Inter-item correlations coefficients; r<0.75 considered within acceptable criteria
Ability to detect change	• Item is not sensitive (i.e., does not change when there is a known change in the concepts of interest)	Present data does not allow assessment of change to therapy; estimated by correlation of DID items with MADRS items that are assumed to assess the same construct; r >0.75 considered within acceptable criteria
Item discrimination	• Item is highly correlated with measures of concepts other than the one it is intended to measure	Correlation of DID items with MADRS items that are assumed to assess the same construct; r >0.75 considered within acceptable criteria
	• Item does not show variability in relation to some known population characteristics (i.e., severity level, classification of condition, or other known characteristic)	IRT used to determine relation to overall depressive severity (total MADRS score)
Redundancy	• Item duplicates information collected with other items that have equal or better measurement properties	Inter-item correlations coefficients; r<0.75 considered within acceptable criteria
Recall period	• The population, disease state, or application of the instrument can affect the appropriateness of the recall period	Not directly assessed; previous literature suggests that recall period is appropriate in this population

*Adapted from references #17 and #19

n/a: not available; PRO: patient reported outcome; DID: depression inventory development project; MADRS: Montgomery–Åsberg Depression Rating Scale; IRT: Item Reponse Theory

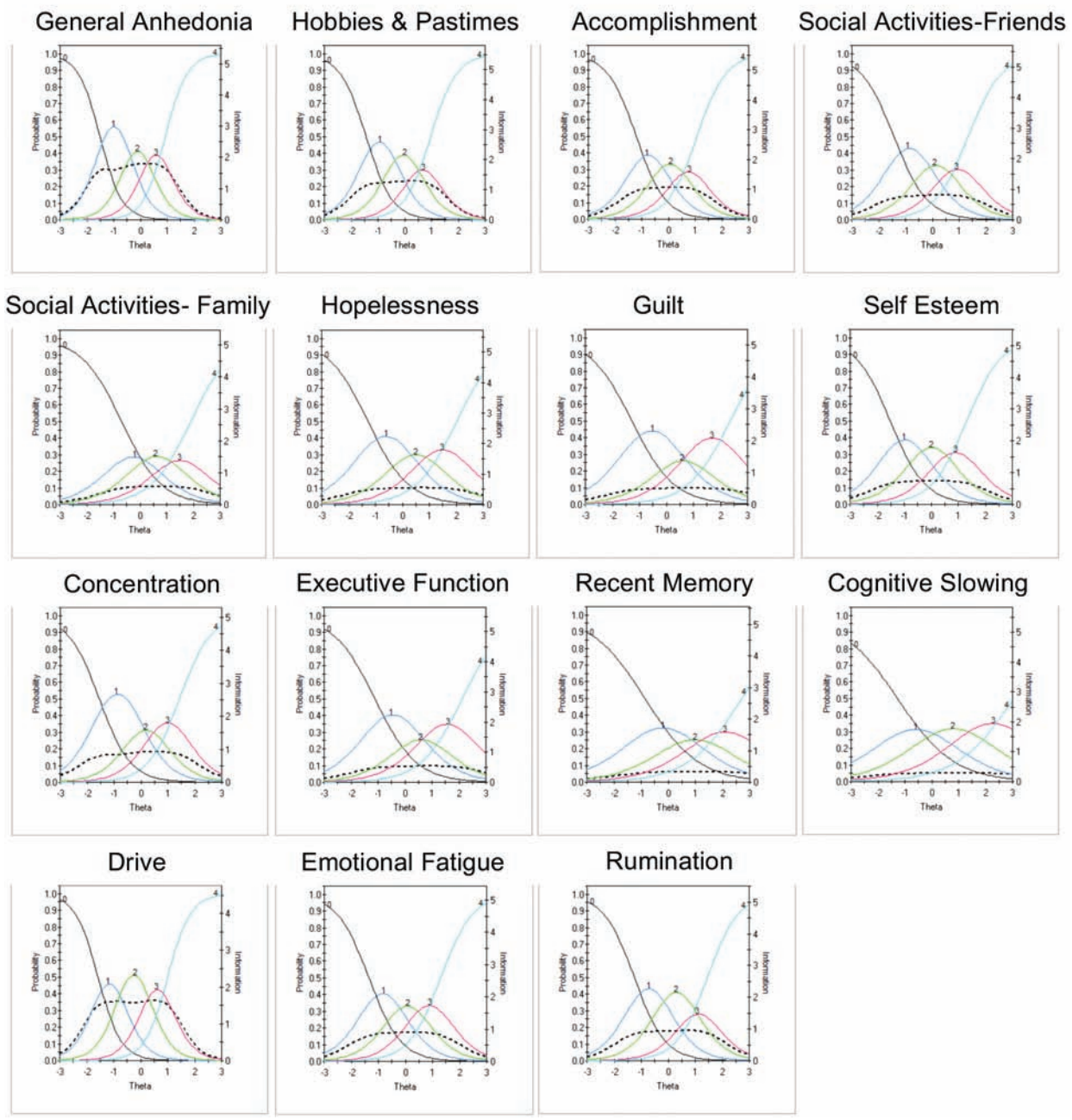


FIGURE 4. Item characteristic curves (ICCs) (smooth lines, left axis) and item information curves (IICs) (dotted line and right axis) for items showing good discriminative properties that will be advanced to the next round of testing.

changes (if necessary) were made on empirical grounds and the modified items were then tested during subsequent stages of the study. To guide decisions about item modification or deletion, we considered the recommendations outlined in the FDA Guidance for PRO Measures.[17] This provided criteria to assess the measurement properties of the DID scale and help determine which items should be kept, deleted, or modified.

Validation. Once all individual items have been developed and field tested, it is necessary to include the final instrument in treatment studies to assess sensitivity to change and ability to separate drug from placebo and to conduct large-scale validation. In the past, pharmaceutical company membership has agreed to include the instrument in future drug development trials to meet this need. This is expected to facilitate discussions with regulatory bodies to ensure the scale is acceptable from that perspective as well.

CURRENT STATUS: RESULTS OF THIRD ITERATION OF DID ITEMS

Building on previous iterations,[36] the DID initiative has now completed its third round of testing. We report here the results of the third iteration of DID items, which includes evaluation of items to assess symptoms related to anhedonia, cognition, fatigue, general

malaise, motivation, anxiety, negative thinking, pain, and appetite.

METHODS

Item administration and data collection. The present study was included as part of a larger multi-site, open label study conducted by the Canadian Biomarker Integration Network in Depression (CAN-BIND) to investigate biomarkers of antidepressant treatment response in patients with depression.[37] Male and female outpatients (18–60 years old) whose symptoms met the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (*DSM-IV-TR*) criteria for a major depressive episode in MDD as determined by the Mini International Neuropsychiatric Interview[38] and had a minimum MADRS score of 24 participated in the study. The CAN-BIND study included clinic visits over a 16-week period during which patients (and healthy controls) underwent clinician-administered scales and self-reports and cognitive testing and neuroimaging assessment (structural and functional magnetic resonance imaging [MRI] and electroencephalogram [EEG] and provided blood and urine samples. The CAN-BIND study was carried out in accordance with the Declaration of Helsinki and the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidelines, and the study design and procedures were reviewed by the appropriate ethics committees; informed consent was obtained from participants after full explanation of the nature of the procedures.

Psychometricians with experience rating depressive symptoms in clinical trial research administered the test items. Patients were administered 34 DID test items using a semi-structured interview adapted from the GRID-HAMD.[16] All raters received training in standardized conventions for scoring, anchor, and item definitions and in the use of the DID structured interview guide. Patients were also administered the MADRS, which allowed the DID items to be evaluated against these existing "benchmark" items. Data (DID, MADRS, demographics) were captured electronically in OpenClinica Enterprise (Waltham, Mass., USA) using the Brain-CODE Platform. Brain-CODE is an extensible informatics platform that manages the acquisition and storage of multidimensional data collected from patients with a variety of brain disorders. Brain-CODE is housed at the Centre for Advanced Computing in Kingston Ontario (www.braincode.ca). The data presented here are from an interim data release (CBN01_1_A01) that included demographic, DID, and MADRS measures for 85 patients on each of two post-screen visits (Baseline [n=85] and Week 8 [n=74]). In total, there were 159 patient visits that included DID and MADRS measures. Data were also supplied for 49 healthy controls. No treatment code or information that could be used to identify a subject was included in the data. DID items were scored on a grid of intensity and frequency, with the combination thereof comprising a composite score.[16] The DID item composite score was the subject of analysis. To ensure a broad range of coverage of depressive severity data were pooled across the visits.

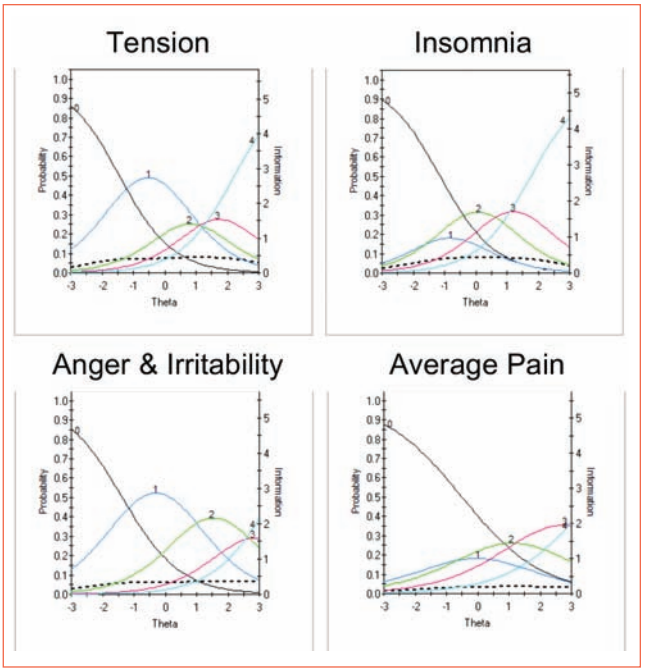


FIGURE 5. Item characteristic curves (ICCs) (smooth lines, left axis) and item information curves (IICs) (dotted line and right axis) for items found to discriminate across levels of severity, with visual examination of these ICCs providing data on which decisions can be made to improve item performance

Items analysis and review. Using FDA recommendations as guidance,[17] standard psychometric criteria were applied to identify which DID items should be eliminated from future iterations and which should be advanced (Table 1).[17,32,39] In addition, IRT was used to assess how informative an individual DID item was as a measure of the depressive severity. In this setting, we assume one underlying latent trait: depression severity as assessed by MADRS total score. Individual DID items were analyzed one at a time against the 10 MADRS items. IRTPro Version 3.0 software (Scientific Software International, Skokie, Ill., USA) was used to generate item characteristic curves (ICCs) and item information curves (IICs). ICCs display the probability of a particular option endorsement for each DID item as a function of trait level (MADRS). An item is considered informative if characterized by a clear identification of the range of severity scores over which an option is most likely to be endorsed, rapid changes in the curves that correspond to changes in severity, and an orderly relation between the weight assigned to the option and the region of severity over which an item is likely to be endorsed. IICs are also useful as they provide insight into the measurement precision of an item. Increased slope of the line indicates the item provides more information, leading to smaller standard errors of measurement. IICs also illustrate at what trait level the

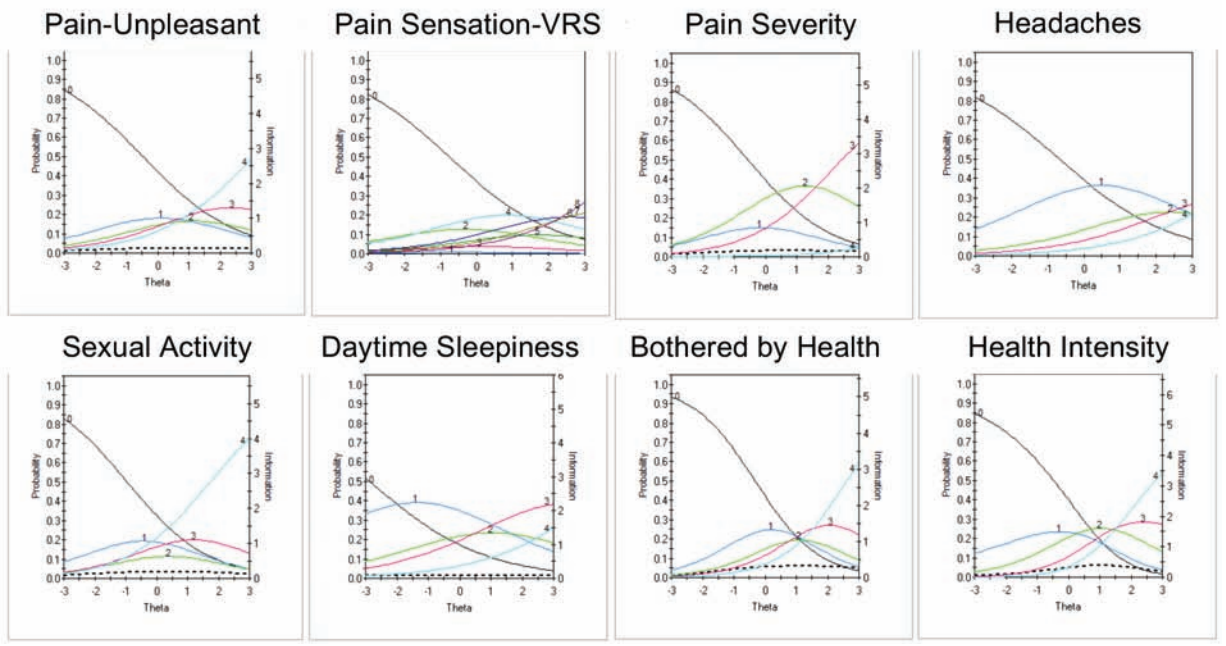


FIGURE 6. Item characteristic curves (ICCs) (smooth lines, left axis) and item information curves (IICs) (dotted line and right axis) for items showing poor discriminative properties

item is most informative; such that an item's psychometric qualities can vary across different levels of severity. Items contributing little information are considered to contribute little precision to a test. All analyses were conducted within the Brain-CODE Platform.

RESULTS

Patient characteristics. Baseline demographics and clinical characteristics are shown in Table 2.

Option scoring frequencies. Option scoring frequencies for each individual DID item were examined to determine the presence of floor or ceiling effects (>80% endorsement of zero or maximum score, respectively), low scoring variability (single option frequency of endorsement >50%), and skewed scoring (aggregate scoring frequency <10%) (Table 1). When these criteria were applied, unacceptable item scoring was observed for seven items and these were excluded from further analyses: Eating-Increase (floor effect and skewed), Appetite-Increase (low scoring variability and skewed), Prolonged Nighttime Sleep (low scoring variability), Appetite-Decrease (low scoring variability), Eating-Decrease (low scoring variability), Physical Weakness (low scoring variability), and Pain unpleasantness–Verbal Rating Scale (skewed). Option scoring frequencies for all DID items are available as supplemental material online.

Item response theory. The remaining 27 items were subject to additional psychometric evaluation using IRT. Of the 27 DID items that remained, 15 items showed good discriminative properties across a broad range of severity, including ICCs with rapid changes in the curves that correspond to changes in severity and an orderly relationship between the weight assigned to the option and the region of severity over which an item is likely to be endorsed (General Anhedonia, Hobbies and Pastimes, Accomplishment, Social Activities with Friends, Social Activity with Family, Hopelessness, Guilt, Self-Esteem, Concentration, Executive Function, Recent Memory, Cognitive Slowing, Drive, Emotional Fatigue, Rumination) (Figure 4). Examination of IICs revealed differences in the amount of information each item contributed. For example, within the anhedonia-related items, General Anhedonia provided more information in estimating severity, as compared to Hobbies and Pastimes, Accomplishment, Social Activities with Friends, and Social Activity with Family. Similarly, within the cognitive-related items, Concentration was more informative than Executive Function, Recent Memory, and Cognitive Slowing.

Four additional items discriminated across levels of severity, with visual examination of these ICCs providing data on which to base decisions about improving performance, including changes in wording and/or scoring options (Tension, Insomnia, Anger and Irritability, Average Pain) (Figure 5). For example, examination of the Anger and Irritability item reveals an equal probability of endorsing Options 3 or 4 at higher levels of severity, suggesting that patients do not discriminate between these options as currently defined.

The remaining eight items showed poor discriminative properties that may limit their clinical utility: Pain Unpleasant, Pain Sensation-Verbal Rating Scale, Pain Severity, Headaches, Sexual Activity, Daytime Sleepiness, and Bothered by Health (Health Intensity) (Figure 6). Although these items generally showed poor

discriminative properties, some were endorsed at higher levels of severity and thus may be useful for assessment in more severe depression.

DISCUSSION

Data were reviewed by the DID team, and decisions were made with respect to items that should be advanced and items that should be excluded from further development. It was agreed that the 19 items that showed favorable measurement characteristics would be advanced to the next rounds of testing, including some item modifications as required (Figures 4 and 5). Although it is clear that some of the items are measuring the same construct and thus can be considered redundant (e.g., anhedonia-related items), additional item modification will be required to determine which item displays better measurement properties and thus should be advanced. This may include combining items that are likely assessing the same construct. The next iteration of items will also include those that have already demonstrated good psychometric properties in the previous iteration, including Depressed Mood and Anxiety.[36]

In the present iteration, five pain-related items were tested to assess different aspects of pain. Although not considered a core diagnostic feature for depression, painful somatic symptoms are commonly reported in MDD,[19,20,40,41] including in primary care setting,[42] and are associated with poorer clinical outcome.[43,44] The MADRS does not assess pain, and the HAMD, although designed to include the assessment of somatic symptoms, does not differentiate pain from other somatic symptoms. In the present study, all DID pain items scored higher in MDD than in healthy controls (all $p<0.001$) and significantly correlated with total MADRS score. Although these pain items did not discriminate at lower levels of depression, the Average Pain item (Figure 5) and to a lesser degree the remaining pain items (Figure 6), were found to discriminate at higher levels of depression severity. Inclusion of items that are sensitive in more severe depression will be required in the final scale to ensure broad symptom coverage and sensitivity.

The development and validation of DID is an iterative process in which item retention, modification, and deletion are based on empirical evidence obtained during field testing. It is unknown at present how many of the items currently under investigation will display sufficiently good psychometric properties to justify their inclusion in a final instrument. However, a key practical consideration in this regard is the time taken to complete the assessment; the DID scale should not take longer than other standard interviews used to assess depression severity (e.g., the HAMD, which should take up to 30 minutes to complete). There is a practical limit with regard to how many items can be included in the final instrument and it is possible that a number of items will either be collapsed into other items or removed (e.g., when two items appear to be equally efficient at assessing a given range of severity or participant profile).

In the initial stages of the DID initiative, we decided to give priority to developing clinician-rated rather than patient-rated questions for several reasons. Many of the symptoms and functions to be assessed involve concepts that may be incompletely understood by a non-clinician unless the wording is clear, precise, and at an appropriate literacy level; this latter point is stressed in the FDA PRO guidance.[17] In a self-report scale, there is no possibility to counter ambiguities in wording or interpretation or to ensure comprehension other than to obtain covariate measures of literacy. With an interview format, on the other hand, a trained interviewer/rater can probe each scale item to ensure the concept is understood. In addition, interviewers can provide feedback as they gain experience with each of the newly developed interview questions, and this can be informative as to the utility of the questions, concepts and definitions.

The DID team has followed the recommendations outlined in the FDA PRO guidance to assess and establish the measurement properties of the final scale.[17] Validation is an ongoing process, and initial steps have been taken to address content and face validity of the instrument through the involvement of clinical experts and patients during the item development process. Additional patient input will also be sought as required, including cognitive debriefing, to assess comprehension and ensure concepts are fully understood and determine symptom saturation from the patient's perspective. Similarly, cross-cultural differences in understanding and reporting of symptoms will need to be taken into account, especially since a universally deployed rating scale will need to be available in many languages. Most aspects of a full validation of the scale, including aspects of reliability (internal consistency, inter-rater, test-retest) and validity (concurrent, discriminant, convergent) will require a mature version of the developing instrument and thus will be completed near the end of the project. The ultimate goal is to achieve construct validity with

TABLE 2. Baseline demographic and clinical characteristics

DEMOGRAPHICS AND CHARACTERISTICS	PATIENTS WITH MDD (n=85)	HEALTHY CONTROLS (n=49)
Female, n (%)	50 (59%)	32 (65%)
Age in years, Mean (range)	36.05 (19–61)	32.53 (20–57)
MADRS, Mean±SEM	29.92±0.66	0.44±0.16
CGI-S, Mean±SEM	4.70±0.08	1.00±0.00

MADRS: Montgomery–Åsberg Depression Rating Scale; CGI-S: Clinical Global Impressions-Severity

the measurement tool and to ensure the ability of the scale to detect change in randomized, controlled trials of individuals with MDD.

ACKNOWLEDGMENTS

We would like to acknowledge the individuals and organizations that have made data used for this research available, including CAN-BIND, the Ontario Brain Institute, the Brain-CODE platform, and the government of Ontario. We thank the CAN-BIND coordinators and raters.

DID working group members. R Michael Bagby, Stephen Brannan, David DeBrota, David Daniel, Andrew Culter, Judith Dunn, Nina Engelhardt, Kenneth Evans, Maurizio Fava, Alan Feiger, James Ferguson, Alastair Flint, Susan Gilbert-Evans, John Greist, James Hartford, Phillip Harvey, Mojib Javadi, Amir Kalali, Joel Katz, James Kennedy, Sidney Kennedy, Ken Kobak, Joseph Kwentus, Raymond Lam, Joshua Lipsitz, Heather McDonald, Roger McIntyre, Ronald Melzack, James Mundt, Phil Ninan, Jason Olin, Jay Pearson, William Potter, Penny Randall, James Russel, Darcy Santor, Alan Schatzberg, David Sheehan, Terrence Sills, Anthony Vaccarino, David Walling, Keith Wesnes, Janet Williams, Glen Wunderlich.

FINANCIAL DISCLOSURES

Drs. Vaccarino, Evans, Kalali, Engelhardt, Greist, Mac Queen, Milev, Placenza, Ravindran, and Williams report no conflicts of interest relevant to the content of this article.

Dr. Kennedy has served on advisory boards or similar committees for Allergan, AstraZeneca, BMS, Janssen, Lundbeck, Lundbeck Institute, Pfizer, Servier, St. Jude Medical, and Sunovion; has participated in clinical trials or studies for Brain Cells Inc., BMS, Clera, Janssen, Pfizer, Servier, and St. Jude Medical; has been a speaker for AstraZeneca, BMS, Eli Lilly, Lundbeck, Pfizer, Servier; and has received research support from BMS, Brain Canada, CIHR, Janssen, Lundbeck ,OBI, OMHF, Pfizer, and Servier.

Dr. Frey has received grant/research support from Alternative Funding Plan Innovations Award, Brain and Behavior Research Foundation, Canadian Institutes of Health Research, Hamilton Health Sciences Foundation, J.P. Bickell Foundation, Ontario Brain Institute, Ontario Mental Health Foundation, Society for Women's Health Research, Teresa Cascioli Charitable Foundation, Eli Lilly, and Pfizer, and has received consultant and/or speaker fees from AstraZeneca, Bristol-Myers Squibb, Canadian Psychiatric Association, CANMAT, Daiichi Sankyo, Lundbeck, Pfizer, Servier and Sunovion.

Dr. Kobak is owner of Center for Telepsychology, which offers online clincian training. Dr. Lam has received speaker honoraria from AstraZeneca, Canadian Psychiatric Association, Canadian Network for Mood and Anxiety Treatments, Lundbeck, Lundbeck Institute, Otsuka, and Servier; has served on consulting/ advisory boards for Bristol Myers Squibb, Canadian Depression Research and Intervention Network, Canadian Network for Mood and Anxiety Treatments, Eli Lilly, Johnson and Johnson, Lundbeck, Mochida, Pfizer, and Takeda; has received research funding (through UBC) from Brain Canada, Bristol Myers Squibb, Canadian Institutes of Health Research, Canadian Network for Mood and Anxiety Treatments, Coast Capital Savings, Lundbeck, Pfizer, St. Jude Medical, University Health Network Foundation, and Vancouver Coastal Health Research Institute; and holds patent and co,pyright for Lam Employment Absence and Productivity Scale (LEAPS).

REFERENCES

1. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23:56–62.
2. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry.* 1979;134:382–389.
3. Reidel M, Möller, H-J, Obermeier M, et al. Response and remission criteria in major depression: a validation of current practice. *J Psychiatr Res.* 2010;44(15):1063–1068.
4. Bagby RM, Ryder AG, Schuller DR, et al. The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *Am J Psychiatry.* 2004;161(12):2163–2177.
5. Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol.* 2006;16(8):601–611.
6. Faries D, Herrera J, Rayamajhi J, et al. The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res.* 2000;34(1):3–10.
7. Santor DA, Coyne JC. Examining symptom expression as a function of symptom severity: item performance on the Hamilton Rating Scale for Depression. *Psychol Assess.* 2001;13(1):127–139.
8. Evans KR, Sills T, DeBrota DJ, et al. An item response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. *J Psychiatr Res.* 2004;38(3):275–284.
9. Zimmerman M, Posternak MA, Chelminski I. Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? *J Clin Psychopharmacol.* 2005;25(2):105–110.
10. Furukawa TA. Assessment of mood: guide for clinicians. *J Psychosom Res.* 2010;68(6):581–589.
11. Mulder RT, Joyce PR, Frampton C. Relationships among measures of treatment outcome in depressed patients. *J Affect Disord.* 2003;76(1–3):127–135.
12. Adler M, Hetta J, Isacsson G, et al., An item response theory evaluation of three depression assessment instruments in a clinical sample, *BMC Med Res Methodol.* 2012;12:1–12.
13. Davidson J, Turnbull CD, Strickland R, et al. The Montgomery-Asberg Depression Scale: reliability and validity. *Acta Psychiatr Scand.* 1986;73(5):544–548.
14. Iannuzzo RW, Jaeger J, Goldberg JF, et al. Development and reliability of the HAM-D/MADRS interview: an integrated depression symptom rating scale. *Psychiatry Res.* 2006;145(1):21–37.
15. Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry.* 1988;45(8):742–747.
16. Williams JB, Kobak KA, Bech P, et al. The GRID-HAMD: standardization of the

Hamilton Depression Rating Scale. *Int Clin Psychopharmacol.* 2008;23(3):120–129.
17. United States Department of Health and Human Services. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. *Fed Reg.* 2009;74:65132–65133.
18. Vaccarino AL, Evans KR, Sills TL, et al. Symptoms of anxiety in depression: assessment of item performance of the Hamilton Anxiety Rating Scale in patients with depression. *Depress Anxiety.* 2008;25(12):1006–1013.
19. Vaccarino AL, Sills TL, Evans KR, et al. Prevalence and association of somatic symptoms in patients with major depressive disorder. *J Affect Disord.* 2008;110(3):270–276.
20. Vaccarino AL, Sills TL, Evans KR, et al. Multiple pain complaints in patients with major depressive disorder. *Psychosom Med.* 2009;71(2):159–162.
21. Kobak KA, Lipsitz J, Williams JB, et al. Are the effects of rater training sustainable? Results from a multicenter clinical trial. *J Clin Psychopharmacol.* 2007;27(5):534–535.
22. Müller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol.* 2002;22(3):318–325.
23. de Oliveira IR, Seixas C, Osório FL, et al. Evaluation of the psychometric properties of the Cognitive Distortions Questionnaire (CD-Quest) in a sample of undergraduate students. *Innov Clin Neurosci.* 2015;12(7-8):20–27.
24. Sills T, Wunderlich G, Pyke R, et al. The Sexual Interest and Desire Inventory-Female (SIDI-F): item response analyses of data from women diagnosed with hypoactive sexual desire disorder. *J Sex Med.* 2005;2(6):801–818.
25. Tabuse H, Kalali A, Azuma H, et al. The new GRID Hamilton Rating Scale for Depression demonstrates excellent inter-rater reliability for inexperienced and experienced raters before and after training. *Psychiatry Res.* 2007;153(1):61–67.
26. Vaccarino AL, Sills T, Anderson KE, et al. Assessment of cognitive symptoms in prodromal and early Huntington disease. *PLoS Curr.* 2011;3:RRN1250.
27. Vaccarino AL, Sills T, Anderson KE, et al. Assessment of day-to-day functioning in prodromal and early Huntington's disease. *PLoS Curr.* 2011;3:RRN1262.
28. Vaccarino AL, Sills T, Anderson KE, et al. Assessing behavioural manifestations prior to clinical diagnosis of Huntington's disease: "anger and irritability" and "obsessions and compulsions." *PLoS Curr.* 2011;3:RRN124.
29. Vaccarino AL, Sills T, Anderson KE, et al. Assessment of motor symptoms and functional impact in prodromal and early Huntington's disease. *PLoS Curr.* 2011;2:RRN1244.
30. Vaccarino AL, Sills T, Anderson KE, et al. Assessment of depression, anxiety and apathy in prodromal and early Huntington's disease. *PLoS Curr.* 2011;3:RRN1242.
31. Nakagawa A, Mitsuhiro S, Dai M, et al. Effectiveness of cognitive behavioural therapy augmentation in major depression treatment (ECAM study): study protocol for a randomised clinical trial. *BMJ Open.* 2014;4(10):1–12.
32. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther.* 2014;36(5):648–662.
33. Hays, RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care.* 2000;38(9 Suppl):II28–II42
34. Kim Y, Pilkonis PA, Frank E, et al. Differential functioning of the Beck Depression Inventory in late-life patients: use of item response theory. *Psychol Aging.* 2002;17(3):379–391.
35. Vaccarino AL, Anderson K, Borowsky B, et al. An item response analysis of the motor and behavioral subscales of the unified Huntington's disease rating scale in Huntington's disease gene expansion carriers. *Mov Disord.* 2011;26(5):877–884.
36. Kalali A, Vaccarino A, Evans K, et al. The depression inventory development workgroup: results from the second validation study. *Euro Neuropsychopharm.* 2008;18 Suppl 4:S298.
37. Lam RW, Milev R, Rotzinger S, et al. Discovering biomarkers for antidepressant response: protocol from the Canadian Biomarker Integration Network in Depression (CAN-BIND) and clinical characteristics of the first patient cohort. *BMC Psychiatry.* 2016;16:105.
38. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry.* 1998;59 Suppl 20:22–33.
39. Lamping DL, Schroter S, Marquis P, et al. The community-acquired pneumonia symptom questionnaire: a new, patient-based outcome measure to evaluate symptoms in patients with community-acquired pneumonia. *Chest.* 2002;122(3):920–929.
40. Bair MJ, Robinson RL, Katon W, et al. Depression and pain comorbidity: a literature review. *Arch Intern Med.* 2003;163(20):2433–2445.
41. Currie SR, Wang J. Chronic back pain and major depression in the general Canadian population. *Pain.* 2004;107(1-2):54–60.
42. Tylee A, Gandhi P. The importance of somatic symptoms in depression in primary care. *Prim Care Companion J Clin Psychiatry.* 2005;7(4):167–176.
43. Kroenke K, Shen J, Oxman TE, et al. Impact of pain on the outcomes of depression treatment: results from the RESPECT trial. *Pain.* 2008;134(1-2):209–215.
44. Ohayon MM, Schatzberg AF. Using chronic pain to predict depressive morbidity in the general population. *Arch Gen Psychiatry.* 2003;60(1):39–47. ■