Universiteit Leiden
The Netherlands

Revisiting the theoretical and methodological foundations of depression measurement

Fried, E.I.; Flake, J.K.; Robinaugh, D.J.

Citation

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology, 1*, 358–368. doi:10.1038/s44159-022-00050-2

Version:	Publisher's Version
License:	Licensed under Article 25fa Copyright Act/Law (Amendment Taverne)
Downloaded from:	https://hdl.handle.net/1887/3515035

Note: To cite this publication please use the final published version (if applicable).

Check for updates

Revisiting the theoretical and methodological foundations of depression measurement

Eiko I. Fried, Jessica K. Flake and Donald J. Robinaugh

Abstract | Depressive disorders are among the leading causes of global disease burden, but there has been limited progress in understanding the causes of and treatments for these disorders. In this Perspective, we suggest that such progress depends crucially on our ability to measure depression. We review the many problems with depression measurement, including the limited evidence of validity and reliability. These issues raise grave concerns about common uses of depression measures, such as for diagnosis or tracking treatment progress. We argue that shortcomings arise because the measurement of depression rests on shaky methodological and theoretical foundations. Moving forward, we need to break with the field's tradition, which has, for decades, divorced theories about depression from how we measure it. Instead, we suggest that epistemic iteration, an iterative exchange between theory and measurement, provides a crucial avenue for progressing how we measure depression.

Major depressive disorder (MDD) — a prevalent, debilitating and often recurrent mental disorder of an episodic nature — is one of the most frequently measured constructs in the scientific literature. More than 280 measures of this mental health condition have appeared in the literature in the past century[1]. These include three scales which are among the 100 most cited papers across all fields of science[2]: the Hamilton Rating Scale for Depression (HRSD)[3], the Beck Depression Inventory (BDI)[4] and the Centre for Epidemiological Studies Depression Scale (CES-D)[5]. These papers have a combined total of 81,000 citations since 1960 (see Supplementary Figure 1 and Supplementary Note 1); according to Web of Science, each has been cited in more than 140 distinct disciplines. Papers introducing abbreviated, translated and adapted versions of these scales contribute thousands more citations.

With so much empirical research on depression, one would expect there to have been considerable advances in understanding depression and the ability to treat it. Unfortunately, progress has been limited. The prevalence and global disease burden of MDD have not decreased over the past three decades[6]. Despite sizeable efforts, researchers have been unable to identify actionable biomarkers for MDD that explain sufficient variance in diagnosis to be useful in clinical settings[7,8]. Further, the efficacies of both psychological and pharmacological treatments remain limited[9,10].

In this Perspective, we take the position that progress in understanding, predicting and treating depression depends crucially on the ability to measure it. We first provide a brief history of depression measurement. Next, we describe the many problems with depression measurement, including limited evidence of validity and reliability. We argue that these problems arise because the measurement of depression rests on shaky methodological and theoretical foundations. We conclude by offering ideas for moving methodological and theoretical aspects of depression measurement into the twenty-first century.

A brief history

In the middle of the twentieth century, psychoanalytic theory and practice dominated psychiatry[11,12]. Diagnoses were defined by narrative descriptions and assessed by unstructured interviews, leaving considerable room for subjectivity[13]. Perhaps not surprisingly, the agreement of two psychiatrists on whether a patient had a given mental disorder was barely above chance[14,15].

During the 1960s and 1970s, there was a concerted effort to increase diagnostic reliability by developing diagnostic criteria sets: lists of readily observable or reportable experiences with explicit algorithms for determining the presence or absence of a disorder on the basis of these signs and symptoms[16]. This effort culminated in 1980 with the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III), American psychiatry's authoritative publication on the diagnosis of mental disorders[17]. DSM-III leaned heavily on recently developed criteria sets and aimed to be theoretically agnostic, focusing on symptoms rather than aetiology or underlying mechanisms.

In this context, the most commonly used depression scales such as the HRSD, BDI and CES-D were developed and began to take hold in the field. Diagnostic interviews (for example, based on DSM criteria) aim to determine the presence or absence of MDD. By contrast, scales such as the HRSD, BDI and CES-D were developed to assess the severity of current depressive symptoms, with scores above a certain threshold indicating potential MDD. The various depression scales that arose during this era differ in several ways, including the number and nature of assessed symptoms and mode of assessment; for example, the BDI and CES-D are self-rated, whereas the HRSD is observer-rated. However, they share a common approach: measuring the severity of depression by adding together the symptoms of depression, such as depressed mood, sleep problems and suicidal ideation. This approach based on symptoms and sum scores is identical in virtually all of the depression instruments that have appeared in the literature, including self-report and observer-rated scales as well as clinician-rated diagnostic criteria[1,18].

In the decades since this shift towards diagnostic criteria and standardized scales, depression research has thrived, but the measurement of depression has remained

strikingly unchanged. Since the HRSD was published over half a century ago, we have put a man on the Moon, invented the Internet and created powerful computers small enough to fit in people's pockets.

Yet the HRSD remains the gold-standard scale for depression, used in more than 90% of antidepressant trials[19]. Given the enormous amount of depression research and the substantial gains made in psychological measurement practices in the past few decades, it is worth taking stock of depression measurement. We focus our investigation on the most important aspects of validity and reliability. These and other key terms are defined in BOX 1.

Box 1 | **Key terms and definitions**

Alpha (coefficient alpha, Cronbach's alpha)
Internal consistency is often summarized with coefficient alpha. Alpha ranges from 0 to 1, with higher numbers indicating more consistency. Alpha does not provide information about scale validity, and is often not appropriate for depression instruments owing to strict assumptions that are rarely met[147].

Depression instrument
A depression instrument is a measure of depression. Common instruments include self-rated and observer-rated scales typically used to assess depression severity, and structured or semi-structured clinical interviews typically used to assess the presence of major depressive disorder (MDD).

Depression scale
A depression scale is a particular type of instrument to measure the severity of depression. Depression scales can be self-rated or observer-rated. These scales typically include a list of depression symptoms rated on a brief ordinal scale indicating frequency (how common is a symptom), intensity (how severe is a symptom), relativity (compared with usual, what is the symptom expression) or a mix of the above.

Diagnostic interview
A diagnostic interview is a particular type of instrument to measure the presence of MDD. Diagnostic interviews are usually structured or semi-structured. They typically include a list of depression symptoms coded as present or absent, and a question about impairment of functioning. A specific algorithm determines presence of the disorder.

Dimensionality
A unidimensional instrument is one that can aptly describe or summarize the relations among items of a construct with only one score (that is, one dimension, factor or component). In such instruments, it is defensible to add up all items to one total score, which reflects the single dimension. Depression instruments are often multidimensional, meaning that more than one score is required to describe the relations among items adequately.

Inter-rater reliability
In the context of depression measurement, inter-rater reliability is the degree to which independent observers (usually two) agree on whether a person should receive a diagnosis of MDD or not.

Internal consistency
Internal consistency quantifies how consistent responses to items on a scale are. A scale is internally consistent if all of its items produce similar scores.

Measurement invariance
If an instrument measures the same construct in the same way across populations or time, it has the psychometric property of measurement invariance. This property is necessary to accurately compare scores across populations or time.

Kappa coefficient
Inter-rater reliability is commonly assessed using the kappa coefficient, which ranges from 0 to 1. Higher numbers indicate more agreement.

Reliability
Reliability or precision denotes the consistency of scores across instances of the testing procedure, such as raters, time, items and context. Reliability is necessary but not sufficient for validity.

Response process
The response process denotes the cognitive processes engaged in by people using an instrument. In depression research, these people can be the participants filling out self-rated instruments or being interviewed, observers scoring observer-rated scales or clinicians administering an interview.

Test score
The test score is the resulting score from depression instruments, usually a continuous sum score indicating depression severity, or a categorical score with two groups, healthy and depressed.

Validity
As defined by the Standards for Educational and Psychological Testing[25], validity "… refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests […]. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations."

Validity and reliability

The most fundamental question for any measure concerns its validity: whether it measures what it purports to measure[20]. This question turns out to be difficult to answer, and there are many theoretical and methodological frameworks for evaluating validity[16,21–24]. Here, we adopt the perspective taken by the Standards for Educational and Psychological Testing and consider validity to be the degree to which the evidence supports the interpretation and use of a test score for a specific purpose[25]. Examples of such purposes include using DSM-5 criteria to diagnose a person with an episode of MDD or using the BDI to track treatment progress over time. Validation entails accumulating evidence to form a sound scientific basis for using instruments for these purposes.

We will consider three sources of evidence for validity — content, response processes and internal structure — and will evaluate whether the evidence supports common uses of depression instruments. These three sources do not represent distinct types of validity; instead, they together support the intended interpretation of scores for a given use[25]. We also discuss reliability — the consistency of instruments across raters, contexts and time. Reliability does not provide evidence of validity, given that a score can be consistent but not hold the intended interpretation. Reliability is therefore necessary, but not sufficient, for validity.

Content

One source of evidence for the valid use of depression instruments is measure content. A valid score must reflect all of the content needed to describe a construct, avoiding construct under-representation (omitting important content) and construct contamination (including construct-irrelevant content)[26]. Evidence of adequate content is critical for many uses of depression instruments, such as communication: if a patient who has been diagnosed is referred from one therapist to another, the diagnosis is useful to the new therapist only if the instrument used for diagnosis actually captures content relevant to depression. Appropriate content coverage is also required for many other purposes,

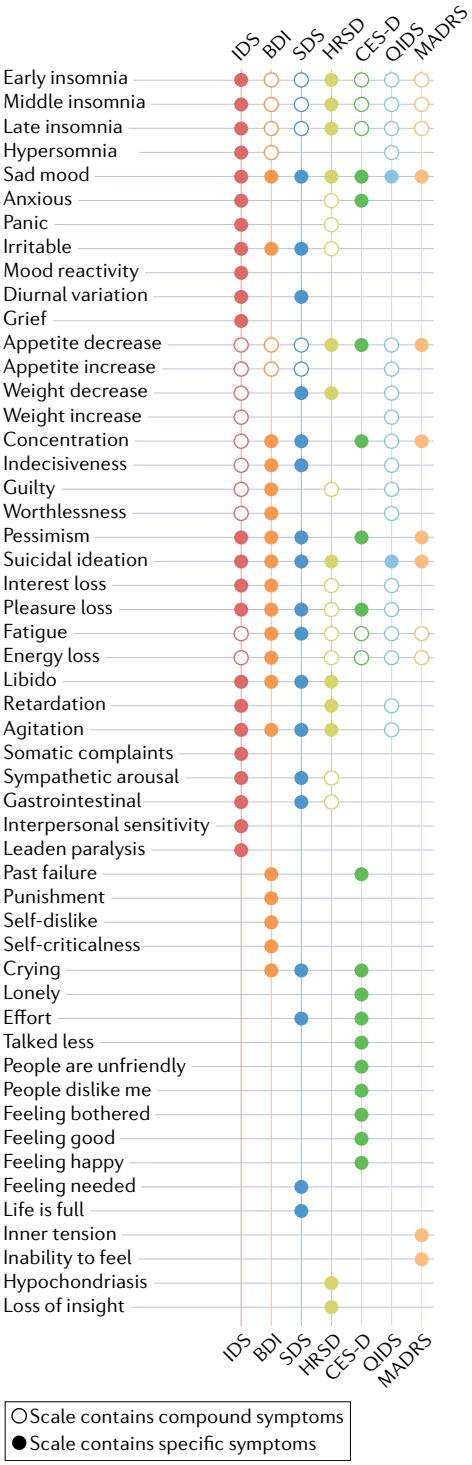

Fig. 1 | **Co-occurrence of 52 depression symptoms across 7 depression rating scales.** Coloured circles for a symptom indicate that a scale directly assesses that symptom, whereas empty circles indicate that a scale indirectly measures a symptom. For instance, IDS assesses 'hypersomnia' directly; BDI measures 'hypersomnia' indirectly via a general question on sleep problems; and SDS does not capture 'hypersomnia' at all. Note that the nine QIDS items analysed correspond to DSM-5 criterion symptoms for MDD. BDI, Beck Depression Inventory[119]; CES-D, Centre of Epidemiological Studies Depression Scale[5]; DSM-5, *Diagnostic and Statistical Manual of Mental Disorders* fifth edition; HRSD, Hamilton Rating Scale for Depression[3]; IDS, Inventory of Depressive Symptoms[80]; MADRS, Montgomery–Åsberg Depression Rating Scale[35]; QIDS, Quick Inventory of Depressive Symptoms[81]; SDS, Zung Self-Rating Depression Scale[145]. Figure adapted with permission from REF. [18], Elsevier.

such as accurately determining whether treatment is needed or progressing well.

The development of diagnostic criteria sets and scales in the middle of the twentieth century provided substantial clarity about the content being assessed relative to earlier, unstructured, interviews. Accordingly, these instruments supported clearer communication and provided standard criteria for determining the need for treatment. However, there is a surprising level of disagreement about the content that depression measures ought to assess. A review of 7 commonly used scales for depression[18,27], including the CES-D, BDI and HRSD, found that they contain 52 disparate symptoms, 40% of which appear in only 1 of the scales. The CES-D — the most used depression scale in history (see Supplementary Fig. 1) — has the lowest mean overlap with other scales (Jaccard similarity index of 31%), with half of all CES-D items not appearing in any of the six other scales. Content overlap between common scales and the DSM-5 criteria for MDD is only moderate (FIG. 1).

If the 52 distinct symptoms are considered to reflect the full breadth of the depression construct for diagnosing, monitoring and studying depression, no single scale can be said to have adequate content converage[25]. Also, the 7 scales mentioned above are just a small minority: more than 280 instruments have been developed to assess depression[1], and a recent review of 30 clinical trials in adolescents found 19 different primary outcome measures for depression[28]. Additionally, there is evidence that none of these scales assesses important features of depression. In a recent study of more than 3,000 patients, informal caregivers and healthcare professionals from 52 countries, mental pain was commonly mentioned as an important feature of depression[29]; but common depression scales do not include this experience[18].

Scales therefore seem to measure different 'depressions'. This conclusion is supported by their different content and the fact that correlations among scales are often only around 0.5 (and regularly much lower)[30–33]. Small-to-moderate correlations among scales are not surprising, as depression instruments were constructed absent a unifying theory and by scholars working in distinct settings and towards distinct goals. The HRSD was developed for inpatients with severe depression who had already been diagnosed and relies heavily on clinical, observable signs such as weight loss and slowing of speech rather than self-reported symptoms. The BDI focuses on cognitive and affective symptoms, such as worthlessness and pessimism, which are central to Beck's theory of depression[34]. The CES-D was developed for depression screening in general population settings and captures problems such as feeling bothered or lonely that are more common in non-clinical settings than BDI or HRSD symptoms. Items on the Montgomery–Åsberg Depression Rating Scale (MADRS)[35] were selected because they were found to change during treatment, providing a scale sensitive to change[35].

Despite these key differences in content, scales such as the BDI, HRSD and MADRS are used interchangeably to, for example, track treatment progress in clinical trials. Clinical trials usually report how many patients respond to and remit during treatment. However, there are systematic differences in the measurement of pharmacological interventions (mostly observer-rated; the HRSD and MADRS most common) versus behavioural interventions (mostly self-rated; the Patient Health Questionnaire (PHQ-9) most common)[36]. Comparing treatments based on different measures is problematic owing to content differences and because observer-rated scales result in larger pre–post treatment effect sizes than self-report scales[36]. Different treatments are therefore confounded with different types of measurement, biasing their comparison.

Another problem of interchangeable use is when scales are used to diagnose participants. The PHQ-9 developers, for instance, encourage doing so[37], despite evidence that scales such as the PHQ-9 produce substantially higher rates of MDD prevalence than clinical interviews[36,38]. Important decisions about whether people are diagnosed with MDD, enrolled into a clinical trial or considered remitted after treatment depend to a considerable degree on the instruments used by researchers and clinicians. This state of affairs leaves much to be desired.

Response processes

A second important source of validity evidence is the response processes involved when people complete a measure. A score is

valid when people respond to an instrument in a way that corresponds to the construct being assessed[39]. For example, when developing a test of mathematical reasoning, questions should not be so rote as to render the test a measure of one's memory for facts, or so verbose as to render the test a measure of reading comprehension.

There is very little research on the processes engaged when people score self-rated or observer-rated depression instruments, but there is reason to think they influence depression measurement. Generally, scores on self-reported depression scales tend to differ from those on observer-rated instruments: studies based on observer ratings find greater efficacy of depression treatments than studies relying on self-report[31,36], and self-reported symptoms tend to be more severe than observer ratings[38,40]. These differences may be due to different response processes. Clinicians may not score some symptoms endorsed in self-report scales if these symptoms can be attributed to external causes. For example, a single parent getting little sleep due to a newborn may endorse sleep problems in a self-report scale, but a clinician may not score that as a depression symptom in a diagnostic interview, leading to differences in scores. Alternatively, it might be that participants report more honestly in self-rated instruments and are less candid in clinical interviews[41,42]. Or differences might arise because observers are vulnerable to certain kinds of cognitive bias, including overconfidence bias (overestimating one's knowledge and therefore acting without sufficient information), confirmation bias (selectively engaging with confirming rather than refuting evidence), attribution error (when a medical condition is misdiagnosed as a psychiatric condition) or diagnosis momentum (even if a diagnosis is erroneously attached to a patient, it tends to stick)[43].

Although the DSM-5 is explicitly atheoretical, clinical judgements are made within the context of implicit and explicit conceptual frameworks, which influence response and measurement processes[44]. A 2007 survey showed that clinicians are acutely aware that these frameworks contribute to clinical decision-making: 86% of participating clinicians stated that psychiatric diagnoses are unreliable. When asked about the reasons for this lack of reliability, clinician-related factors such as differences in training, biases and interview style were the most common explanations for discrepancies between raters (63.5%), rather than patient characteristics (21.6%) or nosological issues (14.9%)[45].

Overall, we know very little about how participants or observers interpret items or select responses on depression instruments. Differences in response processes could explain consistent differences between scale types (such as self-report versus observer report), but this is an area that urgently needs more research.

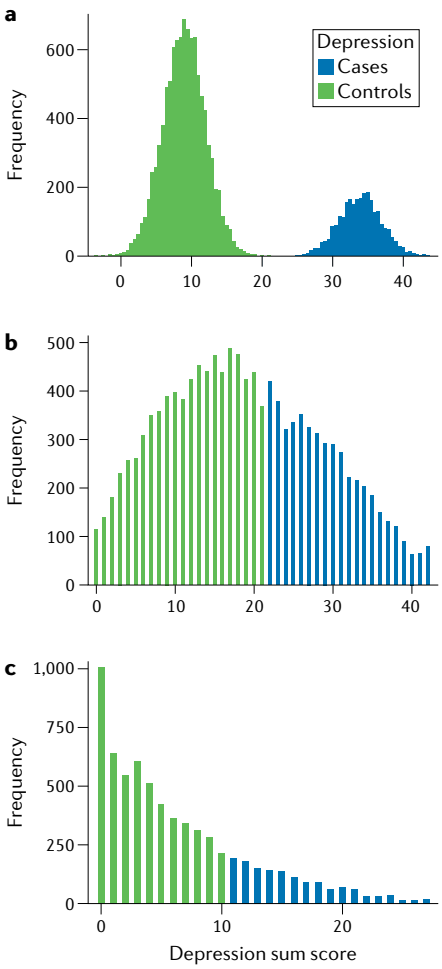

Fig. 2 | **Expected and empirical distributions of depressive symptoms. a** | Expected distribution if a disorder is categorical, featuring a 'zone of rarity' between cases and controls. Such distributions are not observed in empirical depression data. **b** | Empirical distribution of depressive symptoms in an online sample, $n = 12{,}613$, based on 14-item depression subscale of Depression Anxiety Stress Scale (DASS-42)[146]; owing to the nature of an online convenience dataset, depression scores are considerably higher than those in representative samples. **c** | Empirical distribution of depressive symptoms from an individual participant data meta-analysis consisting of 54 datasets with $n = 12{,}613$, based on Patient Health Questionnaire (PHQ-9)[37]. See Supplementary Note 2 for further details and for a link to code and data for reproducing the figure.

Internal structure

The third source of validity evidence comes from an instrument's internal structure: the extent to which the relationships among test items are consistent with one's theory about the construct being assessed[39]. Unfortunately, efforts to evaluate whether the relationships among depression symptoms are consistent with theoretical expectations face an immediate challenge: such expectations are often unclear. The measurement of depression has developed largely independently from theories. The DSM, which is explicitly atheoretical, is one of many examples.

However, many tacit theoretical assumptions about the nature of depression are evident in clinical and research practices. Accordingly, whether an instrument's internal structure conforms to expectations can be evaluated on the basis of assumptions that underlie how this instrument is used. We discuss assumptions of taxonicity, dimensionality, measurement invariance and inter-rater reliability.

Taxonicity. Researchers and clinicians use instruments to separate people with MDD from people without MDD, implying a belief that depression is taxonic (that is, categorical) in nature. In the DSM-5, MDD is diagnosed if participants meet five of nine symptom criteria, for 2 weeks, along with significant impairment of functioning[46]. Similarly, on the HRSD, BDI and CES-D, researchers commonly sum all items and use thresholds to determine the presence of MDD[36]. However, there is considerable evidence that depression is not categorical but, rather, exists on a continuum from healthy to severely depressed[47–50]. If MDD was categorical, one would expect an area of the distribution of depression severity with relatively few individuals — a zone of rarity — that divides those with and without MDD (FIG. 2a). However, such a zone of rarity is not present in observed data (FIG. 2b,c).

The data shown in FIG. 2 explain why many different MDD thresholds exist for each of the most commonly used scales. It is much more difficult to clearly demarcate healthy from sick in the real data presented in FIG. 2b,c, compared with the common conceptualization of depression as a taxon in FIG. 2a that is not supported by evidence. An analysis of 350 clinical trials for depression using the MADRS identified that 28 different thresholds were used to determine whether patients have MDD and should therefore be included in a given trial; thresholds ranged from total scores of 5 to 34 points[36]. Similarly, a review of

29 trials for adolescent MDD identified 47 unique definitions of remission, with only a minority of trials providing a rationale for their cut-offs[51]. Overall, this illustrates that the internal structure of the most commonly used depression instruments does not support the use of cut-offs to identify the presence or absence of MDD.

Dimensionality. Another common use for depression scales is adding up equally weighted symptoms to derive a single score that represents depression severity. This practice is valid if all items measure one construct (rather than three or five) and if items are interchangeable — that is, if they contribute roughly equally to depression severity[52,53]. If these assumptions are met, data from depression instruments will be best described by a unidimensional factor model in which item loadings are roughly equal. However, statistical analyses often do not produce this result. Many depression scales are multidimensional, measuring more than one construct. In fact, one to seven factors have been extracted for the BDI, HRSD and CES-D[33,54,55]. Notably, replication of factor structures is poor across instruments[56], poor within instruments across samples[33,54–61] and poor even within the same instrument in different subsets of the same sample[62]. Thus, not only do depression sum scores often fail to measure a single construct, but the number and nature of those constructs shift across context, time and people. The assumption that items are interchangeable is similarly untenable given a broad set of empirical studies showing that individual symptoms differ in their relations to risk factors[63,64], impact on functioning[65–67] and biological markers[68–74], and are differentially predicted by life events[75–79]. Together, these findings severely limit the use of sum scores to denote one underlying construct[52].

Further evidence against the validity of sum scores comes from the internal consistency of depression instruments: the extent to which people respond similarly across all instrument items. For depression scales, there is a wide range of internal consistency estimates, with reported alpha coefficients from as low as 0.4 to as high as 0.9 (REFS[31,33,80,81]). Acceptable internal consistency (alpha > 0.7) is usually observed in general population samples, whereas alpha is often substantially lower in clinical populations[31,33,80,81]. This phenomenon is particularly visible in clinical trials for depression where, using the same scale in the same sample, alpha often increases considerably within a few weeks as the sample gets healthier (for the HRSD, often from 0.4 to 0.8)[80,81].

Issues of multidimensionality and inadequate internal consistency might be related to the heterogeneity of the MDD phenotype. An analysis of depression symptoms of 3,703 patients with MDD identified more than 1,000 unique empirical symptom profiles; around half of these profiles were endorsed by only a single individual[82] (see REF.[83] for similar findings). There have been attempts to tackle the massive heterogeneity of MDD by proposing more homogeneous depression subtypes or specifiers, such as melancholic or atypical depression, that come with specific symptoms[84–86]. However, subtyping efforts have largely failed to result in categories that support clear demarcation of patients, higher treatment specificity or higher temporal stability[84–96]. Seasonal affective disorder is a possible exception[97–99].

Some of the limits of total scores were understood a long time ago. Hamilton referred to the sum of symptoms assessed by the HRSD as the 'total crude score' and focused his analysis on four subscales assessing narrower phenotypes, such as 'anxiety' and 'agitated depression'.[3] Yet, today, the total crude score is used in nearly all studies that use the HRSD. As reviewed here, the six decades since Hamilton developed his scale have provided ample psychometric evidence regarding the dimensionality and internal consistency of depression instruments, raising questions about the common practice of adding depression symptoms to a single score.

Measurement invariance. Researchers use depression scales to compare scores for different groups of people, implying a belief that depression is invariant across contexts. Accordingly, the measurement of depression should be similarly invariant. Measurement invariance is necessary for common research questions, such as whether depression rates are similar in women and men. If an instrument does not measure the same construct in two groups, it cannot be used to compare groups regarding this construct. Measurement invariance across groups has several levels, including invariance at the structural level (the same number of factors can be extracted in different groups) and invariance of factor loadings (factor loadings of items in one group are similar to those in another group). The more the psychometric properties of an instrument (such as factor loadings) remain consistent across groups, the more the instrument can be said to exhibit measurement invariance[100], that is, to measure the same construct across groups.

Some level of measurement invariance has been established for certain depression instruments across certain groups; for example, the nine DSM-IV depression symptoms exhibit measurement invariance in women assessed across samples collected in the United States, Europe and China[101], as do PHQ-9 scores across women and men in a community sample in Hong Kong[102]. However, in other situations and data, depression instruments do not meet the level of measurement invariance required to compare groups on depression scores. Significant differences in the psychometric properties of depression instruments have been observed across groups defined by socio-economic status[103], ethnicity[104], sex[105] and age[106], among others. Thus, common uses of depression instruments, such as comparing depression scores across groups, might not be valid, depending on the instrument and situation.

The same issue applies to measurement invariance across time: some studies found that so-called temporal measurement invariance of depression scales held, which means that a score in a sample at time 1 holds the same meaning as a score in the same sample at time 2 because the same construct is measured[107]. But many other studies have demonstrated a lack of temporal invariance[33,58,108]. This raises serious concerns about using depression scales to track treatment progress. If temporal measurement invariance is violated, a BDI score of 20 points for a sample at treatment entry and 10 points for the same sample 8 weeks later does not measure the same construct, limiting the ability to assess treatment efficacy[33].

Inter-rater reliability. Finally, diagnoses in clinical or research settings are usually given by one rater, implying a belief that diagnoses are sufficiently reliable that multiple assessors are unnecessary. Indeed, much of the motivation to move towards diagnostic criteria and standardized scales was to enhance reliability: the consistency of scores obtained in depression instruments across raters, contexts and time[25]. Here, we focus on inter-rater reliability (the extent to which independent raters produce similar scores), which is required to support the common clinical and research practice of using one rater to assign a depression diagnosis. Inter-rater reliability is important because prevalence rates derived from diagnoses inform mental health policy, and because both overdiagnosis and under-diagnosis of

MDD can have dramatic consequences for a person's life. There are three broad sets of findings related to inter-rater reliability.

Some studies for DSM diagnoses and observer-rated scales have noted very high agreement among raters, at times exceeding 0.90 (REFS[109,110]). However, such high agreement is usually obtained when interviews are not conducted independently (for example, both raters watch the same interview tape), which inflates agreement among raters. To properly assess inter-rater reliability, different clinicians must conduct their interviews separately.

Studies using separate, structured clinical interviews show moderate agreement between raters. For example, a study in which different clinicians conducted structured interviews reported a kappa coefficient of 0.62 (REF.[111]). Although MDD had the lowest inter-rater reliability of the 20 assessed diagnoses, the result suggests that such interviews can produce substantial agreement. Unfortunately, only an estimated 15% of clinical psychologists and psychiatrists make use of structured interviews[112].

Finally, studies examining how diagnostic criteria perform in routine clinical practice paint a troubling picture; the DSM-5 field trials are a prominent example. Such field trials are conducted when new versions of official psychiatric nosologies such as the International Classification of Diseases (ICD) or DSM are released, with the goal of assessing the reliability of psychiatric diagnoses in clinical practice. In the DSM-5 field trials, interviewers had a minimum of 2 years of psychiatric postgraduate training, and for each participant, independent psychiatric assessments were conducted by two interviewers within 4–48 h of each other; interviews relied on usual clinical interview procedures to "mirror the circumstances in which most diagnosing takes place"[113]. Strikingly, despite using criteria designed explicitly to promote reliability[114], inter-rater reliability for a diagnosis of MDD was just 0.28, placing it among the least reliable diagnoses in the DSM. We illustrate the severe impact of this level of inter-rater reliability on diagnostic outcomes (both false positives and false negatives) in FIG. 3. For comparison, kappa values for bipolar disorder and post-traumatic stress disorder were 0.56 and 0.67, respectively[111]. Reliability for MDD is even lower ($\kappa = 0.16$) when interviews are carried out by general practitioners[115], who are responsible for a substantial proportion of MDD diagnoses worldwide[116].

In summary, the historical shift of DSM-III towards more objective criteria improved reliability, especially when structured interviews are used to assess signs and symptoms. However, the results from studies attempting to approximate typical measurement of MDD in clinical contexts are discouraging, to the degree that the head of the DSM-5 task force had to concede that the "relatively low reliability of major depressive disorder [...] is a concern for clinical decision-making"[113].

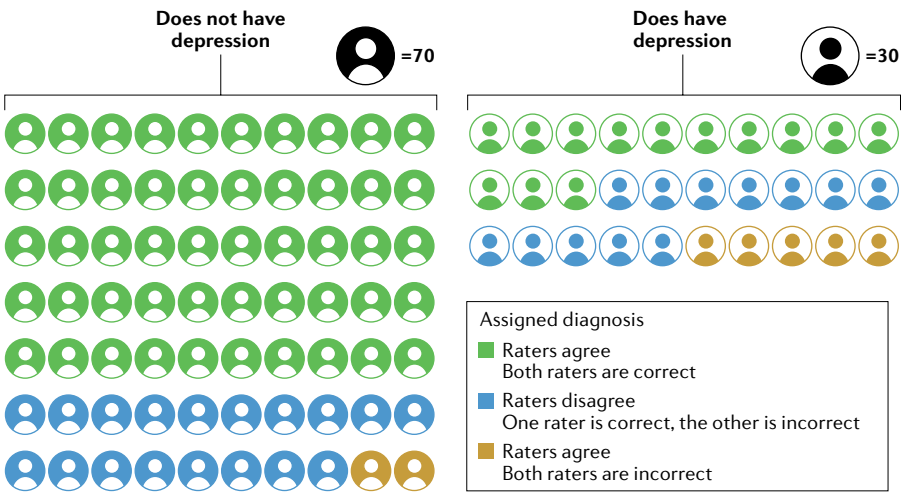

Fig. 3 | **Simulated impact of low inter-rater reliability on diagnostic outcomes.** On the basis of the inter-rater reliability (kappa coefficient) of 0.28 for major depressive disorder (MDD) reported in *Diagnostic and Statistical Manual of Mental Disorders* fifth edition (DSM-5) field trials, we simulated data for 2 clinicians and 100 people, 30 of whom have depression. The left column shows accurate agreement (green) between clinicians in 71% of 70 non-depressed cases, 26% disagreement (blue) and 3% inaccurate agreement (orange). For 30 depressed participants (right column), diagnostic performance was particularly poor, with only 43% accurate agreement (green), 40% disagreement (blue) and 17% inaccurate agreement (orange). See Supplementary Note 3 for a complete description of the simulation underlying this figure.

Shaky foundations

By the standards commonly applied in psychological research, the evidence does not support many of the common uses of depression instruments. Next, we discuss two potential explanations of these shortcomings rooted in the foundations of depression measurement.

Methodological foundations

One explanation for the validity and reliability problems of common depression instruments is that they were not developed following modern best practices. Today, the development and validation of psychometric instruments is a thorough process that occurs in three phases[117]. In phase one, the substance of the construct is explored (for example, clarifying its nature, breadth and depth). In phase two, the structure of the instrument is investigated (for example, using item analyses and factor analysis). Finally, in phase three, the relation between the instrument and external constructs is researched (for example, by testing its ability to discriminate between groups known to be distinct). Critically, each phase includes iterative revisions until the instrument meets the desired criteria[118]. These practices were not widely established when common depression instruments were developed. Although much individual validation research has been published, for example regarding the factor structure of depression instruments[33,54,55,107], it remains a cause for concern that the majority of studies focused on phases two and three rather than the foundational phase one, and that findings have not led to substantial iterative development of depression measurement (the BDI is one of the few scales that have been updated over the years[119]).

Rather than a rigorous exploration of the construct to be measured via item development, expert reviews or focus groups, scale developers often developed depression instruments on the basis of their clinical experiences and personal views. For example, Hamilton developed the HRSD on the basis of his experiences and knowledge working with inpatients, not from an explicated theory that defined depression[3]. Other scale developers sidestepped theoretical considerations altogether, often via statistical procedures. For example, to obtain an "empirically founded scale,"[35] the MADRS items were chosen by dropping items from a prior scale on which 64 patients

did not improve significantly after receiving antidepressant drugs. This procedure raises questions about what the scale measures. As a recent historical analysis notes, the MADRS is really a measure of emotions that change over a few weeks in drug trials, rather than a measure of depression[120].

Furthermore, the content of depression instruments was often shaped not by theory or to support valid scores but by what Lilienfeld called "centrifugal antiscientific forces,"[121] including practical constraints such as ease of clinical use as well as continuity (for example, with previous DSM editions). The former emphasizes brief instruments, and thereby presents a considerable obstacle towards adequate content representation, whereas the latter severely constrains opportunities for iterative development.

Overall, given these various shortcomings and constraints, it should not be surprising that validity evidence does not support many of the common uses of depression measurement.

Theoretical foundations

A second, related explanation for problems with validity and reliability of common depression instruments is that most depression instruments were developed without clear and explicit theories about the nature of depression. Without a clear theory of depression, it is unclear what researchers aim to measure, and how we can evaluate whether they have succeeded[122–124]. However, the absence of explicit theories does not mean depression measurement is atheoretical: many implicit beliefs about depression, such as taxonicity, are embedded in measures. Unfortunately, these latent theories[125] do not align well with decades of empirical research, and one of the most fundamental latent theories is the notion that depression symptoms arise from a common cause.

Mental health research leans heavily on medicine and psychology in its approach to measurement. Medicine has symptom checklists that indicate diseases, and psychologists measure constructs such as mathematical ability or personality via tests and questionnaires. Critically, using items to indicate underlying diseases or constructs works only under the assumption that constructs cause item responses. Medical symptoms can indicate an underlying disorder because symptoms are caused by the disorder (for example, measles causes Koplik's spots). Similarly, tests for mathematical ability use items such as '17 × 39' because an individual's performance on this item is thought to be caused by their mathematical ability.

The common cause theory — all symptoms have a shared origin — implicitly underlies the scoring and use of nearly all psychiatric assessments, including depression instruments, and justifies practices such as summing items[126,127]. The field uses the term 'symptom' to refer to items of depression instruments, implying (by definition of the word symptom) an independent disease entity that gives rise to the symptoms[128]. The common cause theory presupposes a homogeneous, categorical, unitary construct with interchangeable symptoms. The common cause theory also explains common statistical practices in psychiatric research, including the notion that 'good' scales should be unidimensional (they measure one construct) and have high internal consistency (items measure the same construct)[125].

In contrast to the common cause framework, there is increasingly widespread recognition that depression is a highly heterogeneous, multifactorial and complex phenotype[129,130]. Depression has fuzzy boundaries, and features both multi-finality (the same constellation of variables can lead to different outcomes) and equifinality (different constellations can lead to the same outcome)[131]. Depression also shows pronounced inter-individual differences, such that two people diagnosed with MDD may not share a single symptom[82,83], and the disorder may be categorical for some but continuous for others[132]. Indeed, the lack of validity evidence for commonly used instruments for measuring depression can be interpreted as evidence that there is a mismatch between the nature of depression and the common cause theory implicit in these instruments. Importantly, if the common cause theory had been explicit when these instruments were developed, the failure to observe evidence for validity would have immediately signalled that something was wrong, either with the instrument or with the theory of depression. However, because theories about depression have largely remained latent and are only implied through research and clinical practices, these discrepancies were less salient, and opportunities to improve depression instruments were missed.

Improving depression measurement

The state of depression measurement today resembles that of thermometry in the seventeenth century. Although objective measures of temperature are now taken for granted, just a few centuries ago there were many different thermometers developed by many different scientists, all claiming to measure temperature, with "standards kept by each workman, without any agreement or reference to one another" (Halley, 1693, referenced elsewhere[133]). Everyone could agree that these thermometers were assessing something, but the precise nature of the thing was unclear.

Progress in thermometry was made possible by epistemic iteration, a series of successive approximations in which advances in thermometry afforded advances in understanding temperature that, in turn, allowed further improvements in thermometry[133]. Central to this framework is the notion that fallible measures, despite their imperfections, can provide enough advance in knowledge that there is an opportunity for further advances in measurement.

Despite the obvious differences between depression and temperature, the idea of an ongoing exchange between advances in knowledge and improvements in measurement provides a crucial framework for considering how the measurement of depression can move forward. We provide a list of concrete suggestions for improving depression measurement (BOX 2), based on two fundamental principles.

First, we cannot divorce our measures of depression from our theories about what depression is. In contrast to current practices, where measures are often expressly atheoretical but infused with implicit theories (such as that MDD is categorical), it will be essential to ground measurement in strong theories that explicate core assumptions about the nature of depression. Grounding measurement in clearly explicated theories will enable researchers to identify the limits of existing measures and take steps to improve them[123]. Developing such theories will be challenging given the complex, dynamic and heterogeneous nature of MDD. But doing so is crucial, owing to the central role of theory development in advancing scientific knowledge. To this end, clinical sciences can draw on tools and frameworks from fields with rigorous approaches to modelling processes of interest[123,125,134].

Second, improving depression measurement requires iterative development. Despite evidence of the shortcomings of common instruments that have been in use for many decades, there has been minimal effort to move beyond these measures. Evidence of shortcomings is not a criticism of original scale developers; we doubt that Hamilton would have wanted his scale used uncritically and without any

Box 2 | **Towards better depression measurement**

We suggest several steps for iterative improvement of depression measurement.

Development and iteration

Develop explicit theories of depression. Without a clear theory, it is unclear what we ought to measure, and how to evaluate whether we have succeeded in doing so. Explicit theories spell out core beliefs or assumptions about the nature of depression in detail and, in the best case, do so in formalized ways[123,125,148].

Epistemic iteration. Progress in depression measurement comes from successive approximations in which each stage moves us closer to our epistemic goals[133]. Fallible depression instruments, such as the Hamilton Rating Scale for Depression (HRSD) or Beck Depression Inventory (BDI), can provide advances in knowledge which, in turn, enable advances in measurement. This iterative exchange between theory and measurement provides an avenue for science to progress, but critically relies on having explicated theories in the first place.

Experience experts and cross-cultural aspects. Common depression instruments were predominantly designed by WEIRD (western, educated, industrialized, rich and democratic) clinicians, and validated in WEIRD samples. It is crucial to involve people with lived experiences and their caregivers, and people from non-WEIRD cultures and countries, in this process[29].

Response processes. There is a lack of research on how self-rated and observer-rated scales are scored. Response processes should be investigated when developing new (or improving existing) instruments via tools such as the response process evaluation method, a type of cognitive interview that elucidates how participants interpret items and select responses[149–153].

Use

Use scales for appropriate purposes. Not all instruments are appropriate for all purposes[154]. Hamilton wrote in 1960 that his scale ought to be used only in patients already diagnosed as a measure of severity[3], but the HRSD is commonly used today to distinguish participants with MDD from healthy participants. Researchers and clinicians should use instruments for the purposes for which they were developed and validated, and justify their choice of depression instrument. In the immediate future, this might mean developing and using different instruments for different uses, for example, one for determining whether treatment is warranted and another for tracking progress. We note that this suggestion is opposed to recent initiatives by the National Institutes of Health (NIH) and the Wellcome Trust to mandate the Patient Health Questionnaire (PHQ-9) as a universal depression measure for all contexts and uses[154].

Robustness. Especially for data-driven research, researchers should consider using multiple depression instruments and investigating whether they lead to robust results, or whether results depend on the use of one particular instrument[52,154].

Symptomics. Depression severity and MDD are highly heterogeneous phenotypes, such that it can be unclear what scores on these phenotypes represent. Individual symptoms on depression scales, such as insomnia or suicidal ideation, might represent more valid and reliable phenotypes than symptom sum scores or categorical diagnoses[52,127,129].

Continuous analyses. Consistent with psychometric evidence that depression data are best described as continuous[47–50] (FIG. 2), researchers should avoid arbitrary cut-offs whenever not strictly necessary, and conceptualize and analyse depression as a continuum rather than a taxon.

Reporting

Increase transparency of measure use. The 12 versions of the HRSD differ in the number of items (from 6 to 36)[155,156], and some have dozens of translations. Although these versions differ in crucial aspects such as content and psychometric properties, approximately half of the studies using the HRSD provide no information about the version used[155]. In intervention trials published in clinical psychology journals, only one in seven studies preregister their measures[157], leaving these studies vulnerable to selecting which measures to report post hoc, a practice especially prevalent in studies with industry funding[158]. Similarly, only 18 of 32 reviewed randomized controlled trials of adolescent depression featured an identifiable, single, primary outcome[159]. This lack of transparency when administering instruments creates fertile ground for questionable measurement practices, and muddies the inferences that can be drawn. We recommend answering the six questions to promote transparent reporting of measurement listed in REF.[122].

revisions for more than 60 years. Moreover, reluctance to move beyond these measures is not unfounded. There is clear value in having consistency of measurement across time and contexts for an applied field such as psychiatry. Nonetheless, whatever advantages there are to be gained by adherence to precedent, these are outweighed by the gains to be made by genuine progress in our ability to measure and, therefore, understand, diagnose, prevent and treat depression. Given the shortcomings reviewed here, we should develop better depression measures, but these must be rooted in what we have learned from existing instruments[133].

To illustrate these core principles, consider the theory that depression syndrome emerges from a complex system of causal interactions among the physiological, cognitive, emotional and behavioural experiences we commonly refer to as symptoms[135–138]. This theory emerged because data gathered from existing measures were inconsistent with the common cause theory on which they are implicitly based. Now, to measure depression as a complex system we need new measures, which will require at least two innovations.

First, each component in the system must be measured rigorously, which is not the case in current depression measures, which typically provide only very rough assessments of individual elements[52]. Common depression symptoms such as guilt, suicidal ideation and sleep problems are themselves complex phenotypes, but are usually assessed with only a single item each. In addition, measures could encompass a broader set of elements than symptoms alone, including variables conceptualized as risk factors, maintenance factors and outcomes, such as stress, adversity, impairment and quality of life[139–142]. It will be critical to engage scientific and experiential experts in characterizing the system of elements that drive depression[29,143].

Second, according to this systems theory, symptoms (such as sad mood) do not merely indicate depression; they are active causal agents that influence other symptoms (such as sleep, concentration or suicidal ideation). Thus, individual system elements and their relationships must be measured, necessitating a move away from static, retrospective assessments and towards instruments that can assess the dynamic unfolding of depression within individuals over time. For example, smartphone apps and other digital tools that utilize ecological momentary assessment to query people multiple times per day regarding their thoughts, feelings, behaviours and experiences[144] have the potential to reveal dynamic information about depression, including the development of individual system elements and relationships among them.

Importantly, establishing the components of a system and their relations can promote new insights into depression. From a systems

perspective, someone is at risk for depression if the system, once sufficiently perturbed, is likely to fall into a self-sustaining depressed state. The key to this determination is in the system's attractor states. An attractor state can be thought of as a valley in a landscape, with a ball representing the system's current state resting on the surface. If an individual is healthy, the landscape is flat and only has a single valley, which is their healthy attractor state where elements of depression are absent. Perturbations (such as life stress) may push the ball up the slope of this valley, but will always return to this healthy attractor state. By contrast, if an individual is at risk for depression, the landscape features a second valley in which many elements of depression are active. In this landscape, a perturbation can push the ball up the slope, out of the healthy valley and into the depressed one. Critically, system elements and their causal relationships determine the shape of a person's stability landscape (and, thus, the presence of attractor states). Accordingly, accurately measuring elements and relationships can identify the presence of a harmful attractor, providing a novel measure of depression. From this perspective, depression is determined by the presence of a harmful attractor state, as well as the shape of the stability landscape (for example, how steep the valleys are), rather than just the number of symptoms. This shift has substantial implications for how we think about measuring depression risk, depression severity and depression recovery. Identifying people vulnerable to depression means measuring the system thoroughly to determine whether a depressed attractor is present before the person ever falls into it; measuring depression severity means to assess the shape of the stability landscape in detail; and assessing treatment efficacy might involve measuring the flattening or elimination of the harmful attractor, changing the stability landscape into one that has a single healthy valley. Working from a theory that clearly specifies the nature of the phenomenon we are assessing affords clear new paths for how to measure it.

Our example is not meant to show that the systems approach is the one right theoretical path forward but, rather, that grounding measurement in theories can provide insight into how to advance depression measurement. Measuring depression from a systems perspective would initially exhibit substantial shortcomings, but advances in the theory would enable improvements in measurement, which, in turn, may equip us to interrogate further and advance the theory. Through this iterative exchange we can improve the measurement of depression[133], and in doing so, improve our ability to study, diagnose, treat and prevent it.

Data availability

Data underlying FIGS 1,2 and 3 can be found at https://osf.io/7dp5s/.

Code availability

Code to reproduce FIGS 1 and 2 (minus graphical edits performed by the journal art editor), and run the simulation underlying FIG. 3, can be found at https://osf.io/7dp5s/.

Eiko I. Fried[1]✉, *Jessica K. Flake*[2] *and Donald J. Robinaugh*[3,4]

[1]*Department of Clinical Psychology, Leiden University, Leiden, The Netherlands.*

[2]*Department of Psychology, McGill University, Montreal, Québec, Canada.*

[3]*Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA.*

[4]*Department of Applied Psychology, Northeastern University, Boston, MA, USA.*

✉*e-mail: eikofried@gmail.com*

https://doi.org/10.1038/s44159-022-00050-2
Published online: 14 April 2022

1. Santor, D. A., Gregus, M. & Welch, A. Eight decades of measurement in depression. *Measurement* **4**, 135–155 (2006).
2. van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* **514**, 550–553 (2014).
3. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **23**, 56–62 (1960).
4. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. An inventory for measuring depression. *Arch. Gen. Psychiatry* **4**, 561–571 (1961).
5. Radloff, L. S. The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* **1**, 385–401 (1977).
6. Jorm, A. F., Patten, S. B., Brugha, T. S. & Mojtabi, R. Has increased provision of treatment reduced the prevalence of common mental disorders? Review of the evidence from four countries. *World Psychiatry* **16**, 90–99 (2017).
7. Kapur, S., Phillips, A. G. & Insel, T. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17**, 1174–1179 (2012).
8. Scull, A. American psychiatry in the new millennium: a critical appraisal. *Psychol. Med.* https://doi.org/10.1017/S0033291721001975 (2021).
9. Cuijpers, P. et al. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. *Acta Psychiatr. Scand.* **144**, 288–299 (2021).
10. Khan, A. & Brown, W. A. Antidepressants versus placebo in major depression: an overview. *World Psychiatry* **14**, 294–300 (2015).
11. Kendler, K., Munōz, R. & Murphy, G. The development of the Feighner criteria: a historical perspective. *Am. J. Psychiatry* **167**, 134–142 (2010).
12. Spitzer, R. L. Psychiatric diagnosis: are clinicians still necessary? *Compr. Psychiatry* **24**, 399–411 (1983).
13. Horwitz, A. V. in *The Encyclopedia of Clinical Psychology* (eds Cautin, R. L. & Lilienfeld, S. O.) https://doi.org/10.1002/9781118625392.wbecp012 (Wiley, 2015).
14. Beck, A. Reliability of psychiatric diagnoses: 1. A critique of systematic studies. *Am. J. Psychiatry* **119**, 210–216 (1962).
15. Ash, P. The reliability of psychiatric diagnoses. *J. Abnorm. Soc. Psychol.* **44**, 272–276 (1949).
16. Feighner, J. P. et al. Diagnostic criteria for use in psychiatric research. *Arch. Gen. Psychiatry* **26**, 57–63 (1972).
17. APA. *Diagnostic and Statistical Manual of Mental Disorders* 3rd edn (American Psychiatric Association, 1980).
18. Fried, E. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J. Affect. Disord.* **208**, 191–197 (2017).
19. Cipriani, A. et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* **391**, 1357–1366 (2018).
20. Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302 (1955).
21. Robins, E. & Guze, S. B. Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *Am. J. Psychiatry* **126**, 983–987 (1970).
22. Bandalos, D. L. *Measurement Theory and Applications for the Social Sciences* (Guilford, 2018).
23. Kane, M. T. Validating the interpretations and uses of test scores. *J. Educ. Meas.* **50**, 1–73 (2013).
24. Mokkink, L. B. et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **63**, 737–745 (2010).
25. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for Educational and Psychological Testing* (Joint Committee on Standards for Educational and Psychological Testing, 2014).
26. Messick, S. Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* **18**, 5–11 (1989).
27. Fried, E. I. Corrigendum to "The 52 symptoms of major depression: lack of content overlap among seven common depression scales" [Journal of Affective Disorders, 208, 191–197]. *J. Affect. Disord.* **260**, 744 (2020).
28. Mew, E. J. et al. Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. *J. Clin. Epidemiol.* **126**, 71–79 (2020).
29. Chevance, A. M. et al. Identifying outcomes for depression that matter to patients, informal caregivers and healthcare professionals: qualitative content analysis of a large international online survey. *Lancet Psychiatry* **7**, 692–702 (2020).
30. Wittkampf, K. et al. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen. Hosp. Psychiatry* **31**, 451–459 (2009).
31. Sayer, N. N. A. et al. The relations between observer-rating and self-report of depressive symptomatology. *Psychol. Assess.* **5**, 350–360 (1993).
32. Furukawa, T. A. et al. Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking. *Epidemiol. Psychiatr. Sci.* **29**, E24 (2019).
33. Fried, E. et al. Measuring depression over time … or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychol. Assess.* **28**, 1354–1367 (2016).
34. Beck, A. T., Rush, A. J., Shaw, F. S. & Emery, G. *Cognitive Therapy of Depression* (Guilford, 1979).
35. Montgomery, S. A. & Asberg, M. A new depression scale designed to be sensitive to change. *Br. J. Psychiatry* **134**, 382–389 (1979).
36. von Glischinski, M., von Brachel, R., Thiele, C. & Hirschfeld, G. Not sad enough for a depression trial? A systematic review of depression measures and cut points in clinical trial registrations: systematic review of depression measures and cut points. *J. Affect. Disord.* **292**, 36–44 (2021).
37. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
38. Levis, B. et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. *J. Clin. Epidemiol.* **122**, 115–128.e1 (2020).
39. Whiston, S. *Principles and Applications of Assessment in Counseling* (Brooks/Cole, Cengage Learning, 2009).
40. Thombs, B. D., Kwakkenbos, L., Levis, A. W. & Benedetti, A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *Can. Med. Assoc. J.* **190**, 44–49 (2018).
41. Lavender, J. M. & Anderson, D. A. Effect of perceived anonymity in assessments of eating disordered behaviors and attitudes. *Int. J. Eat. Disord.* **42**, 546–551 (2009).

42. Keel, P. K., Crow, S., Davis, T. L. & Mitchell, J. E. Assessment of eating disorders: comparison of interview and questionnaire data from a long-term follow-up study of bulimia nervosa. *J. Psychosom. Res.* **53**, 1043–1047 (2002).
43. Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Med.* **78**, 775–780 (2003).
44. Kim, N. S. & Ahn, W. Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *J. Exp. Psychol. Gen.* **131**, 451–476 (2002).
45. Aboraya, A. Clinicians' opinions on the reliability of psychiatric diagnoses in clinical settings. *Psychiatry* **4**, 31–33 (2007).
46. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR* (American Psychiatric Association, 2000).
47. Ruscio, J., Zimmerman, M., McGlinchey, J. B., Chelminski, I. & Young, D. Diagnosing major depressive disorder XI: a taxometric investigation of the structure underlying DSM-IV symptoms. *J. Nerv. Ment. Dis.* **195**, 10–19 (2007).
48. Haslam, N. Categorical versus dimensional models of mental disorder: the taxometric evidence. *Aust. N. Z. J. Psychiatry* **37**, 696–704 (2003).
49. Haslam, N., Holland, E. & Kuppens, P. Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychol. Med.* **42**, 903–920 (2012).
50. Nettle, D. in *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory* (eds Adriaens, P. R. & De Block, A.) 192–209 (Oxford Univ. Press, 2011).
51. Courtney, D. B. et al. Forks in the road: definitions of response, remission, recovery and other dichotomized outcomes in randomized controlled trials for adolescent depression. A scoping review. *Depress. Anxiety* **38**, 1152–1168 (2021).
52. Fried, E. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13**, 1–11 (2015).
53. McNeish, D. & Wolf, M. G. Thinking twice about sum scores. *Behav. Res. Methods* **52**, 2287–2305 (2020).
54. Gullion, C. M. & Rush, A. J. Toward a generalizable model of symptoms in major depressive disorder. *Biol. Psychiatry* **44**, 959–972 (1998).
55. Helmes, E. & Nielson, W. R. An examination of the internal structure of the Center for Studies-Depression Scale in two medical samples. *Pers. Individ. Dif.* **25**, 735–743 (1998).
56. Shafer, A. B. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J. Clin. Psychol.* **62**, 123–146 (2006).
57. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* **10**, 156 (2012).
58. Quilty, L. C. et al. The structure of the Montgomery–Åsberg Depression Rating Scale over the course of treatment for depression. *Int. J. Methods Psychiatr. Res.* **22**, 175–184 (2013).
59. Elhai, J. D. et al. The factor structure of major depression symptoms: a test of four competing models using the Patient Health Questionnaire-9. *Psychiatry Res.* **199**, 169–173 (2012).
60. Wardenaar, K. J. et al. The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *J. Affect. Disord.* **125**, 146–154 (2010).
61. Wood, A. M., Taylor, P. J. & Joseph, S. Does the CES-D measure a continuum from depression to happiness? Comparing substantive and artifactual models. *Psychiatry Res.* **177**, 120–123 (2010).
62. Furukawa, T. et al. Cross-cultural equivalence in depression assessment: Japan–Europe–North American study. *Acta Psychiatr. Scand.* **112**, 279–285 (2005).
63. Lux, V. & Kendler, K. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol. Med.* **40**, 1679–1690 (2010).
64. Fried, E., Nesse, R. M., Zivin, K., Guille, C. & Sen, S. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychol. Med.* **44**, 2067–2076 (2014).
65. Faravelli, C., Servi, P., Arends, J. & Strik, W. Number of symptoms, quantification, and qualification of depression. *Compr. Psychiatry* **37**, 307–315 (1996).
66. Tweed, D. L. Depression-related impairment: estimating concurrent and lingering effects. *Psychol. Med.* **23**, 373–386 (1993).
67. Fried, E. & Nesse, R. M. The impact of individual depressive symptoms on impairment of psychosocial functioning. *PLoS ONE* **9**, e90311 (2014).
68. Hasler, G., Drevets, W. C., Manji, H. K. & Charney, D. S. Discovering endophenotypes for major depression. *Neuropsychopharmacology* **29**, 1765–1781 (2004).
69. Myung, W. et al. Genetic association study of individual symptoms in depression. *Psychiatry Res.* **198**, 400–406 (2012).
70. Kendler, K., Aggen, S. H. & Neale, M. C. Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. *Am. J. Psychiatry* **70**, 599–607 (2013).
71. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & Van Der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun.* **9**, 905 (2018).
72. Hilland, E. et al. Exploring the links between specific depression symptoms and brain structure: a network study. *Psychiatry Clin. Neurosci.* **74**, 220–221 (2020).
73. Fried, E. et al. Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. *Psychol. Med.* **50**, 2682–2690 (2020).
74. Eeden, W. A. V. et al. Basal and LPS-stimulated inflammatory markers and the course of individual symptoms of depression. *Transl. Psychiatry* **10**, 235 (2020).
75. Keller, M. C. & Nesse, R. M. Is low mood an adaptation? Evidence for subtypes with symptoms that match precipitants. *J. Affect. Disord.* **86**, 27–35 (2005).
76. Keller, M. C. & Nesse, R. M. The evolutionary significance of depressive symptoms: different adverse situations lead to different depressive symptom patterns. *J. Pers. Soc. Psychol.* **91**, 316–330 (2006).
77. Keller, M. C., Neale, M. C. & Kendler, K. Association of different adverse life events with distinct patterns of depressive symptoms. *Am. J. Psychiatry* **164**, 1521–1529 (2007).
78. Cramer, A. O. J., Borsboom, D., Aggen, S. H. & Kendler, K. The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychol. Med.* **42**, 957–965 (2013).
79. Fried, E. et al. From loss to loneliness: the relationship between bereavement and depressive symptoms. *J. Abnorm. Psychol.* **124**, 256–265 (2015).
80. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26**, 477–486 (1996).
81. Rush, A. J. et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major dDepression. *Biol. Psychiatry* **54**, 573–583 (2003).
82. Fried, E. & Nesse, R. M. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J. Affect. Disord.* **172**, 96–102 (2015).
83. Zimmerman, M., Ellison, W., Young, D., Chelminski, I. & Dalrymple, K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Compr. Psychiatry* **56**, 29–34 (2014).
84. Lichtenberg, P. & Belmaker, R. H. Subtyping major depressive disorder. *Psychother. Psychosom.* **79**, 131–135 (2010).
85. Baumeister, H. & Parker, J. D. Meta-review of depressive subtyping models. *J. Affect. Disord.* **139**, 126–140 (2012).
86. Bech, P. Struggle for subtypes in primary and secondary depression and their mode-specific treatment or healing. *Psychother. Psychosom.* **79**, 331–338 (2010).
87. Lam, R. W. & Stewart, J. N. The validity of atypical depression in DSM-IV. *Compr. Psychiatry* **37**, 375–383 (1996).
88. Davidson, J. R. T. A history of the concept of atypical depression. *J. Clin. Psychiatry* **68**, 10–15 (2007).
89. Arnow, B. A. et al. Depression subtypes in predicting antidepressant response: a report from the iSPOT-D trial. *Am. J. Psychiatry* **172**, 743–750 (2015).
90. Paykel, E. S. Basic concepts of depression. *Dialogues Clin. Neurosci.* **10**, 279–289 (2008).
91. Rush, A. J. The varied clinical presentations of major depressive disorder. *J. Clin. Psychiatry* **68**, 4–10 (2007).
92. Melartin, T. et al. Co-morbidity and stability of melancholic features in DSM-IV major depressive disorder. *Psychol. Med.* **34**, 1443 (2004).
93. Fried, E., Coomans, F. & Lorenzo-luaces, L. The 341 737 ways of qualifying for the melancholic specifier. *Lancet Psychiatry* **7**, 479–480 (2020).
94. Oquendo, M. A. et al. Instability of symptoms in recurrent major depression: a prospective study. *Am. J. Psychiatry* **161**, 255–261 (2004).
95. Coryell, W. et al. Recurrently situational (reactive) depression: a study of course, phenomenology and familial psychopathology. *J. Affect. Disord.* **31**, 203–210 (1994).
96. Pae, C. U., Tharwani, H., Marks, D. M., Masand, P. S. & Patkar, A. A. Atypical depression: a comprehensive review. *CNS Drugs* **23**, 1023–1037 (2009).
97. Magnusson, A. & Boivin, D. Seasonal affective disorder: an overview. *Chronobiol. Int.* **20**, 189–207 (2003).
98. Meyerhoff, J., Young, M. A. & Rohan, K. J. Patterns of depressive symptom remission during the treatment of seasonal affective disorder with cognitive-behavioral therapy or light therapy. *Depress. Anxiety* **35**, 457–467 (2018).
99. Lam, R. W. et al. Efficacy of bright light treatment, fluoxetine, and the combination in patients with nonseasonal major depressive disorder a randomized clinical trial. *JAMA Psychiatry* **73**, 56–63 (2016).
100. Meredith, W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **58**, 525–543 (1993).
101. Kendler, K. et al. The similarity of the structure of DSM-IV criteria for major depression in depressed women from China, the United States and Europe. *Psychol. Med.* **45**, 1945–1954 (2015).
102. Yu, X., Tam, W. W. S., Wong, P. T. K., Lam, T. H. & Stewart, S. M. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Compr. Psychiatry* **53**, 95–102 (2012).
103. Nguyen, H. T., Kitner-Triolo, M., Evans, M. K. & Zonderman, A. B. Factorial invariance of the CES-D in low socioeconomic status African Americans compared with a nationally representative sample. *Psychiatry Res.* **126**, 177–187 (2004).
104. Crockett, L. J., Randall, B. A., Shen, Y.-L., Russell, S. T. & Driscoll, A. K. Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *J. Consult. Clin. Psychol.* **73**, 47–58 (2005).
105. Baas, K. D. et al. Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *J. Affect. Disord.* **129**, 229–235 (2011).
106. Williams, C. D. et al. CES-D four-factor structure is confirmed, but not invariant, in a large cohort of African American women. *Psychiatry Res.* **150**, 173–180 (2007).
107. Stochl, J. et al. On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment* **29**, 355–366 (2022).
108. Fokkema, M., Smits, N., Kelderman, H. & Cuijpers, P. Response shifts in mental health interventions: an illustration of longitudinal measurement invariance. *Psychol. Assess.* **25**, 520–531 (2013).
109. Bagby, R. M., Ryder, A. G., Schuller, D. R. & Marshall, M. B. Reviews and overviews the hamilton depression rating scale: has the gold standard become a lead weight? *Am. J. Psyc* **161**, 2163–2177 (2004).
110. Trajković, G. et al. Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. *Psychiatry Res.* **189**, 1–9 (2011).
111. Regier, D. A. et al. DSM-5 field trials in the United States and Canada, part II: test–retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).
112. Bruchmüller, K., Margraf, J., Suppiger, A. & Schneider, S. Popular or unpopular? Therapists' use of structured interviews and their estimation of patient acceptance. *Behav. Ther.* **42**, 634–643 (2011).
113. Kupfer, D. J. & Kraemer, H. C. Field trial results guide DSM recommendations. *Huffington Post* http://www.huffingtonpost.com/david-j-kupfer-md/dsm-5_b_2083092.html (2013).
114. Clarke, D. E. et al. DSM-5 field trials in the United States and Canada, part I: study design, sampling strategy, implementation, and analytic approaches. *Am. J. Psychiatry* **170**, 43–58 (2012).
115. Fernández, A. et al. Is major depression adequately diagnosed and treated by general practitioners? Results from an epidemiological study. *Gen. Hosp. Psychiatry* **32**, 201–209 (2010).
116. Huxley, P. Mental illness in the community: the Goldberg–Huxley model of the pathway to psychiatric care. *Nord. J. Psychiatry, Suppl.* **50**, 47–53 (1996).
117. Flake, J. K., Pek, J. & Hehman, E. Construct validation in social and personality research: current practice and recommendations. *Soc. Psychol. Personal. Sci.* **8**, 370–378 (2017).

118. de Vet, H. C. W., Terwee, C. B., Mokkink, L. B. & Knol, D. L. *Measurement in Medicine: A Practical Guide* (Cambridge Univ. Press, 2011).
119. Beck, A. T., Steer, R. A., Ball, R. & Ranieri, W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *J. Pers. Assess.* **67**, 588–597 (1996).
120. McPherson, S. & Armstrong, D. Psychometric origins of depression. *Hist. Hum. Sci.* https://doi.org/10.1177/09526951211009085 (2021).
121. Lilienfeld, S. O. DSM-5: centripetal scientific and centrifugal. *Clin. Psychol. Sci. Pract.* **21**, 269–279 (2014).
122. Flake, J. K. & Fried, E. Measurement schmeasurement: questionable measurement practices and how to avoid them. *Adv. Methods Pract. Psychol. Sci.* **3**, 456–465 (2020).
123. Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. & Waldorp, L. J. Invisible hands and fine calipers: a call to use formal theory as a toolkit for theory construction. *Perspect. Psychol. Sci.* **16**, 725–743 (2021).
124. Robinaugh, D. J. et al. Advancing the network theory of mental disorders: a computational model of panic disorder. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/km37w (2019).
125. Fried, E. Lack of theory building and testing impedes progress in the factor and network. *Psychol. Inq.* **31**, 271–288 (2020).
126. Van Bork, R., Wijsen, L. D. & Rhemtulla, M. Toward a causal interpretation of the common factor model. *Disputatio* **9**, 581–601 (2017).
127. Fried, E. Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Front. Psychol.* **6**, 1–11 (2015).
128. Fried, E. & Cramer, A. O. J. Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspect. Psychol. Sci.* **12**, 999–1020 (2017).
129. Fried, E. Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Rev. Neurother.* **17**, 423–425 (2017).
130. Fried, E. & Robinaugh, D. J. Systems all the way down: embracing complexity in mental health research. *BMC Med.* **18**, 1–4 (2020).
131. Cicchetti, D. & Rogosch, F. A. Equifinality and multifinality in developmental psychopathology. *Dev. Psychopathol.* **8**, 597–600 (1996).
132. Borsboom, D. et al. Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychol. Med.* **46**, 1567–1579 (2016).
133. Chang, H. *Inventing Temperature: Measurement and Scientific Progress* (Oxford Univ. Press, 2004).
134. Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. & Haig, B. Theory construction methodology: a practical framework for theory formation in psychology. *Perspect. Psychol. Sci.* **16**, 756–766 (2020).
135. Borsboom, D. A network theory of mental disorders. *World Psychiatry* **16**, 5–13 (2017).
136. Kendler, K., Zachar, P. & Craver, C. What kinds of things are psychiatric disorders? *Psychol. Med.* **41**, 1143–1150 (2011).
137. Olthof, M., Hasselman, F., Maatman, F. O. & Bosman, A. M. T. Complexity theory of psychopathology. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/f68ej (2021).
138. Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R. & Borsboom, D. The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychol. Med.* **50**, 353–366 (2020).
139. Hammen, C. Stress and depression. *Annu. Rev. Clin. Psychol.* **1**, 293–319 (2005).
140. Kendler, K., Karkowski, L. M. & Prescott, C. A. Causal relationship between stressful life events and the onset of major depression. *Am. J. Psychiatry* **156**, 837–841 (1999).
141. Mazure, C. M. Life stressors as risk factors in depression. *Clin. Psychol. Sci. Pract.* **5**, 291–313 (1998).
142. McKnight, P. E. & Kashdan, T. B. The importance of functional impairment to mental health outcomes: a case for reassessing our goals in depression treatment research. *Clin. Psychol. Rev.* **29**, 243–259 (2009).
143. Brouwer, M. E. et al. Psychological theories of depressive relapse and recurrence: a systematic review and meta-analysis of prospective studies. *Clin. Psychol. Rev.* **74**, 101773 (2019).
144. Myin-Germeys, I. & Kuppens, P. *The Open Handbook of Experience Sampling Methodology: A Step-by-step Guide to Designing, Conducting, and Analyzing ESM Studies* (Katholieke Universiteit Leuven, 2021).
145. Zung, W. W. K. A self-rating depression scale. *Arch. Gen. Psychiatry* **12**, 63–70 (1965).
146. Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W. & Swinson, R. P. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychol. Assess.* **10**, 176–181 (1998).
147. Sijtsma, K. On the use, the misuse, and the very limited usefulness of cronbach. *Psychometrika* **74**, 107–120 (2009).
148. Smaldino, P. in *Computational Social Psychology* (eds Vallacher, R. B., Read, S. J. & Nowak, A.) (Taylor & Francis, 2017).
149. Presser, S. et al. Methods for testing and evaluating survey questions. *Public Opin. Q.* **68**, 109–130 (2004).
150. Gordon Wolf, M., Ihm, E., Maul, A. & Taves, A. Survey item validation. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/k27w3 (2019).
151. Hawkes, N. & Brown, G. in *Assessment in Cognitive Therapy* (eds Brown, G. & Clark, D.) 243–267 (Guilford, 2015).
152. Willis, G. B. *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (Sage, 2004).
153. Brown, G., Hawkes, N. & Tata, P. Construct validity and vulnerability to anxiety: a cognitive interviewing study of the revised Anxiety Sensitivity Index. *J. Anxiety Disord.* **23**, 942–949 (2009).
154. Patalay, P. & Fried, E. Editorial Perspective: Prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *J. Child. Psychol. Psychiatry* **8**, 1032–1036 (2021).
155. Neumann, L. *Transparency in Measurement: Reviewing 100 Empirical Papers Using the Hamilton Depression Rating Scale* (Leiden Univ., 2020).
156. Williams, J. B. W. Standardizing the Hamilton Depression Rating Scale: past, present, and future. *Eur. Arch. Psychiatry Clin. Neurosci.* **251**, 6–12 (2001).
157. Cybulski, L., Mayo-Wilson, E., Grant, S., Corporation, R. & Monica, S. Improving transparency and reproducibility through registration: the status of intervention trials published in clinical psychology journals. *J. Consult. Clin. Psychol.* **84**, 753–767 (2016).
158. Ramagopalan, S. V. et al. Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res.* **3**, 77 (2018).
159. Monsour, A. et al. Primary outcome reporting in adolescent depression clinical trials needs standardization. *BMC Med. Res. Methodol.* **20**, 1–15 (2020).

Acknowledgements
The authors thank M.G. Wolf, N. Butcher and Z. Cohen for comments on earlier versions of this manuscript. E.I.F. is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 949059). D.J.R. was supported by funding from the National Institute for Mental Health (K23 MH113805). The content is solely the responsibility of the authors and does not necessarily represent the views of any funding agency.

Author contributions
E.I.F and D.J.R. developed the idea and outline for the manuscript, E.I.F. and D.J.R. conducted background research for the manuscript and all authors contributed to writing and revision of the manuscript.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Psychology thanks Ioana Cristea, Kenneth Kendler and Suneeta Monga for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s44159-022-00050-2.

© Springer Nature America, Inc. 2022