

Psychometric origins of depression

History of the Human Sciences

1–17

© The Author(s) 2021



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09526951211009085

journals.sagepub.com/home/hhs

Susan McPherson 
University of Essex, UK

David Armstrong
King's College London, UK

Abstract

This article examines the historical construction of depression over about a hundred years, employing the social life of methods as an explanatory framework. Specifically, it considers how emerging methodologies in the measurement of psychological constructs contributed to changes in epistemological approaches to mental illness and created the conditions of possibility for major shifts in the construction of depression. While depression was once seen as a feature of psychotic personality, measurement technologies made it possible for it to be reconstructed as changeable and treatable. Different types of scaling techniques (Likert versus dichotomous scales) enabled the separation of depressive personality from reactive depression, paving the way for measuring the severity and intensity of emotions. Techniques to test sensitivity to change provided a means of demonstrating the efficacy of new psychoactive drug treatments. Later, more advanced techniques of precision scaling enabled the management of a new measurement problem, clinician unreliability, associated with the growing number of professionals involved in mental health care. Through statistical management of unreliability, the construct of depression has dramatically reduced over this period from hundreds of questionnaire items to potentially just two. Exploring the history of depression through this lens produces an alternative narrative to those that have emerged as a result of medicalisation and the actions of individuals and pressure groups.

Keywords

depression, personality, psychometrics, sensitivity to change, social life of methods

Corresponding author:

Susan McPherson, University of Essex, School of Health and Social Care, Wivenhoe Park, Colchester, CO4 3SQ, UK.

Email: smcpher@essex.ac.uk

Introduction

Histories of the growing prevalence of (clinical) depression recognise the importance of shifts in definition and more intensive case-finding – in other words, there seems agreement that the depression epidemic has been ‘constructed’ in some way rather than reflecting a significant real increase in mental illness in the population. For most of these accounts, the main drivers of the epidemic are professionalisation and medicalisation (Rapley, Moncrieff, and Dillon, 2011). Psychiatrists have been held responsible for extending their domain – ‘All professions strive to broaden the realm of phenomena subject to their control’ (Horwitz and Wakefield, 2007: 213) – while the drug industry and the profit motive have been identified as the driving force (Healy, 2004; Hirshbein, 2009; Shorter, 2013). For other authors, responsibility lies with ‘a too-powerful medical-industrial complex comprising Big Pharma, insurance companies, testing laboratories, equipment and device makers, hospitals, and doctors’ (Frances, 2013: 71) or ‘researchers, physicians, and patients; advertisers, lobbyists, and public-relations experts; consumer advocates, antidrug crusaders, feminists, and consumers of popular media’ (Herzberg, 2009: 192). Expressed another way, it was ‘the unprecedented number of interest groups that have stakes in considering a wide variety of behaviors as pathological’ (Horwitz, 2020: 218).

All these accounts have in common explanations based on various actors who have extended the reach of the depression diagnosis. As Herzberg (2009: 203) emphasises, this is ‘a story about people as much as about technologies and drugs’, in which politics lies behind apparently neutral scientific discoveries. Yet rather than joining this chorus of blame, this article attempts to make the case that developments in methodologies themselves have had an agentic role in the conceptual shifts that have emerged in the construct of depression. Following the idea that methods have a social life (Savage, 2013), this article sets out to study the psychometric origins of depression over a period of about a hundred years, an aspect that has been ignored or glossed over in existing historical accounts.

By focusing on measurement technology, we try to make visible an aspect of the construction of depression that has been largely invisible, or rather, regarded only through the lens of how successfully or not psychometric measures perform their supposed function. Inverting the lens on measurement and method has been used in sociological studies in related areas. For example, Bowker and Star (1999) have examined the use of categorisation in a wide range of social and economic domains, including classification of diseases, arguing that the extent of visibility or invisibility of categorisation within a process can contribute to the fashioning of political and social order. Similarly, Espeland and Stevens (2008) consider the social act of quantifying constructs in numerical form, noting that ‘numbers, like words, should be regarded as deeds: acts of communication whose meaning and functions cannot be reduced to a narrow instrumentality and which depend deeply on “grammars” and “vocabularies” developed through use’. In quantifying constructs, they argue that constructs can be ‘remade’ and thus direct social behaviour and generate new forms of authority. Like these authors, this article attempts to examine the potential power of neutral-seeming instruments to enable or even generate significant shifts in the conceptualisation of depression.

The very word, *instrument*, suggests something in the laboratory rather than a mundane questionnaire. Accounts of the history of psychiatry sometimes refer to instruments but in the sense of machines: ‘By the end of the 1980s, the MRI had replaced CAT scans as the primary instrument of psychiatric research’ (Lieberman and Ogas, 2015: 151). Questionnaires are mentioned only in passing, if at all. But if psychiatric measurement has a life of its own, if it creates a reality that human actors then respond to, then it represents an important, if ignored, aspect of psychiatric history, and the rise of depression diagnoses in particular. We will revisit some of the secondary literature in our conclusion, but the main analysis is concerned with the ways in which psychometrics changed our view of the world of mental illness. As a result, our focus is largely on primary sources, since secondary sources necessarily start from a different explanatory point (as indicated above).

Our analysis aims to document the emergence of psychometric approaches to clinical depression as found in primary source English-language scientific literature, and to select as exemplars those instruments and their accompanying narratives that were produced and discussed at the time in leading academic journals, rather than narratives of measurement devices written from a retrospective stance. Our sources include influential journals in the fields of psychiatry and psychology, such as the *American Journal of Psychiatry* and the *British Journal of Psychiatry*, understood to represent the most powerful narratives in these disciplines across Britain and North America. Measurement tools such as questionnaires travelled across geographical spaces by means of these sources but also the perspective they helped create. A report on the high prevalence of depression revealed by using a particular questionnaire established both knowledge of mental illness and the means of discovering it.

We focus on the narrative and discourse within primary academic texts rather than on authors, individuals, institutions, and their political and professional affiliations, since these designations and retrospective understandings of their import, we argue, would have been produced by the prevailing narratives and their respective dominance or otherwise. In doing so, we set out to produce an analysis that takes the position that this context is itself a product of certain retrospective narratives, rather than a determinant, and therefore obfuscates the analytic frame. The aim of the analysis that follows is to produce an alternative explanation for the development of the construct of depression in Britain and North America that can then be weighed against other accounts.

The new measurement regime

Approaches to measuring and documenting depression over the last century have seemingly evolved gradually, appearing as a natural iterative process of improved understanding of the phenomena associated with depression. Yet we begin by noting the significant difference between measurement approaches at the start and end of the period under study. This enables us to frame the question guiding our analysis, which is about how developments in psychometric technologies may have impacted on such a shift.

The wider context for our analysis is the development of clinical method from the end of the 18th century, when listening to the patient’s account of their illness began to be superseded by a theory of pathology that prioritised the clinical examination of the

patient's body, to see whether disease could be directly detected (Foucault, 1973). This innovation was gradually extended during the 19th and 20th centuries as various technological supports for examining patients' bodies were introduced, ranging from various 'scopes' through to complex imaging and blood analyses. These devices mediated between the clinician's senses and the patient's illness and objectified the problem as an inscription. Yet despite these revolutionary changes in clinical practice that gained increasing headway during 19th-century medicine, the work of psychiatrists in the asylum continued with the older methods of clinical practice that relied on words rather than physical investigations. While it would have been rare for a patient to report their own insanity, friends and relatives could make a provisional diagnosis and invite the clinician to witness the expression of the patient's disturbed thoughts.

By the early 21st century, however, a large proportion of mental health assessment instead depended on a new mediating device, the patient-completed questionnaire. Instead of a psychiatrist listening to and interpreting the patient's words, the patient could now be invited to complete a questionnaire that would diagnose without the need for psychiatric expertise. The PHQ2 (Kroenke, Spitzer, and Williams, 2003), for example, consists of two questions. Patients are invited to consider the extent to which, over the last two weeks, they have been bothered by 'little interest or pleasure in doing things' and 'feeling down, depressed or hopeless'. A simple scoring system and look-up table offers an estimate of whether or not the patient is clinically depressed.

How is it that unmediated psychiatric diagnoses of the 19th century have been superseded by the results of the psychiatric 'test'? The intervening period was characterised not only by the introduction of self-completion questionnaires but also by fundamental changes in the classification of mental disease and in the form and organisation of psychiatric care. There are numerous narratives seeking to explain these latter broad shifts in the field of mental illness in terms of the actions and interests of powerful individuals and groups, as described earlier. This article, however, using the 'social life of methods', considers how the psychometric questionnaire changed the epistemological landscape of mental illness and created the conditions of possibility for revolutions in psychiatric classification and care in the second half of the 20th century.

Note cards and psychological tests

Reliance on reports from others and interpretation of patients' words and conduct meant that psychiatric diagnosis in the 19th century depended on whatever classificatory frame was used by the clinician – with the implication that diagnoses of insanity would not have been made consistently by all clinicians. Clinical records, which were becoming more important for physical illness in the late 19th century, had less value in the asylum, where decades of incarceration did not require close monitoring and where a quarterly 'no change' entry in the case book would suffice (Andrews, 1998; Turner, 1992). Indeed, given that records were kept in chronological case books, the emphasis was on representing the overall numbers of asylum inmates more than the trajectories of individual patients. It was only in the final decade of the 19th century, when Kraepelin, a notable German psychiatrist, reported using *Zahlkarten* (note cards) on every patient, that a systematised record of insanity began to emerge.

The note card was a new technology that was interpolated between the patient and the clinician and originated from ‘census cards for the mentally ill’ devised by the Royal Statistical Bureau of Prussia in Berlin (Guttstadt, cited in Weber and Engstrom, 1997: 377). *Zahlkarten* allowed alienists to systematically record ‘remarks on aetiology and heredity, medical history, age of first and actual onset, duration of treatment, psychopathological status, course of symptoms, correct diagnoses and diagnostic errors’ (Weber and Engstrom, 1997: 379). Psychiatrists no longer needed to debate diagnoses using narratives reconstructed from successive case books nor parade the patient for all to make their judgements. Armed only with *Zahlkarten* as representations of the patient’s illness, psychiatry had a common experiential base on which to base diagnosis and classification. It is therefore unsurprising that Kraepelin’s classification of insanity carried an authority that none of his predecessors had acquired and became a framework for psychiatric classification in the West for the next century.

About the same time, a new brand of empirical psychology was also seeking a reliable method for accessing the mind. Using technology taken largely from the anthropometric laboratory, it was claimed that some measurements, such as reaction times, could be construed as both physiological and psychological. The new psychological laboratory therefore examined variations in individuals’ mental functioning by using psychophysiological tests: attributes such as ‘keenness of sight, the color sense, judgment of eye (estimation and discrimination of lengths, forms, etc.), touch (discrimination, weight, pain, etc.), movement (discrimination, rate), time-sense, reaction time, mental fatigue, memory, association, etc.’ (Titchener, 1893: 187) formed the basis for a new empirical psychology. Yet, despite attempts to relate these measures to ‘abnormal’ mental functioning (insanity), it was clear a different technology was needed for that purpose. The origins of that new approach emerged in the late 19th century, with experiments using a then-novel method of accessing the mind, the questionnaire.

Although the questionnaire was to become the basis for a new mediating device in the diagnosis of mental illness, the first questionnaires had their origins in the earlier psychological method of introspection. Until the end of the 19th century, psychologists had considered the best method of examining mental functioning to be through a process of thinking about one’s own thoughts. But if a psychologist could do this, why not the psychological subject? The technology for effecting this shift was the questionnaire, which could extend the laboratory outward but also supplement – and in time replace – the psychological method of introspection:

Great as have been the contributions of the laboratory to recent psychology, many most fascinating and important problems as yet resist experimental solution. For the study of these the investigator is thrown back upon introspection and observation, and, so far as his introspection is to have extraneous confirmation, upon the questionnaire. (Miles, 1895: 534)

But what questions should be asked, and how should they be framed ‘without at the same time prejudicing the answers to be received’ (Miles, 1895: 534)? Of particular concern was the ability of subjects to use introspection, a method in which psychologists had been specially trained. Questions such as ‘How do you know your right hand from your left?’, ‘How do you go to sleep when sleepless?’, or ‘What is your favourite colour and

why do you like it?’ depended on respondents being able to reflect on their own mental processes: ‘The inability of some respondents to tell how they recall a forgotten name, or how they set themselves to work when disinclined, shows that these questions approach the limit of casual introspection’ (ibid.: 558).

The other problem with inviting subjects to conduct their own introspection was how to analyse their responses. A series of different statements on why red was a favourite colour needed to be further distilled if inferences about the nature of mental functioning were to be drawn. In a way, early experiments crystallised both the problems and the potential of using questionnaires as means of accessing the mind, problems that were to be overcome and potential that was to be realised over the early decades of the 20th century.

Scales of emotion: Pushing the limits of introspection

The possibility of quantifying any human attribute by means of questionnaire ‘tests’ opened up psychological constructs to empirical fragmentation. Nineteenth-century ‘character’, for example, could be recast as ‘temperament’ that was held to ‘underlie and influence all instincts, and which are related to anatomical and physiological differences and may in time have correlations therewith demonstrated, such as bodily energy, general sthenic emotionality, tendency to be phlegmatic’ (Folsom, 1917: 436). Temperament scales could therefore include emotional reactions such as depression. In 1917, Washburn and colleagues examined respondents’ immediate emotional reaction to a set of words to identify those who were cheerful and those depressed (Baxter, Yamada, and Washburn, 1917; Morgan, Mull, and Washburn, 1919). Respondents were ‘normal’ individuals who could express a feeling of depression (along with cheerfulness and either optimism or pessimism) simply as part of an emotional repertoire without any hint of pathological melancholia or insanity.

In 1930, Jasper devised another new test for measuring emotions (Jasper, 1930). Forty questions covered three ‘dimensions’ of depression-elation, optimism-pessimism, and enthusiasm-apathy. Subjects were asked questions ranging from those about their attitudes towards the future condition of man and their views on morals, war, government, youth, and the Church; to subjective questions such as ‘I tend to have “blue spells”’ and those concerning their thoughts about committing suicide, tiredness, and ambition. The questionnaires were given to college students, refined, edited, and validated against other measures (such as student grades). The result was a self-report measure of depression-elation that was ‘practicable for use with large groups of “normal” individuals of the college age’ (ibid.: 316).

Although Jasper’s questionnaire was developed on and intended for use with ‘normal’ subjects, it was apparent that it could also capture depressive emotions as expressed in asylum psychiatry. He noted, for example, that diagnoses such as Kretschmer’s hypomanic cycloid and depressive cycloid ‘would correspond to the characteristics of the elative disposition and the depressive disposition’, in effect juxtaposing the label ascribed through clinical judgement with the quantified precision of the standardised questionnaire administered to normal populations. Clinical judgement within the asylum had constructed depressed emotions as secondary features of insanity that could be

clinically observed and reported in clinical notes. The psychometric test, however, could elicit these emotions in an 'objective' way directly from the patient and quantify them on a scale enabling ranking and comparability. A patient's emotional state – cheerfulness or depression – could be rendered to a certain granularity with precise scores in the relevant test. It was then a small step to apply the fine-grained technology of psychometric scales to the dense diagnostic categories of insanity.

The potential for fragmenting and quantifying depressive components of psychiatric diagnoses was further realised by other new measures of temperament devised during the 1930s that were developed using psychiatric patients. Humm and Wadsworth, for example, developed a 'Temperament Scale', based on Rosanoff's personality theory found in his *Manual of Psychiatry*, that categorised personalities into normal, hysteroid, cycloid, schizoid, and epileptoid. The cycloid component was described as

characterized by emotionality, fluctuations in activity, and interferences with voluntary attention. . . . The depressed phase is manifested by some degree of sadness, lessened activity, dearth of ideas, and associated characteristics such as worry, timidity, feelings of malaise, and the like. The manifestations of a general cycloid nature are fluctuations from emotional equilibrium, hot-headedness, difficulty in sleeping, etc. (Humm and Wadsworth, 1935: 165)

When Miles had proposed the use of questionnaires to elicit mental functions, she had been concerned that respondents might struggle with the degree of introspection required. Her concerns were partly mitigated by the simple yes/no format of responses required by the early psychological tests. Further, temperament was a stable part of an individual's identity, so responses to questions such as 'Are you sometimes so "blue" that life seems hardly worth living?' (from the Humm-Wadsworth Scale) were not held to reflect current mood but something about a more permanent disposition towards feeling very sad. There were still concerns that even answering these sorts of questions, and the level of reflection on their own mental processes that entailed, might be beyond the ability of respondents. Humm and Wadsworth had noted, for example, that 'the most serious limitation of their scale was the tendency of some subjects to answer with too many "yes" responses or too many "no" responses'. Subjects might also be 'incooperative' and 'answer haphazardly, contrary to fact in a stereotyped manner' (Humm and Wadsworth, 1935: 174). Traxler summarised the issue, noting, 'Since most personality inventories call for self-estimates on a series of items, the question of the correctness with which individuals ordinarily make judgments concerning their own personality characteristics is an important question' (Traxler, 1941: 62).

Validity and reliability: Separating depressives from depression

Over time, psychometricians developed technologies for evaluating the worth of a questionnaire test that would dissemble the unreliability of subjects with a variety of techniques, such as face validity, discriminant validity, construct validity, test-retest reliability, and so on. These tools were first introduced to provide some corrective to the possibility of respondents being 'unreliable', that is, unable to introspect. But as

confidence in respondents' abilities to access their mental processes increased, so the sophistication of how they could be questioned was also extended. In particular, respondents might have the introspective powers to be able not only to answer questions such as 'Do you find yourself at times very cheerful, and at others very blue?' (from the Humm-Wadsworth Scale) but also to offer a more nuanced response in terms of either severity or frequency. Jasper, for example, could invite respondents to choose from a range of responses: 'My most characteristic mood or temperament is: Greatly depressed; Pleasant and fairly happy; Extremely happy and elated; Very happy; Somewhat depressed'. This style of response was later formalised by Likert in his eponymous scale (Likert, 1932).

The effect of using Likert scales to measure aspects of temperament dissolved the immutable character of those mental processes that could be construed as being the essence of identity. A person might have the temperament of being 'sad', but also have degrees of being sad. Moreover, a person who was not sad by nature might yet have moments of sadness. While lengthy personality tests produced a set of personality types that were either present or absent as permanent traits, Likert-type scaling added dimensions of severity and frequency, focusing on a person's current state. This created the possibility for depression to be conceived of as either a fixed personality or a changeable scalable condition.

The questions of test reliability and validity, which had focused on the ability of respondents to access their mental processes and render these accurately in a questionnaire, then began to address the underlying constructs themselves. If the responses to two items or two tests tended to agree, then this might indicate not only that the respondents were answering 'truthfully' but also that the underlying construct (depression, say) was a real entity that the items or tests were accessing. In effect, the test began to reify the construct; the constructs themselves had become real entities, and the tools had become fuzzy devices needing refining in order to better weigh the psychological construct and circumvent subjectivity.

Although developed primarily outside the asylum, the new scales had the potential to undermine the solidity of the asylum diagnosis of insanity (and its variants). Patients could still be labelled as insane, but they could also be classified according to their emotions through the instrument of the psychological test or questionnaire. Emotional states, expressed in numbers, could be shown to vary in severity and over time: aspects of mental illness could be construed as labile. Moreover, test results from normal populations could be shown to overlap with those from psychiatric patients. The implication was that the binary nature of mental functioning (sane or insane) on which the asylum system was based was undermined by the psychological test. Interwar changes in the asylum system, such as the growth of psychiatric outpatients, can be seen as manifestations of the emerging shifts in classification made possible by the new science of psychometrics.

Psychiatric scales: Consolidating clinical depression as changeable

The early 20th-century depression measures, developed mostly outside the asylum, had taken theories of abnormal personality employed within asylum psychiatry and

combined these with psychometric measures of temperament in ‘normal’ populations (mostly college students). In the 1940s, a questionnaire emerged that specifically sought to measure abnormal personality in medical settings. The Minnesota Multiphasic Personality Inventory (MMPI) was intended as a psychiatric measuring device for general medical practice – a comprehensive personality inventory to measure clinically important features ‘without regard to the particular phase of personality upon which the item might bear’ (Hathaway and McKinley, 1940: 43). This would test the limits of the technological communion between clinical judgement and psychometrics.

The MMPI authors developed various subscales: first hypochondriasis, then depression, followed by others. The depression scale (Hathaway and McKinley, 1942) initially had 60 items and was later factor analysed into nine dimensions. Items took the form of statements that had to be answered ‘Yes’, ‘No’, or ‘Cannot say’ in the Thurstone style. For the depression scale, the authors wished ‘to avoid the identification of the term depression with anything other than the presence at the time of testing a clinically recognisable, general frame of mind characterised by a poor morale, lack of hope in the future and dissatisfaction with the patient’s own status generally’ (ibid.: 74). Whereas classical psychiatric theory viewed depression, particularly in the form of melancholia, as a premorbid temperament, the new approach of asking subjects to directly report their own thoughts characterised it as a changeable or ‘reactive’ condition – one that arose in response to external conditions or events. ‘It is well recognized that a few patients with a marked degree of depression on one day may change toward normal within 24 hours’ – a phenomenon that made it difficult to obtain ‘a group of patients clearly depressed at the time of testing’ (ibid.). This concept of changeability made it possible to then consider all the many external factors that an individual might ‘react’ to with a depressive mood:

Such a clinical picture might result from economic or vocational frustration, from personal problems, from a depressive phase of a cycloid personality, or from any one of the other commonly known clinical backgrounds of depression. As seen in this way, the measured depression might represent a less stable trait in the individual than . . . most other measured personality characteristics. (ibid.)

In freeing depression from melancholic psychosis (psychosis also having been construed as a personality type), the psychometricians had established depression as a potentially unstable component of psychological functioning that might react to external events and situations. Indeed, the term *reactive depression* had begun to emerge and develop in the psychiatric literature concurrently with the growth of psychometrics during the 1930s, 1940s, and 1950s. An early use of the term shows the emerging separation of ‘reactive’ depression from psychotic (assumed to be premorbid characterological) depression:

Although there were 24 cases classified as affective reactions only six of these could be counted unquestionably psychotic. The remainder were mild disorders – depressive in character – where the mood was more in keeping with a real life situation and where the capacity for social adaptation was only partly affected. These are the cases sometimes classified as reactive depressions. (Jefferson, 1933: 833)

In summary, the emergence of psychometric techniques applied to 'normal' populations enabled 'mild' forms of temporary ('reactive') depression (so far loosely defined) to be identified among non-incarcerated patients. Psychiatric concepts of depression that derived from clinical observation of inmates, combined with clinical theory and clinical judgement, had typically concluded that depression constituted one face of a double-sided 'cycloid' personality. During the 1950s, however, the concept of 'reactive depression' (and 'neurotic' depression) became increasingly separated from the concept of a permanent, incurable, psychotic personality.

Sensitivity to change

In the early post-war years, imipramine, originally tested for antipsychotic effects, was later believed to have 'antidepressant' properties. This new class of drugs did not offer a 'cure' but an amelioration of symptoms, and therefore testing for efficacy involved the application of questionnaire technology rather than cruder clinical judgements. In the first published study, for example, the effect of imipramine was studied using a fairly primitive form of psychometric measurement:

The criteria of improvement consisted of 4 items: symptoms' disappearance (subjective comfort); ward management; ability to go home; and ability to go to work (social effectiveness). The realization of 4, 3 or 2 of these items was categorized as marked, moderate or slight improvement respectively. (Azima and Vispo, 1958: 245)

Given the small and often subtle effects of drugs on mood, sensitive techniques were needed to detect change. The worth of a scale, therefore, would come to be determined not by how well it reflected current theories or classifications in psychiatry, but by its sensitivity to change over a short period. This in turn meant that the content of measures deemed sensitive to change would begin to mark out the boundaries of depression, and thereby shape classification. The Beck Depression Inventory (BDI; Beck *et al.*, 1961), the Hamilton Depression Rating Scale (HRSD; Hamilton, 1960), and the Wechsler Depression Rating Scale (DRS; Wechsler, Grosser, and Busfield, 1963), for example, emerged in the 1960s, all using Likert systems, with a view to detecting changes that might be induced by either antidepressant drugs or psychological therapies.

Doubts and disagreements about depression types and causes became less relevant as the new methods allowed depression as a symptom to be universally scaled. Given that the 'abnormal' population was now merging with the 'normal' population as a result of asylum closures (which also reduced the opportunity for clinical observation and hence clinical judgement), scaling technologies were now the defining feature of psychiatric research. These scales could at once diagnose (without the need for prolonged daily observation) and measure intensity of emotions. Their key attribute was the ability to be sensitive to change. 'The problem of assessment becomes most crucial in the studies attempting to evaluate treatment effects, since they require an assessment not only of the patient's condition but also of change in that condition' (Wechsler, Grosser, and Busfield, 1963: 335).

Initially, other than use of the Likert system, the new psychiatric scales for measuring depression were relatively diverse in content. All tended to have some common items. The BDI, HRSD, and DRS, for example, all asked about mood, suicidality, loss of interest or energy, changes in appetite, and weight loss. But then each instrument also contained additional items that were not shared, such as loss of insight, guilt, low self-esteem or self-hate, social withdrawal, indecisiveness, sleep disturbance, lack of satisfaction, sense of punishment, body image, fatiguability, somatic preoccupation, and loss of libido. The heterogeneous choice of items underpinning each scale did not derive solely from the perceived nature of depression and/or depressive illness but from questions that had the potential to be sensitive to change.

Techniques for assessing sensitivity to change had become more sophisticated by the 1970s. The Montgomery Åsberg Depression Rating Scale (MADRS; Montgomery and Åsberg, 1979), for example, included items selected exclusively on the basis of sensitivity to change while also taking into account the correlation between item change and overall change. This meant that the item 'reduced sexual interest' was excluded from the scale as, while it changed significantly over the course of drug treatment, change was in the wrong direction: 'Reduced sexual interest yielded large changes but was less well correlated to general outcome. Inclusion of an item like this in a scale might spuriously inflate the change scores' (ibid.: 384). In effect, the final nine-item questionnaire was not measuring a clinical or theoretical construct of depression but rather measuring emotions that changed over the typical four to eight weeks measured in drug trials.

Most of the individual items in the post-war depression detection scales had equivalents in their predecessor personality questionnaires, reframed as things that could have intensity, frequency, and changeability rather than things that either did or did not reflect one's character. Humm and Wadsworth had asked, 'Do you find yourself at times very cheerful and at others very "blue"?' (yes or no). Beck now asked subjects to rate their 'Mood' as 'I do not feel sad', 'I feel blue or sad', 'I am blue or sad all the time and I can't snap out of it', 'I am so sad or unhappy that it is very painful', or 'I am so sad or unhappy that I cannot stand it'. Yet, despite these continuities in the item content of questionnaires, the fundamental question had changed. For Humm and Wadsworth, and other interwar investigators, the question was whether or not the particular respondent had clinical depression, whereas post-war questionnaires were driven by the need to detect change. Sadness might therefore have remained a central emotion to elicit, but the heterogeneity of items in post-war questionnaires reflected the various attempts to identify new ways of discerning small shifts in mood.

Maintaining internal reliability: Major depressive disorder

When using these new depression scales, individual item scores were summed to create a total score with proposed cut-off points to indicate levels of severity of depression; thus 'mild' and 'major' depression could now be precisely defined. A new term, 'major depressive disorder' (MDD), was introduced in the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III; American Psychiatric Association, 1980), reflecting the new construct of severity made possible by psychometric scales. Although the new depression scales could capture changes in mood over time, these were

mood swings within the range of depressed features – the scales could not accommodate an entirely different dimension of manic symptoms while retaining internal reliability. ‘Bipolar disorder’ (an evolution of ‘cyloid personality’ capturing the idea of extreme cyclical changes from severe depression to mania) therefore had to be separated from MDD in the new classification system. DSM-III also introduced ‘dysthymic disorder’, which was set out in DSM-III as a form of depression that was both more chronic and less severe than MDD – more in line with the older construct of a depressed ‘personality’. Depression scales could detect moderate changes over short periods but were impractical for detecting smaller fluctuations over prolonged periods of years, as implied by ‘dysthymia’.

Depressive personality disorder was not included in DSM-III but reappeared in DSM-IV (American Psychiatric Association, 1994) as a set of research criteria with a view to considering reinstating it in a subsequent version. Items proposed were character descriptors in the yes/no format, such as ‘Is critical, blaming, and derogatory toward oneself’ or ‘Is brooding and given to worry’. The concepts captured by these items all overlapped with the Likert-style items in depression severity scales (such as low mood, low self-esteem, guilt, self-criticalness, worry, pessimism, guilt, etc.), but were framed as traits with yes/no responses. This formal separation of permanent personality versus a treatable condition appears to have come about because of the two different techniques available for scale construction.

Inter-rater reliability: Discovering unreliable clinicians

Although depression scales were intended to capture patients’ moods by facilitating emotional introspection and converting this into scores in a questionnaire, not all instruments precluded psychiatric judgement. While the BDI, for instance, was entirely self-report, the HRSD was based on clinicians asking subjects questions and the clinician making a rating, whereas the DRS was a combination of both of these, while also including a number of additional clinical observation items covering the patient’s physical appearance, observed speech, voice, tension, and so on (the type of observation previously seen in turn-of-the-century asylum anecdotes). Over previous decades, psychiatry had struggled with the potential unreliability of patients’ reports; but was clinical judgement, as in mediating patients’ words, equally at risk of unreliability? The problem was compounded by the considerable expansion of the number of professions involved in mental health care, now that asylums had been emptied and community mental health care established. Could all these clinicians be relied upon to make correct judgements? Concerned about irregular use of the clinician-rated HRSD, Williams published a standardised structured interview to ensure researchers were applying the HRSD interview systematically and consistently (Williams, 1988). Further, inter-rater reliability testing (increasingly used in psychometrics from the 1960s onwards) could be used to demonstrate that trained raters could be relied on to make judgements similar to those of a more expert clinician, the latter remaining the gold-standard criterion.

In 2008, Williams also published a structured interview guide for the MADRS (Williams and Kobak, 2008), as well as a new structured interview guide for the HRSD (Williams *et al.*, 2008):

Accumulating evidence suggests that the quality of ratings can make the difference between a failed trial and one in which drug separates from placebo. Therefore, any method that improves the quality of clinical trial ratings may improve our ability to conduct successful antidepressant trials. (Williams and Kobak, 2008: 52)

The ideal measure remained the one that directly captured patients' reports. Just as the explosion of investigatory techniques in physical medicine in the second half of the 20th century interpolated 'objective' measures of disease, so the questionnaire that captured mood offered a means of bypassing the subjectivity of psychiatric judgement. In the early 21st century, a new measure was developed, the PHQ9 (Kroenke, Spitzer, and Williams, 2001). This new instrument had nine items, all with an almost exact equivalent item in the MADRS (developed to prioritise sensitivity to change within drug trials). The PHQ9 was entirely self-report, requiring no clinician involvement. The purpose was to create a measure that would detect depression in primary care, where the majority of mental illness had now come to be managed. The measure needed to be short, since the HRSD 'can require 15-30 minutes of clinician time to administer and is therefore not feasible in many practice settings' (ibid.: 612). The authors noted that the HRSD required training and had a complicated scoring system; the MADRS required rating by a clinician with special training. They considered that the then-popular self-rated BDI was not adequately sensitive to change. The PHQ9 would, it was claimed, address all of these issues and provide a tool that would enable greater detection of depression across primary care settings in a quick, inexpensive, and reliable way. The PHQ9 has since been used for routine outcome monitoring in the UK primary care psychological therapy service (Improving Access to Psychological Therapies, IAPT) since 2008 and is recommended as a primary care screening tool by various national bodies, including the American Psychological Association.

In 2003, the same group also introduced the PHQ2, 'because even briefer measures might be desirable for use in busy clinical settings or as part of comprehensive health questionnaires' (Kroenke, Spitzer, and Williams, 2003: 1284). Psychometrics now offered a solution to general practitioner unreliability with their limited training in psychiatry: statistical techniques to identify the minimal number of items needed to maximise detection. PHQ2 development work identified two items (low mood and loss of interest or pleasure) that would identify patients who would score for major depression if tested on a longer test such as the PHQ9 or by a trained clinician. While the PHQ2 is not a formal requirement for primary care screening in the UK or USA, it reflects the potential of psychometric advances to impact on the wider conceptualisation of depression.

Conclusions

Our analysis proposes that the shift from clinical judgement and observation in the asylum to the psychometrics of 'normal' populations reordered the spatial distribution of mental illness, both geographically and conceptually. Depression became changeable, indeed treatable, since subjects (now the data source) were inherently unreliable and changed their accounts frequently. Psychometric techniques of correlation, factor

analysis, and means of assessing reliability and validity allowed subject unreliability to be statistically managed and thus to become acceptable for use among psychiatric populations, even though the assumption remained for some time that clinicians were naturally the best judges. Variations in scaling techniques enabled the separation of depressive personality from reactive depression, paving the way for measuring severity and intensity of emotions. Techniques to test sensitivity to change, along with new research designs (such as randomised controlled trials), enabled these severity measures to ‘prove’ the efficacy of the new treatments, which at that time were psychoactive drugs. Latterly, more advanced techniques of precision scaling made it possible to manage the new measurement problem of clinician unreliability that resulted from the growing number of professions and professionals involved in mental health care. This has left the construct of depression reduced from hundreds of items to potentially just two that best manage subject and clinician unreliability and, as a further effect, maximises the number of subjects under the gaze of mental health professionals.

As noted at the start of this article, there are many alternative accounts of the history of depression attempting to explain the manufacture of depression as a form of growing pandemic (e.g. Greenberg, 2010; Rose, 2018). However, in focusing on the social life of methods, the analysis here emphasises an aspect of the history of psychiatry that is often elided in more ‘political’ accounts. The development and spread of ‘objective’ methods of identifying depression, seemingly independent of both practitioners’ and patients’ idiosyncrasies, not only provided the basis for an empirical psychiatry but also appeared to confirm again and again, in thousands of encounters, the reality of depression. This ‘validated’ construct then became the bedrock for the late 20th-century paradigm of depression, one that could be challenged or negotiated from within but not from outside. If huge numbers of patients suffer from depression – as evidenced from questionnaire data – the reality of the construct and the scale of the problem cannot be ignored.

One of the defining characteristics of recent historical accounts is their critical stance towards the way in which more and more people are caught in the net of a depression diagnosis. For some authors, it is sufficient to decry those they hold responsible, often psychiatrists or, more commonly, the pharmaceutical industry. For others, there is a belief that underlying the swathe of depression diagnoses is a ‘real’ depression that may ultimately be identified through some biological characteristics or responsiveness to certain treatments (Shorter, 2009; Shorter and Fink, 2010), reflecting the ‘evolutionary criteria for how human beings are biologically designed to behave’ (Horwitz and Wakefield, 2007). More prosaically, it should be possible to separate out ‘very sick people’ (Hirshbein, 2009: 8). For these critics, it is not the fault of DSM symptom-based diagnostic criteria that confused ‘intense normal sadness’ with depressive disorder (Horwitz and Wakefield, 2007) but the very nature of the continuous distribution of depression scores derived from psychiatric instruments. Indeed, as Frances (2013: 18) suggests, the problem is inherent in the shape of that distribution:

This brings us to the question of the moment – can we use statistics in some simple and precise way to define mental normality? Can the bell curve provide a scientific guide in deciding who is mentally normal and who is not? Conceptually, the answer is ‘why not’, but practically the answer is ‘hell no’... The normal curve tells us a great deal about the

distribution of everything from quarks to koalas, but it doesn't dictate to us where normal ends and abnormal begins.

Our analysis proposes that the modern terrain of depression – the continuity between the abnormal and the normal – is the direct result of psychometric shifts earlier in the 20th century. In other words, the explanatory frameworks adopted by secondary sources all start from varying degrees of engagement with the continuous scales of mood that were invented only a few decades earlier. Once accepted, the outputs from these psychiatric instruments then become the substrate on which to predicate explanatory models and apportion responsibility, a story with 'human action' at its heart (Herzberg, 2009). The social life of methods presents a different picture, one in which the inexorable logic of psychometrics produces its own reality.


Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Susan McPherson  <https://orcid.org/0000-0002-9478-9932>

References

- American Psychiatric Association (1980) *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.). Washington, DC: American Psychiatric Association.
- American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: American Psychiatric Association.
- Andrews, J. (1998) 'Case Notes, Case Histories, and the Patient's Experience of Insanity at Gartnavel Royal Asylum, Glasgow, in the Nineteenth Century', *Social History of Medicine* 11(2): 255–81.
- Azima, H. and Vispo, R. H. (1958) 'Imipramine: A Potent New Anti-depressant Compound', *American Journal of Psychiatry* 115(3): 245–6.
- Baxter, M. F., Yamada, K., and Washburn, M. F. (1917) 'Directed Recall of Pleasant and Unpleasant Experiences', *American Journal of Psychology* 28(1): 155–7.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961) 'An Inventory for Measuring Depression', *Archives of General Psychiatry* 4(6): 561–71.
- Bowker, G. C. and Star, S. L. (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
- Espeland, W. N. and Stevens, M. L. (2008) 'A Sociology of Quantification', *European Journal of Sociology* 49(3): 401–36.
- Folsom, J. K. (1917) 'A Statistical Study of Character', *Pedagogical Seminary* 24(3): 399–440.
- Foucault, M. (1973) *The Birth of the Clinic*, trans. A. M. Sheridan Smith. New York, NY: Pantheon Books.

- Frances, A. (2013) *Saving Normal: An Insider's Revolt Against Out-of-Control Psychiatric Diagnosis, DSM-5, Big Pharma, and the Medicalization of Ordinary Life*. New York, NY: William Morrow.
- Greenberg, G. (2010) *Manufacturing Depression: The Secret History of a Modern Disease*. New York, NY: Simon & Schuster.
- Hamilton, M. (1960) 'A Rating Scale for Depression', *Journal of Neurology, Neurosurgery & Psychiatry* 23(1): 56–62.
- Hathaway, S. R. and McKinley, J. C. (1940) 'A Multiphasic Personality Schedule (Minnesota): I. Construction of the Schedule', *Journal of Psychology* 10(2): 249–54.
- Hathaway, S. R. and McKinley, J. C. (1942) 'A Multiphasic Personality Schedule (Minnesota): III. The Measurement of Symptomatic Depression', *Journal of Psychology* 14(1): 73–84.
- Healy, D. (2004) *Let Them Eat Prozac: The Unhealthy Relationship Between the Pharmaceutical Industry and Depression*. New York, NY: New York University Press.
- Herzberg, D. L. (2009) *Happy Pills in America: From Miltown to Prozac*. Baltimore, MD: Johns Hopkins University Press.
- Hirshbein, L. D. (2009) *American Melancholy: Constructions of Depression in the Twentieth Century*. New Brunswick, NJ: Rutgers University Press.
- Horwitz, A. V. (2020) *Between Sanity and Madness: Mental Illness From Ancient Greece to the Neuroscientific Era*. New York, NY: Oxford University Press.
- Horwitz, A. V. and Wakefield, J. C. (2007) *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow Into Depressive Disorder*. Oxford: Oxford University Press.
- Humm, D. G. and Wadsworth, G. W. (1935) 'The Humm-Wadsworth Temperament Scale', *American Journal of Psychiatry* 92(1): 163–200.
- Jasper, H. H. (1930) 'The Measurement of Depression-Elation and Its Relation to a Measure of Extraversion-Introversion', *Journal of Abnormal and Social Psychology* 25(3): 307.
- Jefferson, R. A. (1933) 'Psychiatric Experiences in a Medical Out-Patient Clinic', *American Journal of Psychiatry* 89(4): 831–8.
- Kroenke, K., Spitzer, R. L., and Williams, J. (2001) 'The PHQ-9: Validity of a Brief Depression Severity Measure', *Journal of General Internal Medicine* 16(9): 606–13.
- Kroenke, K., Spitzer, R. L., and Williams, J. (2003) 'The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener', *Medical Care* 41(11): 1284–92.
- Lieberman, J. A. and Ogas, O. (2015) *Shrinks: The Untold Story of Psychiatry*. New York, NY: Little, Brown.
- Likert, R. (1932) 'A Technique for the Measurement of Attitudes', *Archives of Psychology* 22(140): 1–55.
- Miles, C. (1895) 'A Study of Individual Psychology', *American Journal of Psychology* 6(4): 534–58.
- Montgomery, S. A. and Åsberg, M. (1979) 'A New Depression Scale Designed to Be Sensitive to Change', *British Journal of Psychiatry* 134(4): 382–9.
- Morgan, E., Mull, H. K., and Washburn, M. F. (1919) 'An Attempt to Test Moods or Temperaments of Cheerfulness and Depression by Directed Recall of Emotionally Toned Experiences', *American Journal of Psychology* 30(3): 302–4.
- Rapley, M., Moncrieff, J., and Dillon, J., eds (2011) *De-medicalizing Misery: Psychiatry, Psychology and the Human Condition*. Houndmills: Palgrave Macmillan.
- Rose, N. S. (2018) *Our Psychiatric Future: The Politics of Mental Health*. Medford, MA: Polity.

- Savage, M. (2013) 'The "Social Life of Methods": A Critical Introduction', *Theory, Culture & Society* 30(4): 3–21.
- Shorter, E. (2009) *Before Prozac: The Troubled History of Mood Disorders in Psychiatry*. Oxford: Oxford University Press.
- Shorter, E. (2013) *How Everyone Became Depressed: The Rise and Fall of the Nervous Breakdown*. Oxford: Oxford University Press.
- Shorter, E. and Fink, M. (2010) *Endocrine Psychiatry: Solving the Riddle of Melancholia*. Oxford: Oxford University Press.
- Titchener, E. B. (1893) 'Anthropometry and Experimental Psychology', *Philosophical Review* 2(2): 187–92.
- Traxler, A. E. (1941) 'Current Construction and Evaluation of Personality and Character Tests', *Review of Educational Research* 11(1): 57–79.
- Turner, T. H. (1992) 'A Diagnostic Analysis of the Casebooks of Ticehurst House Asylum, 1845–1890', *Psychological Medicine. Monograph Supplement* 21: 1–70.
- Weber, M. M. and Engstrom, E. J. (1997) 'Kraepelin's "Diagnostic Cards": The Confluence of Clinical Research and Preconceived Categories', *History of Psychiatry* 8(31): 375–85.
- Wechsler, H., Grosser, G. H., and Busfield, B. L. (1963) 'The Depression Rating Scale: A Quantitative Approach to the Assessment of Depressive Symptomatology', *Archives of General Psychiatry* 9(4): 334–43.
- Williams, J. B. W. (1988) 'A Structured Interview Guide for the Hamilton Depression Rating Scale', *Archives of General Psychiatry* 45(8): 742–7.
- Williams, J. B. W. and Kobak, K. A. (2008) 'Development and Reliability of a Structured Interview Guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA)', *British Journal of Psychiatry* 192(1): 52–8.
- Williams, J. B. W., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., Olin, J., Pearson, J., and Kalali, A. (2008) 'The GRID-HAMD: Standardization of the Hamilton Depression Rating Scale', *International Clinical Psychopharmacology* 23(3): 120–9.

Author biographies

Susan McPherson is Professor of Psychology and Sociology in the School of Health and Social Care at the University of Essex.

David Armstrong is Emeritus Professor of Medicine and Sociology at Kings College London. He is the author of *Political Anatomy of the Body* (1983) and *A New History of Identity* (2002).