British Journal of Psychiatry (1996), **168**, 594–597

Present Use of the Hamilton Depression Rating Scale: Observations on Method of Assessment in Research of Depressive Disorders

R. P. SNAITH

Background. The Hamilton Depression Rating Scale retains its primacy in research. There have been recent important critiques. It is clear that instructions provided by its author are widely overlooked.

Method. A survey of the present use of the HDRS was conducted by inspection of five major journals publishing studies in the field of psychiatry. Note was especially made of whether a recognised version of the Scale was quoted; also of whether authors had selected specific scores on one or other of the versions to indicate a criterion for inclusion of a subject in a study, and likewise whether a specific score had been selected as an indication of recovery following some procedure or treatment.

Results. One hundred and fourteen articles were reviewed in which 71 had used a depression scale. This was the HDRS in 66% of the studies. There was considerable evidence that the instruction that the HDRS was only to be used in situations where the patient had received a diagnosis of a primary depressive illness had been ignored. There was considerable degree of arbitrary selection of Scale scores.

Conclusions. The survey causes concern about the methodology of much research in the field of assessment of severity of psychiatric disorder. The rationale of assessment by the rating scale method is considered and suggestion made for improvement in research practice.

The Hamilton Depression Rating Scale (HDRS) appeared in 1960 and a subsequent analysis of data obtained from its use was published by its author (Hamilton, 1967). It was rapidly incorporated into research practice and is still widely used. It is now often used in a manner and for studies for which it was not devised. Hamilton had stated that the HDRS was a scale for the assessment of severity of the disorder in studies of patients who had received a diagnosis of primary depressive illness (presently termed major depressive disorder); he specifically stated that use in other circumstances could not be justified. His instructions for its use included the advice on independent rating by two researchers and a taking into account of information from nurses and relatives as well as direct interview with the patient. Such instructions appear to be widely ignored.

There was very little early criticism or questioning of the HDRS and nearly two decades were to pass before serious consideration was given to its psychometric aspects. One such comment (Bech & Rafaelson, 1980) noted a poor internal consistency was a disadvantage. Such occasional criticisms did not seriously affect the use of the Scale.

Now two major publications call for attention to the purpose and the structure of the HDRS. The first of these is a summary of the papers read at a symposium convened for the purpose of a reconsideration of the Hamilton Scales (Bech & Coppen, 1990). The second is a critique by an American research team (Gibbons *et al*, 1993).

Most of the contributors to the symposium had a criticism to make or problems to be discussed. Only the most pertinent points will be mentioned here. Zitman remarked that the different versions of the HDRS caused "uncertainty as to the meaning of scores" and Bech stated that the HDRS "had been released with insufficient operational definition". Paykel commented on the "vagaries" of the HDRS and advised that other measures should be used in addition to it. Williams considered that the provision of details of a structured interview was necessary to "overcome uncertainty of the scoring procedure". Gastpar & Gilsdorf reflected "with astonishment" that, despite careful planning and

https://doi.org/10.1192/bjp.168.5.594 Published online by Cambridge University Press

training of the researchers there was considerable variation between centres which led them to the somewhat lame conclusion that "the HDRS is never completely unreliable but partly reliable everywhere".

The publication by Gibbons *et al* (1993) presents a very much more penetrative comment, not only on the HDRS, but on the practice of assessment of the severity of a construct, 'depression', composed of cognitive, somatic, behavioural and social aspects by an instrument which is also composed of a mixture of the phenomena of disordered mood states; they concluded that this practice led only to "obfuscation and misunderstanding".

Although I did not attend the symposium I was a trainee psychiatrist in Max Hamilton's department at the time when he was undertaking the revision of his Depression Scale; I recall his attitude to the instrument he had created and his certainty of the misuses to which it would be put – despite his clearly stated directions. Although Hamilton agreed with the criticism that patients suffering from other disorders would also achieve high scores on the Scale he declined to alter the item content, which he considered to provide a comprehensive coverage of the symptoms which usually featured in such an illness. It was unfortunate that he did not remove a set of items reflecting less commonly occurring symptoms; consequently there arose two versions of the HDRS, one with 17 items and the other with the four further items, the 21-item version. It was also very unfortunate that, having created the HDRS, he did not pursue work to provide information on the interpretation of the scores or of the sensitivity to change of severity of the depressive disorder. The instructions for the rating procedure remained unchanged. Analysis of the data from the second study confirmed the earlier finding of lack of homogeneity and several factors emerged: (a) 'overall severity', (b) 'psychic *v.* somatic anxiety', (c) 'agitation *v.* retardation' and other more complex factors which differed between the sexes. No age effect emerged but this was to be expected since the study sample was composed largely of middle-aged hospitalised patients.

The purpose of the present study is to review the manner in which the HDRS is used in research practice today; the investigation enquired into the following points: whether HDRS retains pre-eminence as the choice for an assessment scale for severity of depression disorder, and which version is most frequently chosen; whether Hamilton's instruction concerning confinement to use in specific diagnosis is followed; whether particular scores on either of the versions have been adopted to indicate a particular level of severity; and how change scores on the HDRS are interpreted.

Method

The enquiry concentrated entirely on research reports in journals that publish studies in general psychiatry. Five leading journals were selected on the grounds that they covered studies presented in the English language throughout the world. These were *Acta Psychiatrica Scandinavica*, *American Journal of Psychiatry*, *British Journal of Psychiatry*, *Archives of General Psychiatry*, and *Psychological Medicine*. The last complete one year period prior to the review i.e. 1994, was taken and contents lists were examined. Note was made of all studies which contained, within their title, the terms 'depression', 'affective disorder', 'bipolar disorder' and 'dysthymia'. The method of the study was examined, particularly with regard to the use of any scale to assess severity of the depressive disorder or depressive mood disorder.

If the HDRS was used in the study, note was made of whether the version was stated i.e. the 17-item, the 21-item or some other adaptation of the Scale. Citation of a reference to one of the two papers presenting the HDRS did not provide the required information since both presentations allowed either of the versions to be used. The method of the study was then examined in order to determine whether particular scores of the HDRS had been selected for a special purpose e.g. to provide a criterion for inclusion in the study or an indication of recovery from the disorder.

Results

The findings of the survey are summarised in Table 1. The questions posed above may now be answered.

Hamilton Depression Rating Scale retains its pre-eminence as the instrument for first choice in the majority of studies where a researcher-administered depression scale is used in the study. It was the chosen scale, and the one on which conclusions were based in 66% of the studies.

There is uncertainty and confusion concerning the version selected for the study: in 40% of the studies there was no statement of the version used and in 6 (13%) studies some other version or some personal adaptation of the HDRS had been used.

There is a wide tendency to ignore the instruction to use the scale for assessment of severity of the disorder in patients who had already been diagnosed as suffering from a primary depressive illness.

https://doi.org/10.1192/bjp.168.5.594 Published online by Cambridge University Press

Table 1
Survey of a one-year series of articles and the use of the Hamilton Depression Scale

Total number of articles on depression	114
No measure of 'depression'	39
Studies excluded on grounds of same authorship	4
Did not use HDRS	23 (34%)
Used HDRS	47 (66%)
17-item	16 (34%)
21-item	6 (13%)
24-item	4
25-item	1
31-item	1
version not stated	19 (40%)
inclusion score used	22 (47%); range: 10–20
recovery score used	15 (32%); range: 5–10
other 'recovery' criteria:	reduction of 50% of initial score: 7 studies

A selected score on one or other version of the HDRS had been used in 22 (47%) of the studies for a statement of diagnosis of disorder or for justification of inclusion as a case in a study. In most cases the score selected was arbitrary although in a few studies, justification for the choice of score was derived from its use in some other study to which reference was given. Similarly an arbitrary choice of a score on one of the versions of the HDRS was used to justify a statement of 'recovery' from the depressive disorder in nearly a third of the studies surveyed. Frequently 'recovery' was defined in terms of some degree of a percentage fall in the scale score.

Discussion

The findings of this review do not reflect well on present research procedure. Such arbitrary use of a scale devised for a specific purpose cannot be justified. Apart from the ignoring of Hamilton's instruction that the HDRS could only be considered to be a valid instrument in the assessment of severity of disorder in patients who had been diagnosed as suffering from a specific type of depressive disorder there was practically no evidence that his other requirement for the use of the scale, the use of two raters and culling of information from other sources, had been followed. The confusion that has arisen over the existence of two versions of the HDRS and also the fact that no definite scale scores were ever established to indicate different levels of severity must be laid at the door of its author.

The fact that some researchers provide personal selection of a scale score for the purpose of the study without even stating which version of the HDRS had been used is not acceptable scientific practice. Clearly a particular score will have uncertain significance if it is derived from measures composed of different numbers of items. Such a 'moving-of-the-goalpost' procedure enables inclusion and exclusion of subjects to a study in a manner which will have a major effect on conclusions drawn. Hedlund & Vieweg (1979) commented on the difficulties in the integration of different research findings caused by the varied versions of the HDRS and lack of indication of the version used in the published studies.

The practice of a definition of a depressive disorder by a score on a scale composed of a wide variety of components i.e. somatic symptoms, behaviour, personal preoccupations, mood change and other phenomena of mental disorder has become established practice in psychiatric research, both as regards the definition of disorder and the composition of instruments to assess their severity. The comment of Gibbons *et al* (1993) must be heeded; for instance a change score, which may be supposed to reflect improvement, could well be produced by an improvement in somatic symptoms while leaving suicidal preoccupation and motivation for normal activity substantially unaltered or even worse. There is a general impression that the HDRS is the standard for measure of any disorder considered to be a form of 'depression' and that, so long as the Scale is used research procedure and editorial attitude will be in accord with any individual adaptation.

There are, of course, wider implications for research practice than that illustrated by this survey of the use of the HDRS. Similar surveys of other measures of severity of psychiatric disorders would arrive at a similar conclusion unless the measure had been introduced with inflexible statement of the significance of a score. As Gibbons *et al* (1993) have pointed out, the whole procedure of assessing disorders composed of a variety of phenomena by scales similarly composed leads to "obfuscation and misunderstanding". The fact that most research into mental illness does proceed in this manner is a matter requiring serious consideration. Farmer & McGuffin (1989) called for a different approach to the assessment of depressive disorders without indicating the route which should be followed. One direction, recently proposed in this journal (Costello, 1992; van Pragg, 1992) advocates the abandonment of research on multifaceted syndromes and a concentration on well defined psychopathological constructs. Certainly the use of 'depression', as of any other rating scales, calls

https://doi.org/10.1192/bjp.168.5.594 Published online by Cambridge University Press

for a careful scrutiny of what exactly they may be presumed to measure (Snaith, 1993). In the present survey an example of how the choice of rating scale could influence research conclusions was provided by one study (Smith *et al*, 1994): both the HDRS and the Montgomery–Åsberg Depression Rating Scale had been used; the one scale but not the other indicated a significant difference between study and comparison samples prior to the investigation, and at a 14-day follow-up the change in severity was at a much higher level of statistical significance with one scale compared with the other.

References

Bech, P. & Coppen, A. (eds) (1990) The Hamilton Scales. *Psychopharmacology Series*, **9**. New York: Springer Verlag.

——— & Rafaelsen, O. J. (1990) The use of rating scales exemplified by a comparison of the Hamilton and the Bech–Rafaelsen Melancholia Scale. *Acta Psychiatrica Scandinavica*, **62** (Suppl. 285), 128–131.

Costello, C. C. (1992) Research on symptoms versus research on syndromes; arguments in favour of allocating more time to the study of symptoms. *British Journal of Psychiatry*, **160**, 304–308.

Farmer, A. & McGuffin, P. (1989) The classification of depressions: contemporary confusion revisited. *British Journal of Psychiatry*, **155**, 437–443.

Gibbons, R. D., Clark, D. C. & Kupfer, D. J. (1993) Exactly what does the Hamilton Depression Scale measure? *Journal of Psychiatric Research*, **27**, 259–273.

Hamilton, M. (1960) A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, **23**, 56–62.

——— (1967) The development of a scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, **6**, 278–296.

Hedlund, J. & Vieweg, B. W. (1979) The Hamilton Scale For Depression: a comparative review. *Journal of Operational Psychiatry*, **10**, 149–165.

Smith, J., Williams, K., Birkett, S., *et al* (1994) Neuroendocrine and clinical effects of electroconvulsive therapy and their relationship to treatment outcome. *Psychological Medicine*, **24**, 547.

Snaith, R. P. (1993) What do depression scales measure? *British Journal of Psychiatry*, **163**, 293–298.

van Praag, H. M. (1992) Reconquest of the subjective: against the waning of psychiatric diagnosis. *British Journal of Psychiatry*, **160**, 266–271.

R. P. Snaith, FRCPsych, Academic Unit of Psychiatry, Clinical Sciences Building, St James's University Hospital, Leeds LS9 7TF

(First received 12 June 1995, final revision 18 December 1995, accepted 18 December 1995)

https://doi.org/10.1192/bjp.168.5.594 Published online by Cambridge University Press