Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Psychiatric Research 40 (2006) 243–246
JOURNAL OF PSYCHIATRIC RESEARCH
www.elsevier.com/locate/jpsychires

# Validation of an IVRS version of the MADRS

James C. Mundt [a], David J. Katzelnick [a,*], Sidney H. Kennedy [b], Beata S. Eisfeld [b], Beverley B. Bouffard [b], John H. Greist [a]

[a] *Healthcare Technology Systems, Inc., 7617 Mineral Point Road, Madison, WI 53717, United States*
[b] *University Health Network, Department of Psychiatry, 200 Elizabeth Street, Toronto, Ont., Canada M5G 2C4*

Received 2 March 2005; received in revised form 25 April 2005; accepted 6 May 2005

**Abstract**

Interest in self-reported measures of depression in clinical trials has grown in recent years. This study compared the reliability and validity of the clinician-administered Montgomery-Asberg Depression Rating Scale (MADRS) to a computer-administered version administered over the telephone using Interactive Voice Response (IVR) technology.

Sixty subjects were administered both the clinician- and computer-administered versions of the MADRS in a counter-balanced order. A subsample of 20 patients was reassessed 24 h later by both methods.

Mean score differences between IVR and clinician were not statistically significant (<1 point) and a high correlation was found between forms ($r = .815$, $p < .001$). Reliability measures (Cronbach's Alpha and 24-h test–retest) were comparable. Clinicians rated the severity of subjects' sadness and pessimistic thoughts lower than subjects self-report.

The data obtained in this pilot study provide support for the equivalence between the clinician and IVR versions of the MADRS.


*Keywords:* Depression; Self assessment; Outcome assessment; Computers; Validation study; Computer communication networks

## 1. Introduction

Treatment outcomes in antidepressant medication trials have traditionally used clinician-administered rating scales such as the Hamilton Depression Rating Scale (HAMD) (Hamilton, 1960), Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979), and the Inventory of Depressive Symptomotology (IDS) (Rush et al., 1996). Recently, these measures have received increased scrutiny due to the rising rate of failed clinical trials (Khan and Brown, 2001; Walsh et al., 2002). The training and expertise of clinical raters administering such assessments are critical factors influencing the reliability and validity of the measures obtained. Methodological problems such as functional unblinding of raters that may compromise randomization blinds (Greenberg et al., 1992) and inflation of baseline severity measures to meet study enrollment goals (DeBrota et al., 1999; Kobak et al., 2000) may contribute to current concerns that factors exogenous to the unbiased assessment of depression severity and treatment response may influence study results (Robinson and Rickels, 2000).

An alternative to the use of clinician assessments for measuring treatment outcomes is the use of patient self-reported measures of depression severity (Edwards et al., 1984). The use of computer technology to elicit self-report measures has been suggested as a possible means to address current problems in the conduct of randomized clinical trials (Greist et al., 2002). The procedural standardization of computer-based assessments may contribute to more reliable assessments, thus improving subject selection, promoting greater disclosure of personally sensitive information, and controlling clinician biases that may arise due to treatment

* Corresponding author. Tel.: +608 827 2452; fax: +608 827 2444.
*E-mail address:* Dkatzelnick@healthtechsys.com (D.J. Katzelnick).

0022-3956/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jpsychires.2005.05.002

unblinding or expectancy sets. Computer automated versions of the HAMD have been developed and validated for both desktop (Kobak et al., 1990) and interactive voice response (IVR) applications (Kobak et al., 2000). Paper-based self-report versions of the IDS and Quick Inventory of Depressive Symptomotology (QIDS) have been developed and validated (Trivedi et al., 2004), as has a version of the MADRS (Svanborg and Ashberg, 2001). Recent analyses suggest equivalence between clinician assessment of the HAMD and self-report measures of the QIDS obtained by paper methods or IVR technology (Rush, 2004). This report presents data from a study investigating the reliability and validity of an IVR version of the MADRS compared to concurrent clinician assessment of the same scale.

## 2. Methods

Sixty subjects (26 men and 34 women) aged 22–64 years (Mean = 42.7; SD = 10.6 years), were recruited through newspaper advertisements by the Department of Psychiatry at the University Health Network, Toronto, Canada. The sample was 80% Caucasian, and 74% had at least some college. Subjects who endorsed symptoms of depression during a brief telephone screen were invited to participate. They subsequently signed informed consent documents and were enrolled in the study. Study methods and materials were reviewed and approved by the University Health Network Research Ethics Board (Toronto, ON).

Subjects completed both the clinician-administered, face-to-face MADRS and the IVR self-report version of the MADRS in a counter-balanced order at the research office. For the IVR MADRS, patients began by providing an overall rating of their self-perceived severity for each of the 10 MADRS depression items (listed in Table 1) from 0 (no symptom present) to 6 (extremely severe). After providing this rating, the patients were presented with an appropriate anchoring description in a voice matched to the gender of the patient (i.e., women heard a female voice, men heard a male voice to facilitate personal identification) spoken with the affective intonation to convey the intended severity of the symptom being assessed. The patients were then asked whether his or her internal feeling state was less severe, equally severe, or more severe than the presented anchor (they were also allowed to listen to the anchor as many times as they wished). Patients indicating lesser (or greater) severity than the presented anchor were provided the next lower (or higher) descriptor and allowed to indicate the accuracy of that anchor for describing his or her feelings. Thus, regardless of the initial starting place, patients were allowed to 'titrate' up or down the severity scale until they felt the anchoring expression accurately reflected his or her own feelings, or until indicating a feeling state between two anchors. The IVR MADRS uses anchoring descriptions for scale severities of 0, 2, 4, and 6 (same as the original scale); patients indicating greater severity than one descriptor, and lesser severity than the next higher descriptor were assigned scale values of 1, 3, or 5. For example, severity scores for the symptom of "Reported Sadness" (Item 2) were anchored by "I haven't felt sad at all this past week, except when it was appropriate" (score = 0); "I feel a bit sad or low but I brighten up without difficulty" (score = 2); "I am thoroughly sad or gloomy, but things can make me feel a little bit better at times" (score = 4); "I am extremely sad and miserable all the time and cannot snap out of it at all" (score = 6).

After completion of both the clinician and IVR MADRS interviews, an IVR diagnostic interview (Mental Health Screener®) was administered.(Kobak et al., 1997) Clinicians also completed the Clinical Global Impression scale for severity (CGI-S), and patients completed the Patient version of the same scale (PGI-S) (Guy, 1976). Subjects were paid $50 for their participation. A

Table 1
Comparison of item scores (mean ± SD) and internal scale reliability for clinician and IVR administered versions of the Montgomery-Asberg depression rating scale

| Item | Clinician | IVR | Difference *t*-test (59 df) | Intraclass correlation (agreement) |
|---|---|---|---|---|
| Apparent sadness | 2.30 (1.18) | 2.23 (1.49) | −.348 $p = .729$ | .398 $p = .001$ |
| Reported sadness | 2.77 (1.21) | 3.12 (1.53) | 1.936 $p = .058$ | .469 $p < .001$ |
| Inner tension | 2.58 (1.33) | 2.70 (1.27) | .708 $p = .482$ | .520 $p < .001$ |
| Reduced sleep | 2.85 (1.89) | 2.75 (1.74) | −.629 $p = .532$ | .769 $p < .001$ |
| Reduced appetite | 1.68 (1.62) | 1.90 (1.66) | 1.635 $p = .107$ | .800 $p < .001$ |
| Concentration difficulties | 2.92 (1.36) | 2.90 (1.59) | −.089 $p = .930$ | .522 $p < .001$ |
| Lassitude | 2.83 (1.71) | 2.70 (1.58) | −.782 $p = .437$ | .679 $p < .001$ |
| Inability to feel | 2.80 (1.65) | 2.77 (1.59) | −.166 $p = .868$ | .545 $p < .001$ |
| Pessimistic thoughts | 2.63 (1.28) | 3.12 (1.64) | 2.659 $p = .010$ | .505 $p < .001$ |
| Suicidal thoughts | 1.13 (1.42) | 1.12 (1.61) | −.123 $p = .903$ | .764 $p < .001$ |
| Total score | 24.50 (9.09) | 25.30 (9.32) | 1.107 $p = .273$ | .815 $p < .001$ |
| Cronbach's Alpha (internal consistency) | .816 | .796 | | |

Sixty subjects were assessed by both methods on the same day in counter-balanced order.