

# Development of a standardized training program for the Hamilton Depression Scale using internet-based technologies: results from a pilot study

Kenneth A. Kobak<sup>a,b,\*</sup>, Joshua D. Lipsitz<sup>a,c</sup>, Alan Feiger<sup>a,d</sup>

<sup>a</sup>Research Training Associates, 7601 Ganser Way, Madison, WI 53719, USA

<sup>b</sup>Dean Foundation for Health Research and Education, USA

<sup>c</sup>Columbia University, USA

<sup>d</sup>Feiger Health Research Center, USA

Received 18 September 2002; received in revised form 15 April 2003; accepted 21 April 2003

## Abstract

Poor inter-rater reliability is a major concern, contributing to error variance, which decreases power and increases the risk for failed trials. This is particularly problematic with the Hamilton Depression Scale (HAMD), due to lack of standardized questions or explicit scoring procedures. Establishing standardized procedures for administering and scoring the HAMD is typically done at study initiation meetings. However, the format and time allotted is usually insufficient, and evaluation of the trainee's ability to actually conduct a clinical interview is limited. To address this problem, we developed a web-based, interactive rater education program for standardized training to diverse sites in multi-center trials. The program includes both didactic training on scoring conventions and live, remote observation of trainees applied skills. The program was pilot tested with nine raters from a single site. Results found a significant increase in didactic knowledge pre-to-post testing, with the mean number of incorrect answers decreasing from 6.5 (S.D. = 1.64) to 1.3 (S.D. = 1.03),  $t(5) = 7.35$ ,  $P = 0.001$  (20 item exam). Seventy-five percent of the trainees' interviews were within two points of the trainer's score. Inter-rater reliability (intraclass correlation) (based on trainees actual interviews) was 0.97,  $P < 0.0001$ . Results support the feasibility of this methodology for improving rater training. An NIMH funded study is currently underway examining this methodology in a multi-site trial.

© 2003 Elsevier Ltd. All rights reserved.

**Keywords:** Rater training; Assessment; Inter-rater reliability; Depression; Hamilton Depression Scale; Computerized assessment; Internet

## 1. Introduction

In controlled clinical trials, establishing adequate inter-rater reliability is a major concern. Poor inter-rater reliability increases measurement error, which in turn increases the chance for Type II error, i.e. failing to detect *true differences* between active drug and placebo. Additionally, poor reliability decreases statistical power, resulting in the need for larger sample sizes to detect significant differences between drug and placebo. For example, a study whose assessment reliability drops from 1.0 to 0.80 drops in power from 0.80 to 0.71, and requires 25% more subjects to detect a significant difference (Muller & Szegedi, 2002). Given the time and

financial resources involved in drug development, minimizing this source of error variance is of the utmost importance.

Establishing inter-rater reliability in multi-center clinical trials has been difficult for several reasons. Most anti-depressant clinical trials use as their primary outcome measure clinician-administered symptom rating scales whose ratings are based on clinical judgment. However, scales such as the Hamilton Depression Rating Scale (HAMD; Hamilton, 1960, 1967), provide only general guidelines for the administration and scoring of items. No standardized questions or explicit scoring algorithms are provided. As a result, raters vary widely on how they administer the scale and score the items.

In an attempt to solve this problem, several interview guides for the HAMD have been developed, containing more explicit scoring conventions and standardized probe questions (see Williams, 2001 for a review).

\* Corresponding author. Tel.: +1-608-829-1493; fax: +1-608-829-1411.

E-mail address: [kobak@charter.net](mailto:kobak@charter.net) (K.A. Kobak).

However, raters at different sites are often trained using different interview guides and scoring conventions, making inter-rater reliability across sites in multi-center trials difficult to achieve. For example, Grundy and colleagues (1994) found 10 distinct versions of the HAMD, seven derivative versions, and at least five alternative versions.

Industry has reacted to this challenge in a variety of ways. While some sponsors ignore the issue of reliability altogether, most attempt to achieve some training and reliability at start up meetings. These attempts have primarily focused on standardizing raters to the same set of scoring procedures and typically involve a review of scoring conventions and passive watching of videotapes. Bringing raters together for inter-rater reliability training is a costly process, and the format and time allotted at start up meetings do not allow for a comprehensive study of scoring conventions. Of more concern is that little or no effort is made to teach or evaluate the trainee's applied skills—i.e. their ability to conduct a clinical interview. In addition, inter-rater reliability based on trainees passively watching and rating a video produces artificially inflated estimates of reliability compared to trainees independently conducting interviews with the same patient. The former provides little information about how trainees will actually perform during the trial.

For the most part, rater training at startup meetings has proven insufficient to obtain adequate results. For example, in a recent inter-rater reliability training effort for a multi-site clinical trial, the difference in maximum and minimum total HAMD scores (full scale range 0–52) ( $N=86$  raters; 32 sites) evaluating the same subjects on videotaped presentations varied from a spread of 14 points in the best case, to a spread of 21 points in the worst case (Demitrack et al., 1997). The authors concluded that measurement error is large, and that “there was no evidence of improved rating performance across the 6 h of reliability training” (p. 20). Clearly, the development of better training methodologies is needed. This is especially true for the growing number of non-academic sites who are naive to clinical research, and may not have access to the extensive training needed to become proficient on specific rating scales.

### 1.1. Goal of the current study

The goal of this study was to develop a model for rater training designed particularly for raters participating in multicenter clinical trials. Our goal was to develop an educational program that will help raters become both highly competent (in both their didactic knowledge and applied skills) and reliable with each other. The challenge was to provide this interactive training from a centralized location to diverse sites. The use of web-based technologies allows for such an approach.

For purposes of this pilot program, we limited training to the basic 17 HAMD items, as recommended by Hamilton. We used a modified version of the Guy (1976) HAMD anchors (revised to include more specific descriptors of frequency and severity required for each anchor point, and to provide more objective behavioral descriptors for each severity level). These conventions have been used in several multi-center trials, and have shown to provide good inter-reliability (Feiger et al., 2001).

## 2. Methods

### 2.1. Features of the web-based HAMD training model

The web-based HAMD training program consists of two components: a didactic component and an applied component. The didactic component precedes the applied component in order to insure that trainees have an academic understanding of the rules and principles for administering and scoring the HAMD before attempting to apply these rules and procedures in a clinical interview (similar to taking your written driving test before getting behind the wheel). There is a testing and evaluation component of both the didactic and applied training, both during the learning process (to reinforce learning) and at the completion of each phase of the training.

#### 2.1.1. Didactic training

The didactic training component consisted of a *web-based interactive teaching tutorial* that contained the following components:

1. A review of general interviewing techniques and general scoring conventions for the HAMD;
2. A review of the item content for each of the 17 HAMD items;
3. A review of the scoring conventions for each of the 17 HAMD items, and
4. A final “self-exam” to evaluate trainees' didactic understanding of the above.

The tutorial takes advantage of the media to improve learning through features such as interactive testing and feedback throughout the tutorial for immediate reinforcement of learning, multi-modal learning (i.e. audio, video, and text), modeling of good interviewing skills through video examples, video examples of subjects displaying specific scores (anchor points), and video illustrations of scores on non-verbal behaviors (i.e. agitation and retardation). For example, each of the HAMD items begins with a review of the item content and item scoring, followed by a video vignette of a HAMD interview for that item. This is followed by

some commentary on scoring issues observed during the vignette, and observations on interview technique. Then the trainee is asked to rate the vignette, after which he/she is given immediate feedback and a rationale for the correct answer. After all the modules for all the items are reviewed, a final “self-exam” is given to assess and consolidate the trainee’s learning. The exam contains clinical vignettes illustrating certain issues in scoring conventions, and contains both multiple-choice and true–false response formats. The exam also samples content from the general interviewing principles. Trainees are given immediate feedback on their score and can review the entire exam (or just the answers they missed) and explanations for the correct answer. Results are immediately emailed to the trainer for review, and are also stored in a database accessible to the site administrator. Trainees could email questions on the material to the trainer at any time. Trainees could also print out the text of both the teaching modules and final exam for future reference. For trainees without high speed Internet access, the tutorial was available on CD-ROM. The didactic training takes about 2 h.

### 2.1.2. Applied training

As previously discussed, most pharmaceutical sponsored reliability training programs do not involve any instruction in or observation of a trainees applied skills. The applied training component of the training protocol involved *web-based videoconferencing* for

1. Live remote observation of trainees conducting HAMD interviews with feedback from the trainer on the trainee’s applied skills, and
2. Evaluation of inter-rater reliability (IRR).

With this method, trainees interviewed a patient (in our case, a “standardized patient”, e.g. a medically trained actor; see discussion) who was at the central site. The trainee and the trainer (Dr. Kobak) videoconferenced through their laptop computer and web video camera (in our case a Canon Hi 8). Using Netmeeting software, each user logs on, types in the web address, and enters the session information. Both sites can then see and hear each other through the laptop computer. The trainee interviews the subject over the Internet, with the trainer observing (making comments and suggestions either live during the interview or via discussion following the interview). As part of our pilot testing, we also had the trainee and subject together at the same location, with the trainer at a remote location observing and giving feedback. In our pilot test for feasibility, the trainee and the standardized patients were actually in different rooms of the same building. Live videostreaming service was provided by a commercial vendor (Vitalstream).

Our training program includes the use of a structured interview guide. We used the SIGH-D (Williams, 1988),

modified to include additional probes for more precisely determining frequency and severity. The use of a structured interview guide increases reliability by providing standardized probes [which reduce information variance (see below)], and helps insure all required domains are assessed.

### 2.1.3. Pilot study procedures

In order to evaluate the feasibility of the didactic portion of the training protocol, nine trainees (seven study coordinators and two psychiatrists) from a single site went through the tutorial. In order to do pre-and post-testing, two versions of the final exam were developed. Half were given version “A” at pre-test and half were given version “B”. This was done to obviate memory effects and to control for any possible differences in exam difficulty. After going through the tutorial, trainees received the alternative version of the exam that they took the first time. Trainees varied in number of years’ experience with the HAMD, from over 10 years’ experience to almost no experience. A score of 80% or greater (at least 16 of the 20 questions answered correctly) was decided a priori to be considered passing.

In order to evaluate the applied training component, each of the seven trainees interviewed four test patients (the two psychiatrists did not participate in this phase of the study). The trainees’ scores were then compared to scores given by the trainer, which served as the “gold standard”. Standardized patients from the University of Wisconsin Medical School were used for testing purposes. The standardized patients were given scripts to follow in order to remain consistent across trainees and were oriented to the scale and its scoring algorithm in order better accomplish this. In order to compare differences in methodology, three of the seven trainees were trained using the traditional method of in-person observation and feedback, while four of the seven were trained using remote observation and feedback over the Internet (all had previously taken and passed the web tutorial). It was decided a priori that the trainee’s score was considered “passing” if the total score was not more than two points different than the trainers score (Hamilton, 1967).

Finally, we evaluated inter-rater reliability, i.e. whether the training methodology produced raters who obtained similar scores when independently interviewing the same patient (in our case, a standardized patient). Traditional methods of determining IRR typically have trainees watch the same video and then compare ratings. This methodology is flawed, as two people passively watching a video done by a third person (usually an expert) artificially inflates inter-rater reliability by reducing the “information variance” that would occur if the two raters had to interview the patient independently (Spitzer & Williams, 1980; Spitzer et al., 1982). We determined IRR using videoconferencing, which

allows pairs of trainees at different locations to independently interview the same patient at a third (central) location for purposes of determining inter-rater reliability. Such a feature can be utilized in determining IRR between raters at two different sites in a multi-site trial. Inter-rater reliability was determined using the Intra-class Correlation Coefficient (ICC; Bartko, 1966). A total score not more than two points different from each other was chosen a priori as a secondary measure.

### 3. Results

#### 3.1. Didactic training on HAMD conventions

At baseline (prior to completing the tutorial), only one of the nine raters obtained a “passing” score of 80% on the didactic exam. At post-test, 100% passed. While not initially powered for inferential procedures, we also examined the change in the number of incorrect answers pre-and-post teaching. The mean number of incorrect answers (out of 21 items) at pretest was 6.5 wrong, (S.D. = 1.64, range 3–10). At post-test, the mean number of incorrect answers was 1.3 (S.D. = 1.03, range 0–3),  $t(5) = 7.35$ ,  $P = 0.001$ . A graphic illustration of score improvement is shown in Fig. 1.

#### 3.2. Applied training on trainees clinical interviewing skills

Seventy-five percent (21 of 28) of the interviews conducted by the trainees were within two points of the “gold standard” trainer score (see Fig. 2). This two-point standard was a high bar, considering that the methodology involved actively conducting independent interviews as opposed to simply rating a videotape. This figure is also impressive due to the fact that trainees received only one feedback session. Only one of the 28 scores was four points from the gold standard, with the rest of the scores within three points. Subjects trained

using remote feedback achieved equivalent scores to those who were trained in the traditional face-to-face method (i.e. four interviews were more than two points from the gold standard using traditional feedback, compared to three interviews for the Internet group; Table 1).

#### 3.3. Inter-rater reliability based on trainee conducted interviews

The inter-rater reliability coefficient (ICC) between trainees was 0.97,  $P < 0.0001$ . This compares favorably to the literature, where, for example, a correlation of 0.81 was found in the SIGH-D validation study (Williams, 1988).

### 4. Discussion

Results from this pilot study support the feasibility of this methodology in improving rater training. Given the increasing number of failed clinical trials, there is an urgent need in the field to improve clinical trial methodology and improve rater training, competence, and reliability. The novel technologies employed in the current work offer a new way of not only standardizing, but also improving rater training, thus improving rater competence and increasing inter-rater reliability across diverse settings. Given that most clinical trials involve multiple centers, such standardization is essential for decreasing error variance and reducing type II error. The three-stage model used in this study (didactic learning, applied learning, and testing of efficacy of the learning) may be a template for future training and testing, not only for the HAMD, but other clinician-administered scales as well. Ideally, this training should occur prior to the start of the clinical trial. Post-study monitoring of trainee’s interviews may also be valuable in order to reinforce and further learning, determine reliability, and to prevent rater drift. Digital audio recordings of study interviews can be immediately

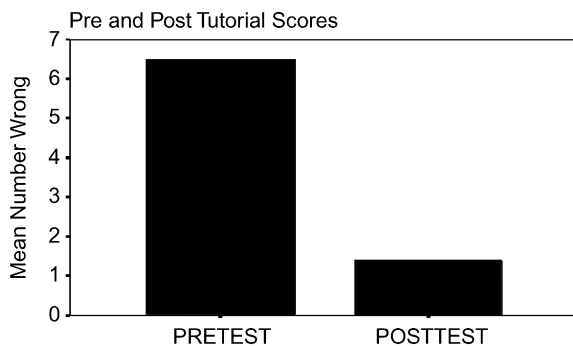


Fig. 1. Pre- and post-training test scores on didactic knowledge of HAMD scoring conventions (number incorrect). Note:  $N = 9$ ; pretest range 3–10, S.D. = 1.64; post-test range 0–3, S.D. = 1.03. Total number of test items = 21.

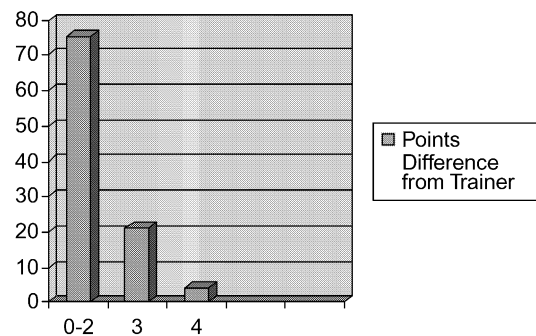


Fig. 2. Post-training evaluation of applied skills: percent of trainees that were within 0–2, 3, and 4 points from trainers total HAMD score (gold standard).

Table 1  
Results of applied training: trainee vs trainer score by patient<sup>a</sup>

	Trainee's HAMD Score	Trainers HAMD Score	Difference
<i>Patient A</i>			
Trainee 1	5	6	1
Trainee 2	5	6	1
Trainee 3	6	6	0
Trainee 4	6	6	0
Trainee 5	6	6	0
Trainee 6	8	6	2
Trainee 7	6	6	0
<i>Patient B</i>			
Trainee 1	12	14	2
Trainee 2	13	14	1
Trainee 3	11	14	3
Trainee 4	11	14	3
Trainee 5	11	14	3
Trainee 6	13	14	1
Trainee 7	13	14	1
<i>Patient C</i>			
Trainee 1	23	26	3
Trainee 2	25	26	1
Trainee 3	23	26	3
Trainee 4	26	26	0
Trainee 5	26	26	0
Trainee 6	24	26	2
Trainee 7	24	26	2
<i>Patient D</i>			
Trainee 1	12	14	2
Trainee 2	11	14	3
Trainee 3	13	14	1
Trainee 4	13	14	1
Trainee 5	10	14	4
Trainee 6	12	14	2
Trainee 7	14	14	0

<sup>a</sup> Trainees 4–7 trained using videoconferencing, trainees 1–3 trained using traditional in person observation.

emailed to the sponsor for review and feedback. Recent collaborative, industry-wide efforts at standardizing the administration and scoring of the HAMD should also help improve reliability (Williams et al., 2002), and provide a common framework for training methodologies. Ideally, some type of industry-wide rater certification should be implemented, so raters are trained and certified across trials and sponsors.

Web-based technologies have several unique features that offer advantages over traditional training methods, including:

1. Providing tools (such as interactive video) that increases the quality of the training program and improves retention of knowledge through immediate reinforcement of learning;
2. Standardizing the training protocol between sites in order to insure each trainee receives similar information, thus improving reliability;

3. Standardizing the quality of the information provided;
4. Enabling remote training of multiple sites from a centralized location, as well as remote evaluation of subjects and trainees; and
5. Providing rater training that is easily accessible, cost effective, and more easily and widely disseminated.

One concern relevant to this study is whether interviews administered remotely via teleconference or videostreaming will obtain results that differ from the same interview administered face-to-face. The literature on telephone versus in-person interviews provides some support for this. Simon et al. (1993) compared telephone to face-to-face administration of the HAMD in 30 outpatients beginning antidepressant treatment. They found the two modes of administration produced nearly identical results, with only 0.2 points difference between the methods, and an intraclass correlation of 0.80. Similar findings have been found with other clinician-administered rating scales, e.g. SCID and SCL-90 (Simon et al., 1993), Center for Epidemiologic Studies Depression Scale (CES-D) (Aneshensel et al, 1982), and the KIDDIE-SADS and Personality Disorders Examination (Rohde et al., 1997). Currently, the NIMH sponsored study MH400498 (Sequenced Treatment Alternatives to Relieve Depression; STAR-D) utilizes HAMD interviews conducted by telephone from a central location as the primary outcome measure. It seems reasonable to expect that live video can achieve similar results as found with telephone administration. In the one study using TeleVideo (Stevens et al., 1999), patients randomly assigned to face-to-face or TeleVideo unstructured psychiatric assessments gave both interviews similarly high ratings satisfaction and ability to develop rapport, providing some indication that patients will accept this method of assessment.

With that said, it is fair to point out some of the study limitations. The major study limitation is that of insufficient power. With only nine subjects, there is a risk of making a Type I error due to small sample size. Therefore, the results should be interpreted with caution, and treated only as preliminary data. In addition, the sample was not randomly selected.

There are also some practical limitations with the technology. In order to view the video vignettes on the web tutorial one must have a high-speed Internet connection (e.g. cable modem, ISDN, DSL, fractional T1, etc). We solved this problem by creating a CD-ROM version of the tutorial. The CD-ROM version is identical to the web version in content, and could be used as a stand-alone alternative to the web. Videoconferencing may also be affected by audio and video latency of up to 1 s introduced by heavy traffic on the Internet between the test sites. We did not find this to be much of a

problem. The applied portion is fairly labor intensive, involving the time of both an expert teacher as well as a standardized or actual patient. However, all good training is labor intensive, as it involves practice over time, with individualized feedback based on ongoing observation. This is true whether the training is conducted remotely or in-person.

All the trainees in this study were from the same site. As this was a feasibility study, we simulated “real world” conditions by having trainees in remote rooms during videoconferencing. Whether similar results are found in real world conditions remains to be determined. The small sample of trainees used to test feasibility limits the generalizability of the findings. A multi-center clinical trial is currently underway that should help address this issue.

Standardized patients were used to train and test raters. This may have inflated the correlation coefficients, as the standardized patients were all trained to portray the same symptoms to each trainee. Real patients may have greater variability when interviewed on repeated occasions. Studies on standardized patients have been found they achieve high level of stability for inter-rater reliability purposes in the assessment of depression (Badger et al., 1995). On the other hand, several studies have demonstrated that trainee competence as evaluated with standardized patients is a good measure of trainee competence with actual patients (Pieters et al., 1994; De Champlain et al., 1997; Colliver & Swartz, 1997; Peabody et al., 2000). A consensus conference on the use of standardized patients convened by the Association of American Medical Colleges in December 1992 reported that experienced physicians were unable to differentiate standardized patients from real patients when sent unannounced into a physicians office, even when the physician was told in advance that this would be occurring (Proceedings of the AAMC’s consensus conference, 1993; Annex to the proceedings of the AAMC consensus conference, 1994). Standardized patients allow trainers to focus on certain points that are especially critical for the trainees to learn. Use of professionally trained actors also has the advantage of eliminating the error variance that is introduced in inter-rater reliability due to changes in the patient between the first and second interview. They also provide added convenience for training and testing purposes.

Overall, results support the use of this technology in improving rater competence and establishing good inter-rater reliability. The Internet provides the means to deliver more meaningful training in an efficient and user friendly manner, and allows for training that includes a focus on applied skills, and not just passive observation of videotapes. Reviews and recalibrations can be easily repeated in the course of a trial to help prevent rater drift. We are currently studying the effi-

cacy of this training methodology in a multicenter, NIMH funded trial, using both audio and videoconferencing, and both real and standardized patients. This will allow a more rigorous test of the feasibility and efficacy of this methodology across diverse sites and backgrounds.

## Acknowledgements

The authors gratefully acknowledge the contributions of Glenn Chung at Visuality, who helped with web design, Dawn Sikich, who conducted the training interview for the website, and the raters at the Dean Foundation for pilot testing the program.

This study was supported by NIMH SBIR contract #N43MH12049 awarded to Kenneth A. Kobak, PhD.

## References

- Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Measuring depression in the community: a comparison of telephone and personal interviews. *Public Opinion Quarterly* 1982;46:110–21.
- Annex to the proceedings of the AAMC consensus conference on the use of standardized patients in the teaching and evaluation of clinical skills. *Teach Learn Med*, 6, 2–35.
- Badger LW, deGruy F, Hartman J, Plant MA, Leeper J, Ficken R, Templeto B, Nutt L. Stability of standardized patient’s performance in a study of clinical decision making. *Family Medicine* 1995;27:126–31.
- Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 1966;19:3–11.
- Colliver J, Swartz MH. Assessing clinical performance with standardized patients. *JAMA* 1997;278:790–1.
- De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patient’s accuracy in recording examinees’ behaviors using checklists. *Academic Medicine* 1997;72:S85–0S87.
- Demitrack MA, Faries D, DeBrotta D, Potter WZ. The problem of measurement error in multisite clinical trials. *Psychopharmacology Bulletin* 1997;33:513.
- Feiger AD, Lipsitz JD, Kobak KA, Evans KR, Sills T. (May, 2001). Impact of a Comprehensive HAMD Inter-Rater Reliability Training in a Multi-Site Trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 41st Annual Meeting, Phoenix, AZ.
- Grundy CT, Lunnen KM, Lambert MJ, Ashton JE, Tovey DR. The Hamilton Rating Scale for Depression: one scale or many? *Clinical Psychology: Science and Practice* 1994;1:197–205.
- Guy, W. 1976. ECDEU Assessment Manual for Psychopharmacology, revised. Rockville, MD, National Institute of Mental Health, US Department of Health, Education, and Welfare publication ADM 76\_338.
- Hamilton M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960;23:56–62.
- Hamilton M. Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology* 1967;6:278–96.
- Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *Journal of Clinical Psychopharmacology* 2002;22(3):318–25.
- Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction. *JAMA* 2000;283:1715–22.
- Pieters HM, Touw-Otten FW, De Melker RA. Simulated patients in

- assessing consultation skills of trainees in general practice vocational training: a validation study. *Medical Education* 1994;28:226–33.
- Proceedings of the AAMC's consensus conference on the use of standardized patients in the teaching and evaluation of clinical skills. *Acad Med*, 68, 437–488.
- Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing axis I and axis II disorders. *American Journal of Psychiatry* 1997;154:1593–8.
- Simon GE, Revicki D, Von Korff M. Telephone assessment of depression severity. *Journal of Psychiatric Research* 1993;27:247–52.
- Spitzer RL, Skodol AE, Williams JBW, Gibbon M, Kass F. Supervising intake diagnosis. A psychiatric 'rashomon'. *Archives of General Psychiatry* 1982;39:1299–305.
- Spitzer RL, Williams JBW. Classification in Psychiatry: Classification of Mental Disorders and DSM-III. In: Kaplan HI, Freedman AM, Sadock BJ, editors. *Comprehensive Psychiaitry*. Baltimore: Williams & Wilkins; 1980.
- Stevens A, Doidge N, Goldbloom D, Voore P, Farewell J. Pilot study of TeleVideo psychiatric assessments in an under serviced community. *American Journal of Psychiatry* 1999;156:783–5.
- Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 1988;45:742–7.
- Williams JBW. Standardizing the Hamilton Depression Rating Scale: Past, present, and future. *European Archives of Psychiatry and Clinical Neuroscience* 2001;251(2):II/6–0II/12.
- Williams JBW, Kobak KA, Kalali A, Lipsitz JD, Englehart N, Evans K, Olin J, Pearson J, Rothman M, Bech P. Using the new GRID HAMD: Results from pilot testing. National Institute of Mental Health. Boca Raton, FL: New Clinical Drug Evaluation Unit, 42nd Annual Meeting; 2002.