

Assessing Depression in Primary Care with the PHQ-9: Can It Be Carried Out over the Telephone?

Alejandra Pinto-Meza, PhD,¹ Antoni Serrano-Blanco, MD,¹ Maria T. Peñarrubia, MD,²
Elena Blanco, MD,² Josep Maria Haro, PhD¹

¹Research and Development Unit, Sant Joan de Déu-SSM, Barcelona, Spain; ²Primary Care Health Center Gavà II, Costa de Ponent's Teaching Unit, Catalan Health Service, Catalonia, Spain.

BACKGROUND: Telephone assessment of depression for research purposes is increasingly being used. The Patient Health Questionnaire 9-item depression module (PHQ-9) is a well-validated, brief, self-reported, diagnostic, and severity measure of depression designed for use in primary care (PC). To our knowledge, there are no available data regarding its validity when administered over the telephone.

OBJECTIVE: The aims of the present study were to evaluate agreement between self-administered and telephone-administered PHQ-9, to investigate possible systematic bias, and to evaluate the internal consistency of the telephone-administered PHQ-9.

METHODS: Three hundred and forty-six participants from two PC centers were assessed twice with the PHQ-9. Participants were divided into 4 groups according to administration procedure order and administration procedure of the PHQ-9: Self-administered/Telephone-administered; Telephone-administered/Self-administered; Telephone-administered/Telephone-administered; and Self-administered/Self-administered. The first 2 groups served for analyzing the *procedural validity* of telephone-administered PHQ-9. The last 2 allowed a test-retest reliability analysis of both self- and telephone-administered PHQ-9. Intraclass correlation coefficient (ICC) and weighted κ (for each item) were calculated as measures of concordance. Additionally, Pearson's correlation coefficient, Student's *t*-test, and Cronbach's α were analyzed.

RESULTS: Intraclass correlation coefficient and weighted κ between both administration procedures were excellent, revealing a strong concordance between telephone- and self-administered PHQ-9. A small and clinically nonsignificant tendency was observed toward lower scores for the telephone-administered PHQ-9. The internal consistency of the telephone-administered PHQ-9 was high and close to the self-administered one.

CONCLUSIONS: Telephone and in-person assessments by means of the PHQ-9 yield similar results. Thus, telephone administration of the PHQ-9 seems to be a reliable procedure for assessing depression in PC.

KEY WORDS: PHQ-9; telephone assessment; depression; primary care.
DOI: 10.1111/j.1525-1497.2005.0144.x
J GEN INTERN MED 2005; 20:738-742.

Depression is one of the most prevalent mental disorders¹ and between one-fourth and one-half of patients with depression are treated at primary care (PC) centers (PCC).^{2,3} There are a number of case-finding instruments for detecting depression in PC, ranging from 2 to 30 items in length, and typically scored as continuous measures of depression severity with established cut-off points above which the probability of major depression is substantially increased. Most of these instruments show good validity.⁴⁻⁶

The Patient Health Questionnaire (PHQ)⁷ was designed for use in PC and to provide a brief self-report diagnostic instrument for the diagnosis of mental disorders using criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision (DSM-IV-TR).⁸ Validation data were published for the original English⁷ and for other language versions (e.g., Spanish,⁹ German,¹⁰ and Arabian¹¹). Its 9-item depression module, the PHQ-9, is a brief self-reported diagnostic and severity measure of depression. Several studies support its validity, feasibility, and its capacity to detect changes of depressive symptoms over time.¹²⁻¹⁴ Additionally, the PHQ-9 is increasingly being used in research and clinical practice, and has demonstrated superior criterion validity with respect to the diagnosis of major depression compared with other established depression-screening questionnaires.¹⁵ Recently, a Spanish version of the PHQ-9 has been shown to be a feasible screen tool for depression.¹³

Kroenke et al.¹² pointed out, regarding PHQ-9, that "having a simple self-administered measure to complete either in the clinic or by telephone administration (e.g., nurse administration or interactive voice recording) would save clinicians the time needed to enquire about the presence and severity of each of the 9 DSM-IV symptoms to assess outcomes" (p. 612). However, to our knowledge, the *procedural validity* of telephone administration of the PHQ-9 has not yet been established. That is, we do not know to which extent telephone-administered PHQ-9 could produce results similar to self-administered PHQ-9. Procedural validity refers to the degree to which a new procedure (e.g., telephone administration) offers results similar to those obtained through a well-known procedure (e.g., self-administration) that is used as a *gold standard*. Thus, procedural validity refers only to the validity of an assessment procedure, not to the validity of the instrument itself.¹⁶

Telephone assessment of depression for research purposes is increasingly being used. To date, scales such as the Hamilton Depression Rating Scale (HDRS) or the Center for Epidemiologic Studies-Depression Scale (CED-D) have been validated for telephone administration.^{17,18} The relative low cost and ease of administration of telephone interviews for assessing depression could enable larger sample sizes to be obtained and facilitate studies over wider geographical areas or with more follow-up ratings. In order to investigate the procedural validity of the telephone-administered PHQ-9, we compared agreement between self-administered and telephone-administered responses to the PHQ-9. Additionally, we compared mean scores between self-administered and

The authors have no conflicts of interest to report.

Address correspondence and requests for reprints to Dr. Pinto-Meza: Carrer Dr. Antoni Pujades, 42, 08830 Sant Boi de Llobregat, Barcelona, Spain (e-mail: apinto@sjd-ssm.com).

Received for publication January 19, 2005

and in revised form January 29, 2005

Accepted for publication February 3, 2005

telephone-administered PHQ-9 in order to investigate systematic bias; analyzed and compared telephone- and self-administered PHQ-9 test-retest reliability; and evaluated the internal consistency of the telephone-administered PHQ-9.

METHODS

Participants

Participants were selected by two PC physicians (PCP) and assessed twice (within a 7-day period) by a clinical psychologist among persons seeking medical assistance ($n=289$) or working at two PCC in Barcelona, Spain ($n=57$). All signed an informed consent. Three hundred and forty-six out of 375 selected participants (aged 18 to 75 years old) were included. Sixteen persons were excluded because they were younger than 18 or older than 75 years old, and 13 were excluded because they did not answer to the second administration of the PHQ-9.

Measurements

The PHQ-9 is a 9-item self-reported questionnaire designed to evaluate the presence of depressive symptoms during the prior 2 weeks. As a severity measure, scores can range from 0 (absence of depressive symptoms) to 27 (severe depressive symptoms). Each of the 9 items, asking for each of the DSM-IV diagnostic criteria, can be scored from 0 (not at all) to 3 (nearly every day). As a diagnostic measure, major depression is diagnosed if 5 or more of the 9 depressive symptom criteria have been present at least "more than half the days" (a score of 2) in the past 2 weeks, and one of the symptoms is depressed mood or anhedonia. Also, before making a final diagnosis, the clinician is expected to rule out physical causes of depression, normal bereavement, and history of manic episode.¹²

Procedure

Participants were divided into 4 groups according to the PHQ-9 administration procedure order and administration procedure: Self-administered/Telephone-administered (ST), Telephone-administered/Self-administered (TS), Telephone-administered/Telephone-administered (TT), and Self-administered/Self-administered (SS). The first two groups served for analyzing the procedural validity of telephone-administered PHQ-9. The last 2 allowed a test-retest reliability analysis of both self- and telephone-administered PHQ-9. Participants were recruited until the intended quota for each group was achieved. Thus, first the ST group was completed, and then the TS, the TT, and finally the SS one. For logistic reasons (risk of high attrition in 2 consecutive visits), while participants for the ST, TS, and TT groups were selected among patients seeking medical assistance at the PCC ($n=289$), participants for the SS group were recruited among PCC staff members ($n=57$).

Selected participants for the ST group first answered self-administered PHQ-9. Within the subsequent 7 days, PHQ-9 was administered over the telephone.

Selected participants for the TS group were first assessed by telephone. In order to increase the probability of answering to the second PHQ-9 administration, selected participants had a prescheduled consultation with the PCP within the 7 days following telephone assessment. Thus, after the telephone evaluation, participants answered the self-administered

PHQ-9 (after the PCP consultation). As the PHQ-9 is a self-administered questionnaire, the evaluator could not interfere with participant's answers.

Selected participants for the TT group were assessed twice by telephone within a 7-day period. To reduce possible interferences with answers to the second telephone interview, the evaluator did not have access to prior results at the time of the second telephone assessment.

Finally, selected participants for the SS group answered twice, within a 7-day period, the self-administered PHQ-9. Participants did not have access to prior answers at the time of second assessment. As mentioned before, in order to increase the probability of answering to the second PHQ-9 administration, participants for the SS group were recruited among PCC staff.

Considering that the instructions of the self-administered version of the PHQ-9 were not suitable for telephone administration, telephone-administered PHQ-9 instructions were modified as follows: "I'm going to ask you about several problems. Please tell me how often you have been bothered, over the last 2 weeks, by any of the following problems. For each problem there are 4 possible answers: Not at all; Several days; More than half of the days, or Nearly every day." Possible answers were repeated as needed after each of the nine items.

Statistical Analysis

Differences between individual scores using the two assessment procedures were analyzed with the Intraclass Correlation Coefficient (ICC), using both consistency and absolute agreement indexes, and the weighted κ statistic (for items analysis). Weights used were 1.00; 0.66; and 0.00 or 1.00; 0.50; and 0.00. On one hand, it has been established that ICC values between 0.41 and 0.75 reveal moderate-to-good agreement, and ICC values over 0.75 reveal excellent agreement. On the other hand, κ values between 0.41 and 0.60 reveal moderate agreement, and κ values over 0.60 show good agreement.¹⁹ Paired t -tests were used to investigate differences in mean scores between telephone- and self-administered assessments. Finally, the internal consistency of the telephone-administered PHQ-9 was evaluated using Cronbach's α .

RESULTS

The socio-demographic characteristics of participants are summarized in Table 1. χ^2 tests and ANOVA (for age) revealed that the groups were different in terms of socio-demographic characteristics. However, when comparing socio-demographic characteristics between the ST and TS groups (i.e., groups used to assess procedural validity), there were no statistically significant differences in terms of gender, marital status, or employment status. Only years of formal education ($P=.006$) and age ($P=.016$) were significantly different between the ST and TS groups. Participants from the ST group were more educated and were younger than those from the TS group.

Intraclass correlation coefficient between each pair of administrations (Table 2) were high either for the procedural validity groups (i.e., ST and TS, or ST plus TS) or for the test-retest reliability groups (i.e., TT and SS). Weighted κ coefficients for each PHQ-9 item (Table 3) for the procedural validity groups were mostly higher than 0.60. Only item 1 (TS group) showed a low κ value. Additionally, when comparing self-administered

Table 1. Socio-Demographic Characteristics of Participants by Groups

| | Groups Formed According to Administration Procedure Order and Administration Procedure | | | | |
|--------------------------------------|--|---------------|---------------|---------------|---------------|
| | ST (n=118) | TS (n=113) | TT (n=58) | SS (n=57) | Total (n=346) |
| Gender* | | | | | |
| Female (%) | 69.5 | 63.7 | 84.5 | 66.7 | 69.7 |
| Age (y)* | | | | | |
| Mean (SD) | 48.01 (16.35) | 53.07 (15.32) | 50.84 (13.23) | 38.70 (13.14) | 48.60 (15.73) |
| Median | 47.00 | 57.00 | 49.00 | 37.00 | 48.00 |
| Marital status (%)* | | | | | |
| Married/living with a partner | 71.2 | 70.9 | 72.4 | 63.2 | 70.0 |
| Never married | 15.3 | 10.6 | 8.6 | 31.6 | 15.3 |
| Divorced/separated/widowed | 13.5 | 18.5 | 19.0 | 5.2 | 14.7 |
| Employment status (%)* | | | | | |
| Working | 33.1 | 31.0 | 44.8 | 80.7 | 42.2 |
| Not working (sick leave, unemployed) | 15.3 | 13.3 | 5.2 | 3.6 | 11.0 |
| Homemaker | 28.0 | 41.5 | 34.5 | 7.0 | 30.1 |
| Other (student, retired) | 23.6 | 14.2 | 15.5 | 8.7 | 16.7 |
| Years of formal education (%)* | | | | | |
| 0-4 | 27.6 | 28.3 | 29.3 | 1.8 | 23.8 |
| 5-8 | 19.0 | 37.2 | 39.7 | 17.5 | 28.3 |
| 9-12 | 27.5 | 21.2 | 15.5 | 12.3 | 20.9 |
| More than 12 | 25.9 | 13.3 | 15.5 | 68.4 | 27.0 |

χ^2 or ANOVA comparisons among groups.

* $P < .05$.

ST, self-administered/telephone-administered PHQ-9; TS, telephone-administered/self-administered PHQ-9; TT, telephone-administered/telephone-administered PHQ-9; SS, self-administered/self-administered PHQ-9; PHQ-9, Patient Health Questionnaire-9.

and telephone-administered PHQ-9 items, regardless of administration procedure order (ST plus TS group), none of the PHQ-9 items showed a κ value below 0.58. The same was observed for the test-retest reliability groups. Weighted κ coefficients for each PHQ-9 item were higher than 0.60.

PHQ-9 mean scores, correlations, and mean differences between each pair of assessments are summarized in Table 4. Correlations between each pair of assessments were large and highly statistically significant ($P < .001$ for all). Comparisons of mean ratings revealed differences for the ST and TT pairs. When comparing differences between self-administered and telephone-administered PHQ-9 for the ST plus TS group, that is, comparing self- and telephone-administered PHQ-9 regardless of administration procedure order, differences emerged between the self- and telephone-administered PHQ-9. Thus, participants showed slightly lower mean scores (0.60 points) in the telephone-administered PHQ-9 when compared with the self-administered PHQ-9.

Table 2. PHQ-9 ICC for Each Group Formed According to Administration Procedure Order and Administration Procedure

| | PHQ-9 | PHQ-9 |
|---------------|-------------------|--------------------------|
| | ICC (Consistency) | ICC (Absolute Agreement) |
| ST (n=118) | 0.94 | 0.93 |
| TS (n=113) | 0.91 | 0.91 |
| ST+TS (n=231) | 0.92 | 0.92 |
| TT (n=58) | 0.93 | 0.92 |
| SS (n=57) | 0.92 | 0.92 |

ST, self-administered/telephone-administered; TS, telephone-administered/self-administered; ST+TS, sum of the ST and TS groups; TT, telephone-administered/telephone-administered; SS, self-administered/self-administered; ICC, intraclass correlation coefficient; PHQ-9, Patient Health Questionnaire-9.

The internal consistency of telephone-administered PHQ-9 was 0.82 (a total of 289 participants who answered to the telephone-administered PHQ-9 were included in the analysis. If it was a TT participant, only the first assessment was considered). Self-administered PHQ-9 showed an internal consistency of 0.86 (a total of 288 participants who answered to the self-administered PHQ-9 were included in the analysis. If it was a SS participant, only the first assessment was considered).

DISCUSSION

On the basis of the results presented here, it is possible to conclude that telephone and in-person assessment, by means of the PHQ-9, yield similar results. Additionally, the internal consistency of the telephone-administered PHQ-9 was similar to the self-administered PHQ-9. Thus, telephone administration of the PHQ-9 seems to be a reliable procedure for assessing depression at PC.

In the present study, questionnaire items were identical but the administration procedure differed. According to Helzer et al.²⁰ we can describe the present study as a study of procedural validity, having characteristics of both reliability, because the same measure was used twice, and validity, since telephone administration was compared with a gold standard (self administration).

Intraclass correlation coefficient between self-administered and telephone-administered PHQ-9 were excellent regardless of administration procedure order (ST or TS) or administration procedure (telephone- or self-administration). Moreover, item concordance analysis (weighted κ) of each group revealed good or moderate agreement for all items, showing an adequate procedural validity for the telephone-administered PHQ-9 and good test-retest reliability for both self- and telephone-administered PHQ-9. Additionally, the internal consistency of telephone-administered items was high

Table 3. κ Coefficient for Each PHQ-9 Item by Groups Formed According to Administration Procedure Order and Administration Procedure

| PHQ-9 Items | ST (n=118) | TS (n=113) | ST+TS (n=231) | TT (n=58) | SS (n=57) |
|--|------------|------------|---------------|-----------|-----------|
| PHQ1 (Little interest or pleasure in doing things) | 0.71 | 0.48 | 0.60 | 0.83 | 0.70 |
| PHQ2 (Feeling down, depressed, or hopeless) | 0.78 | 0.68 | 0.73 | 0.78 | 0.83 |
| PHQ3 (Trouble falling or staying asleep, or sleeping too much) | 0.69 | 0.74 | 0.72 | 0.89 | 0.61 |
| PHQ4 (Feeling tired or having little energy) | 0.64 | 0.68 | 0.66 | 0.70 | 0.59 |
| PHQ5 (Poor appetite or overeating) | 0.60 | 0.58 | 0.59 | 0.73 | 0.74 |
| PHQ6 (Feeling bad about yourself, or that you are a failure or have let yourself or your family down) | 0.69 | 0.70 | 0.70 | 0.68 | 0.67 |
| PHQ7 (Trouble concentrating on things, such as reading the newspaper or watching television) | 0.63 | 0.56 | 0.60 | 0.80 | 0.79 |
| PHQ8 (Moving or speaking so slowly that other people could have noticed, or the opposite, being so fidgety or restless that you have been moving around a lot more than usual) | 0.63 | 0.52 | 0.58 | 0.72 | 0.63* |
| PHQ9 (Thoughts that you would be better off dead or hurting yourself in some way) | 0.68 | 0.77 | 0.72 | 0.77* | 0.57* |

ST, self-administered/telephone-administered; TS, telephone-administered/self-administered; ST+TS, sum of the ST and TS groups; TT, telephone-administered/telephone-administered; SS, self-administered/self-administered; PHQ-9, Patient Health Questionnaire-9. Weights used for the κ analysis were: 1.0; 0.66; 0.33; and 0.0.

*Except for these items, where weights were: 1.0; 0.50, and 0.0.

(between 0.85 and 0.79 depending on the group) and very close to the self-administered items.

A high and significant positive correlation was observed between self-administered and telephone-administered PHQ-9. Furthermore, correlation between both procedures was even higher than the one obtained by Kroenke et al.¹² in their validation study of the PHQ-9 as a depression severity measure. While in that study, correlation between self-administered and telephone reappraisal performed within 48 hours was 0.84, in the present study, no group (i.e., ST, TS, TT, or SS) showed a correlation below 0.90.

Table 4. PHQ-9 Mean Scores, Correlations Between Ratings, and Mean Differences by Groups Formed According to Administration Procedure Order and Administration Procedure

| Groups Formed According to PHQ-9 Administration Procedure Order and Administration Procedure | PHQ-9 | |
|--|---------------|-----------------------------------|
| | Mean (SD) | Pearson's Correlation Coefficient |
| ST (n=118) | | |
| Self-administered (SD) | 6.19 (5.44) | 0.95** |
| Telephone-administered | 5.13 (5.08) | |
| Mean difference: S-T | 1.06** (1.76) | |
| TS (n=113) | | |
| Telephone-administered | 5.62 (5.62) | 0.90** |
| Self-administered | 5.75 (5.67) | |
| Mean difference: T-A | -0.12 (2.43) | |
| ST+TS (n=231) | | |
| Self-administered | 5.97 (5.55) | 0.92** |
| Telephone-administered | 5.37 (5.35) | |
| Mean difference: A-T | 0.60** (2.16) | |
| TT (n=58) | | |
| 1st Telephone-administered | 4.96 (4.90) | 0.93** |
| 2nd Telephone-administered | 4.37 (4.65) | |
| Mean difference: T-T | 0.59* (1.79) | |
| SS (n=57) | | |
| 1st Self-administered | 3.91 (3.86) | 0.92** |
| 2nd Self-administered | 3.61 (3.75) | |
| Mean difference: S-S | 0.30 (1.51) | |

* $P < .05$; ** $P < .001$.

ST, self-administered/telephone-administered; TS, telephone-administered/self-administered; ST+TS, sum of the ST and TS groups; TT, telephone-administered/telephone-administered; SS, self-administered/self-administered; SD, standard deviation; PHQ-9, Patient Health Questionnaire-9.

PHQ-9 mean comparisons revealed a significant tendency toward lower scores for the telephone administration. However, the differences were minor (0.60 points) and probably lacked clinical relevance. In fact, according to Kroenke et al.¹² depression severity measured with the PHQ-9 is considered to change qualitatively every 5 points. Thus, we should be cautious in overstating this point.

Although both procedures considered the same questions (items), it may have been possible that answering individually (i.e., self-administered PHQ-9) may have enhanced personal acknowledgment of certain characteristics that, when answering to someone else (over the telephone), could have been inhibited, either because of distrust or lack of privacy. As Evans et al.²¹ pointed out, "it is less easy to ensure privacy in a telephone interview, because the interviewer does not know who else may be present, possibly inhibiting disclosure by the subject" (p. 161). In the same way, Rohde et al.²² suggested that when scheduling a telephone assessment, the interviewer should try to set up a time when the participant could talk in private.

Additionally, PHQ-9 mean comparisons also revealed a statistically significant tendency toward lower scores on reappraisal assessments, that is, participants showed lower scores on the second PHQ-9 assessment for the ST plus TS group, and for the TT group. Two studies comparing face-to-face and telephone interviews found the same tendency.^{22,23} According to Jorm et al.²⁴ when assessing psychiatric symptoms or personality traits twice, a mean change in scores toward less psychopathology is often observed. This retest artifact does not seem to be related to time lag between occasions and confined to measures assessing negative self-characteristics and administered orally by an interviewer. Some hypotheses intending to explain this are as follows: (1) regression to the mean, (2) therapeutic effects of the first interview, (3) participants trying to create a more favorable impression on retest, or (4) respondents taking the second evaluation less seriously. Any of these hypotheses are plausible for the present study. Unfortunately, our results do not allow us to clarify this point.

Limitations of the present study and of telephone interviewing must be acknowledged. First, participants were not randomized to the 4 groups, and while PCC patients formed the ST, TS, and TT groups, PCC staff members formed the SS one. This may explain the differences in socio-demographic

characteristics among groups. For example, individuals in the SS group were younger, more educated, and most of them were currently working. However, because the analyses were conducted within groups, we believe that these differences do not represent a major methodological concern. Besides, ST and TS groups (those directly related to procedural validity testing) were more similar as they only differed in terms of mean age and years of formal education.

Second, differences emerging from age, educational level, or gender variations were not considered when comparing telephone- and self-administered PHQ-9 because of the small sample size for each socio-demographic category. It could be possible that telephone- and self-administered PHQ-9 could show more or less an agreement according to such differences, and therefore the telephone-administration procedure could be less valid for certain populations. For example, it was stated that telephone responses from older people might be different from face-to-face assessments for the General Health Questionnaire.²¹

Third, as indicated by the low mean PHQ-9 scores (between 3.61 and 6.19), our sample included only a few participants with high levels of depression severity. Therefore, our results might not be representative for patient samples with higher levels of depression severity or samples with a wider range of depressive severity. This potential "bottom effect" may limit generalization of our findings.

Fourth, the brief time interval considered between assessments could have favored recall of initial answers. However, these conditions could also represent an advantage, as a brief time interval could reduce possible changes within subjects.

Fifth, during telephone interview, it is less easy to ensure privacy, because the interviewer does not know who else may be present, possibly inhibiting disclosure by the participant. The importance of developing a rapport between the interviewer and the participant before gathering sensitive information has been pointed out.²² This could be less easy to do over the telephone. In the present study, the PCP requested their patients to participate as a way of favoring confidence. In any case, we do not know the extent to which this was achieved.

Finally, during telephone interview, we have to be aware that we could be selectively excluding participants not having a telephone and therefore biasing our results.

Future research concerning agreement between responses to self-administered and telephone-administered PHQ-9 or other scales could attempt to explore possible differences emerging from gender, educational level, or age. Additionally, reassuring the respondent regarding privacy as much as possible may favor the validity of the assessment.

This study was supported by a grant from the Catalan Agency for Health Technology Assessment and Research (063/26/2000 and it is part of the IRYSS network (FIS G03/202)). The authors wish to express their gratitude to David Suarez for his assistance with data analyses.

REFERENCES

1. **World Mental Health Survey Consortium.** Prevalence, severity, and unmet need for treatment of mental disorders in the World Health

- Organization World Mental Health Survey. *JAMA*. 2004;291:2581-90.
2. **Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK.** The de facto US mental and addictive disorders service system. *Arch Gen Psychiatr*. 1993;50:85-94.
3. **Narrow WE, Regier DA, Rae DS, Manderscheid RW, Locke BZ.** Use of services by persons with mental and addictive disorders. *Arch Gen Psychiatr*. 1993;50:95-107.
4. **Murrow CD, Williams JW, Gerety MB, Ramirez G, Montiel OM, Kerber C.** Case-finding instruments for depression in primary care settings. *Ann Intern Med*. 1995;122:913-21.
5. **Williams JW, Noël PH, Cordes JA, Ramirez G, Pignone M.** Is this patient clinically depressed? *JAMA* 287:1160-70.
6. **McDowell I, Kristjansson E, Newell C.** Depression. In: McDowell I, Newell C., eds. *Measuring Health: A Guide to Rating Scales and Questionnaires*. 2nd edn. New York, NY: Oxford University Press; 1996:238-86.
7. **Spitzer RL, Kroenke K, Williams JB.** Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of the PRIME-MD: the PHQ primary care study. *JAMA*. 1999;282:1737-44.
8. **American Psychiatric Association.** Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR 4th edn Text Revision. Washington, DC: American Psychiatric Association; 2000.
9. **Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL.** Validation and utility of the Patient Health Questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosom Med*. 2001;63:679-86.
10. **Gräfe K, Zipfel S, Herzog W, Löwe B.** Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica*. 2004;50:171-81.
11. **Becker S, Al Zaid K, Al Faris E.** Screening for somatization and depression in Saudi Arabia: a validation study of the PHQ in primary care. *Int J Psychiatr Med*. 2002;32:271-83.
12. **Kroenke K, Spitzer RL, Williams JBW.** The PHQ-9 validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16:606-13.
13. **Wulsin L, Somoza E, Heck J.** The feasibility of using the Spanish PHQ-9 to screen for depression in primary care in Honduras. *Prim Care Companion J Clin Psychiatr*. 2002;4:191-5.
14. **Löwe B, Kroenke K, Herzog W, Gräfe K.** Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *J Affect Disorders*. 2004;81:61-6.
15. **Löwe B, Spitzer RL, Gräfe K, et al.** Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disorders*. 2004;81:61-6.
16. **Spitzer RL, Williams JB.** Clasificación de los trastornos mentales [Classification of mental disorders]. In: Kaplan HI, Sadock BJ, eds. *Tratado de psiquiatría [Comprehensive Textbook of Psychiatry]*. 2nd edn. Barcelona, Spain: Salvat; 1989:585-607.
17. **Simon GE, Revicki D, VonKorff M.** Telephone assessment of depression severity. *J Psychiatr Res*. 1993;27:247-52.
18. **Aneshensel CS, Frerichs RR, Clark VA, Yocopenic PA.** Measuring depression in the community: a comparison of telephone and personal interviews. *Public Opin Q*. 1982;46:110-21.
19. **Doménech JM.** Fundamentos de diseño y estadística. UD 14: Medida del cambio: Análisis de diseños con medidas intrasujeto [Statistics and Design Basis. UD 14: Change Measures: Intra-Subject Measure Design Analysis]. Barcelona, Spain: Signo; 2002.
20. **Helzer JE, Robins LN, McEvoy LT, et al.** A comparison of clinical and diagnostic interview schedule diagnosis. Physician reexamination of lay-interviewed cases in the general population. *Arch Gen Psychiatr*. 1985;42:657-66.
21. **Evans M, Kessler D, Lewis G, Peters TJ, Sharp D.** Assessing mental health in primary care research using standardized scales: can it be carried out over the telephone? *Psychol Med*. 2004;34:157-62.
22. **Rohde P, Lewinsohn PM, Seeley JR.** Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. *Am J Psychiatr*. 1997;154:1593-8.
23. **Fenig S, Levav I, Kohn R, Yelin N.** Telephone vs face-to-face interviewing in a community psychiatric survey. *Am J Public Health*. 1993;83:896-8.
24. **Jorm AF, Duncan-Jones P, Scott R.** An analysis of the re-test artifact in longitudinal studies of psychiatric symptoms and personality. *Psychol Med*. 1989;19:487-93.