Depression scale (CES-D) in a mixed-mode repeated measurements design: sex and age effects in older adults

SANDRA W. GEERLINGS, AARTJAN TE BEERMAN, DORLY J.H. DEEG, WILLEM VAN TILBURG, JAN H. SMIT, Department of Psychiatry, Free University Amsterdam, Netherlands

ABSTRACT In order to obtain repeated measurements of depression in an efficient and relatively inexpensive design, a mixture of face-to-face interviews and mail questionnaires was employed. The aims of the study were to examine mode effects of face-to-face interviews versus mail questionnaires on depression scores and to test potential interactions between mode of data collection and sex and age of the respondents. In the study sample, which at the outset consisted of 327 depressed and 325 non-depressed older adults (55–85 years) drawn from a larger random community based sample in the Netherlands, depression was measured in successive waves (cycles), using the Center for Epidemiologic Studies of Depression scale (CES-D). With mode of data collection and sex and age of the respondents as independent variables, differences in CES-D scores were analyzed. The CES-D scores were higher when collected by mail questionnaires than when face-to-face interviews were used. No systematic interactions between sex and age of the respondents and mode of data collection were found. For the scores based on mail questionnaires, a transformation is proposed, resulting in scores that are comparable to those obtained by interviews. In studying depression in older adults, more cost-effective mail questionnaires may be used in addition to face-to-face interviews, provided that a transformation is performed before embarking on the analysis.

Key words: depression, older adults, longitudinal studies, mode of data collection, transformation.

Introduction Depression is considered to be one of the most prevalent psychiatric disorders among older adults (Blazer, 1994). This has led to a growing number of studies in which the impact of potential risk factors on both the onset and the course of depression in older adults is evaluated. These studies aim at a better understanding of depression, which may help to design more effective strategies of prevention and treatment. In general, the most appropriate design to evaluate the impact of risk factors on both the onset and the course of depression is a large-scale prospective longitudinal community-based study in which the development of depression is investigated by repeatedly measuring respondents’ depression status over time. The mode commonly used for collecting data on depression is the face-to-face interview, but applying this mode repeatedly in large-scale prospective longitudinal community based studies is expensive. Budget constraints have led to designs in which data are collected with mixed modes: less expensive data collection modes such as mail questionnaires and telephone interviews are used in combination with the more expensive face-to-face interviews. However, using mixed-mode data collection procedures may reduce comparability of the data collected, as the mode may influence the respondents’ answer behaviour, introducing different sources of bias while estimating depression (Greve, 1989; De Leeuw, 1992).

The present study focuses on mode effects in data on depression in a large-scale, prospective community-based study among older adults in which data were collected with mixed modes: face-to-face interviews and mail questionnaires. Systematic differences between data concerning depression derived from face-to-face interviews and mail questionnaires may be expected. Reporting depression is related to disclosure mechanisms. Mail questionnaires are answered in a...
those obtained from the face-to-face interviews is proposed, resulting in scores comparable to the ones obtained from the mail questionnaires. However, it is generally assumed that questions on depression are difficult to answer, and face-to-face interviews are superior to mail questionnaires in explaining procedures, solving misunderstandings and reducing (item) non-response (De Leon, 1992). Those advantages of face-to-face interviews over mail questionnaires have led to the opinion that the former mode may be superior in terms of validity in depression research.

Studies on depression comparing face-to-face interviews with mail questionnaires have led to controversial results. In some studies, mail questionnaires yielded significantly higher depression scores (Buurman and Oliver, 1983; Aquilino, 1998), whereas in other studies no such mode effect was found (King and Buchwald, 1982; Bryson and Pylon, 1984). It has also been suggested that there is an interaction between the mode of data collection and the sex of the respondent. In face-to-face interviews males might be less willing to reveal depression because of the fear of receiving a negative evaluation from the interviewer (Buurman and Oliver, 1981). Mode effects on sensitive topics such as depression may also be related to the age of the respondents. Although empirical studies on mode effects among older adults are scarce, there are indications that especially older respondents’ answer behavior differs in mail and face-to-face data collection procedures (Helsing and Kulka, 1989).

In the present study, data on the Center for Epidemiologic Studies Depression scale (CES-D) (Radloff, 1977) were collected at eight successive waves over a period of three years. In the first and the last wave, data were collected using face-to-face interviews, and in the intermediate period mail questionnaires were used. This procedure enables the study of a mode effect. In the case of a mode effect, a transformation procedure of the answers obtained from the mail questionnaires is proposed, resulting in scores comparable to those obtained from the face-to-face interviews.

**Methods**

**Sampling and procedures**

The present study is part of the Longitudinal Aging Study Amsterdam (LASA), which is a 10-year interdisciplinary study on predictors and consequences of changes in autonomy and well-being in the aging population (Ding and Westendorp de Sarra, 1994). Full details on sampling and response have been described elsewhere (Redman, Ding, Van Limbeek, Brumm, De Vries, Van Tilburg, 1997). At baseline (t), 3,056 interviews were available. On the CES-D (Radloff, 1977), 454 respondents (14.9%) had a score of at least 16, which is the generally accepted cut-off score for a clinically relevant depressive syndrome (Berkman, Berkman, Kail, Freeman, Leo, Osterfield, Coroni-Huntly and Brody, 1986; Beckman et al., 1997). All depressed subjects, and a random sample of non-depressed subjects were asked to participate in a side study concentrating on late life depression, which involved a diagnostic interview at the beginning and a series of follow-up measurements by means of mail questionnaires. The sample participating in this side study consisted of a depressed cohort (n = 327) and a non-depressed cohort (n = 328). Non-response was associated with age (p < 0.001), but not with sex. Follow-up data on depressive symptoms were collected at five-monthly intervals in seven successive waves (t, t1, t2, t3, t4, t5, t6, t7), covering a period of three years. At t1, t6 and t7, data on depressive symptoms were collected by mail questionnaires, and in the final wave (t8), the same face-to-face mode of interviewing was used as at t. For the present study, we used CES-D data from the first and the last wave (t and t8, both interview data), and those from the second and the penultimate wave (t1 and t7, both mail questionnaire data). At time t, at t1, and at t8, 74.5%, 66.6% and 72.1%, respectively, of the participants were present. Attrition after t, was associated with age (p < 0.001) and depression (p < 0.001) but not with sex.

**Measures**

Depressive symptoms were measured using the Center for Epidemiologic Studies Depression scale (CES-D) (Radloff, 1977). This is a 20-item self-report scale developed to measure depressive symptoms in the community. The 20 symptom items constitute four subscales: depressed affect, positive affect, somatic complaints, and interpersonal problems (Radloff, 1977; Smit, Van den Eeden, Ding, Broekman, 1993). Subjects were asked how often they experienced each symptom during the previous week. Items were scored...
In the LASA baseline sample (t0) 14.9% of the respondents scored 16 or higher on the CES-D scale (Radloff, 1977). The psychometric properties of the scale were found to be good in older populations (Radloff and Teri, 1986; Beekman et al., 1994). The respondents were requested to answer the questions while seated in a quiet environment at home. In both the face-to-face interviews and the mail questionnaires, the question order, the question wording and the response categories and their codes were identical to those in the face-to-face interviews. In the mail questionnaires, the question order, the question wording and the response categories and their codes were identical to those in the face-to-face interviews. In both the face-to-face interviews and the mail questionnaires, the reliability indexed by Cronbach's alpha was high (α = 0.89).

Data-analysis

In the LASA baseline sample (n = 410), 14.9% of the respondents scored 16 or higher on the CES-D and in the present study sample this was increased to 50%. Due to this sampling procedure, inverse sampling probability weights (respondents scoring ≤16 : respondents scoring ≥16 = 85.1% : 14.9%) had to be applied in order to achieve a sample which is representative for the original sample. The non-depressed subjects were given a weight of 1.00 and those depressed at baseline were given a weight of 0.30. In all stages of the analysis, CES-D (scales) scores, collected by two different modes, were the dependent variables.

CES-D scores were collected by different modes at successive waves with five-monthly intervals over a period of three years, and thus in analyzing mode effects time-related changes in CES-D scores had to be taken into account. Real changes and artefactual changes can be distinguished. In this study, only scores at two successive waves with intervals of five months were compared. CES-D scores at t1 and t6 (both mail questionnaire data) for respondents at t0 as between-subjects factors. Age was dichotomized at the median (a20 = 70). As the CES-D scale is highly skewed with a minimum of zero, all analyses were repeated, using a logarithmic transformation of the scale score plus one.

To examine whether a mode effect exists on each of the four subscales of the CES-D (Radloff, 1977), two analyses were performed. In the first analysis, where potential bias due to retest effects and sequence effects was ruled out, we averaged CES-D scores at t1 and t6 (both mail questionnaire data) for respondents with valid observations at all four occasions (n = 369). Data were analysed, using multivariate analysis of variance (MANOVA) for repeated measurements, with mode (face-to-face interview versus mail questionnaire) as within-subjects factor and sex and age of respondents at t1 as between-subjects factors. Age was dichotomized at the median (a20 = 70). As the CES-D scale is highly skewed with a minimum of zero, all analyses were repeated, using a logarithmic transformation of the scale score plus one.

To examine whether a mode effect exists on each of the four subscales of the CES-D, the same analyses were repeated at the subscale level (depressed affect, positive affect, somatic complaints, and interpersonal problems). In preliminary analysis, both in the face-to-face interview data and in the mail questionnaire data, the presence of the four subscales was examined using factor analysis (analysis of principal components with varimax rotation). In the case of a mode effect (higher depression scores in the mail questionnaires than in the face-to-face interviews), a transformation is proposed in order to make the CES-D data collected by mail questionnaires comparable with the data collected by interviewers. The standard deviation and the percentage scoring ≥16 the

on a four-point scale, ranging from 0 (rarely or none of the time) to 3 (most of or all the time). The values of these response categories were reversed for the positive affect items. The total CES-D score ranges from zero to 60. A score of 16 or higher is generally used to indicate the presence of a clinically relevant depressive syndrome (Beekman et al., 1996; Beekman et al., 1997). The psychometric properties of the scale were found to be good in older populations (Radloff and Teri, 1986; Beekman, Van Limbeek, Deeg, Wouters, Van Tilburg, 1994). At t0, and t7, CES-D scores were compared using face-to-face interviews and at t0, t6, mail questionnaires were used. The face-to-face interviews were administered by trained interviewers at the respondents' residences. The respondents were requested to answer the questions, while seated in a quiet environment at home. In both the face-to-face interviews and the mail questionnaires, reliability indexed by Cronbach's alpha was high (α = 0.89).

Data-analysis

In the LASA baseline sample (n = 410), 14.9% of the respondents scored 16 or higher on the CES-D and in the present study sample this was increased to 50%. Due to this sampling procedure, inverse sampling probability weights (respondents scoring ≤16 : respondents scoring ≥16 = 85.1% : 14.9%) had to be applied in order to achieve a sample which is representative for the original sample. The non-depressed subjects were given a weight of 1.00 and those depressed at baseline were given a weight of 0.30. In all stages of the analysis, CES-D (scales) scores, collected by two different modes, were the dependent variables.

CES-D scores were collected by different modes at successive waves with five-monthly intervals over a period of three years, and thus in analyzing mode effects time-related changes in CES-D scores had to be taken into account. Real changes and artefactual changes can be distinguished. In this study, only scores at two successive waves with intervals of five months were compared. CES-D scores at t1 and t6 were compared with those at t0. As it is unlikely that the average level of depression changes significantly over a period of five months in older adults in the general population, real changes were not expected between two measurements. However, artefactual changes caused by selective attrition during the study and by repeated measurement (retest effects) had to be taken into account. Furthermore, a sequence effect, due to the fact that in the first two waves an interview (t1) was followed by a mail questionnaire (t6) and in the last two waves the sequence was reversed (t6 mail questionnaire; t1 interview), had to be accommodated for. In order to cope with the above, two separate analyses were performed. In the first analysis, where potential bias due to selective attrition was avoided by using the largest possible number of respondents (n = 519), CES-D scores at t7 were compared with those at t0. For the second analysis, where potential bias due to retest effects and sequence effects was ruled out, we averaged CES-D scores at t0, t7, and t6 (both interview data) and at three, four, and t6 (both mail questionnaire data) for respondents with valid observations at all four occasions (n = 369).
percentage scoring at the cut-off score of 16 or above, thereby preserving interpretable scores and not affecting the original inter-individual differences. These conditions are most likely to be fulfilled when using a generalized T-score transformation (Nunnally, 1978; Kohout, Berkman, Evans, Cornoni-Huntley, 1993). This transformation involves three steps: first, the standardizing (z-score) transformation is applied to each score; second, each z-score is multiplied by the standard deviation of the criterion distribution (CES-D at face-to-face interview); and finally, the mean of the criterion distribution is added to each score. In the absence of significant differences between the means and standard deviations of the CES-D at t₀ and t₁, the mean and the standard deviation at t₀ are used as criterion parameters. In the presence of differences, adjustments will have to be made.

In formula:
\[ X_{\text{trans}} = (X_i - \mu_i) \cdot \sigma_i + \mu_t \]
where \( X_{\text{trans}} \) = scores on the transformed scale
\( i \) = wave number of mail questionnaire (1 – 6)
\( X_i \) = scores on the original scale at \( t_i \)
\( \mu_i, \sigma_i \) = mean and standard deviation of \( X_i \)
\( \mu_t, \sigma_t \) = mean and standard deviation of the criterion distribution at \( t_t \)

Results
In Table 1, mean CES-D scores are presented for the weighted sample. Comparison of mean CES-D scores at t₀ (7.81) with those at t₁ (9.85; difference: 2.04) and those at t₆ (10.55) with those at t₇ (7.92; difference: 2.63) suggests a mode effect. The averaged scores of t₀ and t₃ (both interview data) and of t₁ and t₄ (both mail questionnaire data) were 7.07 and 9.75, resulting in a difference of 2.68 (t = 10.07, df = 368, p < 0.001).

Table 2 presents the results of the first MANOVA analysis. Instead of a full factorial model, a model without the interaction term of sex and age is given, as there was no significant interaction between these factors. The demographic variables sex and age at t₀ had an effect on CES-D scores. Women and older people had relatively high depression scores both at t₀ and at t₁. The mode effect was significant: CES-D scores were higher at t₁ than at t₀. No interactions between demographic characteristics and mode of data collection were found. In the logarithmic transformation condition, the results were essentially the same (data not shown). Results of the second MANOVA analysis are presented in Table 3. The results are similar to those of the first analysis. In the logarithmic transformation condition, the results were essentially the same as those of the first analysis (data not shown). To check whether differences in variances among subgroups affected the results, stratified analyses assessing mode effect among younger old and older old males and females were performed. In all analyses, the same (highly significant) mode effect was found (data not shown). For all subscales a significant mode effect was found, the most prominent of which applied to the positive affect subscale (Table 4).

<table>
<thead>
<tr>
<th>Mode</th>
<th>n</th>
<th>mean (std dev)</th>
<th>median</th>
<th>range</th>
<th>% ≥ 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₀ interview</td>
<td>652</td>
<td>7.81(7.82)</td>
<td>5.00</td>
<td>0–48</td>
<td>15.0</td>
</tr>
<tr>
<td>t₁ mail</td>
<td>519</td>
<td>9.83(7.93)</td>
<td>8.00</td>
<td>0–62</td>
<td>18.8</td>
</tr>
<tr>
<td>t₂ mail</td>
<td>541</td>
<td>9.49(9.99)</td>
<td>8.00</td>
<td>0–61</td>
<td>18.8</td>
</tr>
<tr>
<td>t₃ mail</td>
<td>511</td>
<td>10.64(10.12)</td>
<td>9.00</td>
<td>0–67</td>
<td>25.3</td>
</tr>
<tr>
<td>t₄ mail</td>
<td>497</td>
<td>10.27(8.36)</td>
<td>8.00</td>
<td>0–66</td>
<td>23.0</td>
</tr>
<tr>
<td>t₅ mail</td>
<td>463</td>
<td>10.22(8.03)</td>
<td>9.00</td>
<td>0–65</td>
<td>23.1</td>
</tr>
<tr>
<td>t₆ mail</td>
<td>463</td>
<td>10.55(9.33)</td>
<td>9.00</td>
<td>0–62</td>
<td>23.1</td>
</tr>
<tr>
<td>t₇ interview</td>
<td>486</td>
<td>7.92(7.65)</td>
<td>6.00</td>
<td>0–46</td>
<td>14.8</td>
</tr>
</tbody>
</table>

Significance tests are F ratios from MANOVA: mode effect (sex: F₃,515 = 70.70, p < 0.001).
The mail questionnaire data (t_1 to t_6) were transformed by a T-score transformation. As there were no systematic interactions between mode effect and sex and age of the respondents, one transformation was performed for male and female subjects in all age groups. Thus, CES-D scores were transformed, with no adjustment made for sex and age. The mean and the standard deviation at t_0 and at t_7 were essentially the same (Table 1), as only those at t_0 were used as criterion parameters. Results of the transformation are presented in Table 5. As intended, important characteristics of the CES-D scores, such as the mean and standard deviation and the percentage scoring ≥16 have become very similar.

**Discussion**

The need for prospective longitudinal community-based studies on depression in older adults leads to the question of whether there are economical ways of data collection that may be used in addition to face-to-face interviews. Data collected by different modes were therefore compared. In the present study, mode effects of face-to-face interviews versus self-administered mail questionnaires on CES-D scores were examined. As expected, differences in CES-D scores were observed between the two data collection modes. CES-D scores were significantly higher when they were collected by mail questionnaires than when face-to-face interviews were used. These results replicate the findings of an earlier study using the same scale to assess symptom severity in younger adults (Aquilino, 1998). The differences were demonstrated in each of the four subscales, and they were particularly notable in the positive affect items (such as 'I felt hopeful about the future'). It might be argued that positive formulations exert emotional or psychosocial pressure on the respondent to choose an answer in the direction in which the question is formulated, more than negative formulations do (Smit et al., 1995). In the context of the CES-D, and in line with our findings, this may mean that respondents in an interview setting will tend to respond relatively more positively (and thus possibly in a more socially desirable way) on the positive affect items than in mail questionnaires, where no such emotional/psychosocial pressure can be exerted. Concerning the impact of sex and age on mode effects, no systematic interactions between these characteristics and mode of data collection were found. Thus, the mode effect was the same for males and females in all age groups.

Considering the results, it is clear that if CES-D data collected by mail questionnaires and by face-to-face interviews are to be used in one study, a transformation is necessary. A mode effect was found on each of the four subscales of the CES-D, and thus the total scale was transformed. A generalized T-score transformation was shown to be satisfactory (Nunnally, 1978; Kohout et al., 1993). As mean CES-D scores at the beginning and at the end of the follow-up were essentially the same, thereby indicating stability over time, the transformation yielded equal mean depression scores at successive waves. However, the transformation did preserve variability at the individual level, which is the matter of interest in most studies. The mail questionnaire data were transformed using the face-to-face interview data as a criterion. Face-to-face interviews are superior over mail questionnaires in explaining procedures, solving misunderstandings and reducing (item) non-response (De Leeuw, 1992).

There is no external (independent) criterion on the basis of which one can choose between the two modes. Mail
The CES-D in a mixed-mode repeated measurements design

Questionnaires are answered in a more anonymous context, which may result in fewer barriers in revealing sensitive information than in face-to-face interviews. However, using the interview data as a criterion is more in line with clinical practice, and makes the data more comparable with other epidemiological studies in which the measurement of depression is generally incorporated in large-scale interviews. Furthermore, the CES-D was originally designed for use in surveys involving face-to-face interviews; the original norms were obtained within the context of an interview, and the data derived from interviews have generally been used in validation studies (Radloff, 1977; Beekman et al., 1994, 1997). The utility of mail questionnaires is, so far, less well understood. Finally, the proposed transformation is a conservative one (CES-D scores are decreased instead of increased), which precludes overestimations in studies on the onset and the course of depression.

Regarding the present study, some limitations must be acknowledged. First, the results may have been affected by the loss of subjects in the course of the study. It was found that the older subjects and those more depressed at the baseline measurement were more likely to be lost at follow-up. However, the adverse consequences are probably limited, as it was demonstrated that the analyses of the maximum sample lead to essentially the same conclusions as comparable analyses based on the more selective sample. Secondly, the LASA study was not specifically designed to study mode effects in older adults. To reduce confounding and to examine net effects and sequence effects more thoroughly, an experimental design in which respondents are randomly allocated to

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Interview t mean</th>
<th>Mail t mean</th>
<th>Interview t, O mean</th>
<th>Mail t, O mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Depressed affect</td>
<td>1.51</td>
<td>1.70</td>
<td>1.47</td>
<td>1.74</td>
</tr>
<tr>
<td>II Positive affect</td>
<td>2.93</td>
<td>4.93</td>
<td>2.89</td>
<td>4.60</td>
</tr>
<tr>
<td>III Somatization</td>
<td>2.51</td>
<td>3.00</td>
<td>2.45</td>
<td>2.91</td>
</tr>
<tr>
<td>IV Interpersonal problems</td>
<td>0.22</td>
<td>0.31</td>
<td>0.27</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Significance tests are F ratios from MANOVA with mode as within-subjects factor.
*** p<0.001.
** p<0.01.
* p<0.05.

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Interview mean (stddev) median range %</th>
<th>Mail mean (stddev) median range %</th>
</tr>
</thead>
<tbody>
<tr>
<td>I Depressed affect</td>
<td>7.81 (7.82) 5.00 0–48 15.0</td>
<td>7.85 (7.91) 6.00 –4–49 14.5</td>
</tr>
<tr>
<td>II Positive affect</td>
<td>7.96 (7.91) 6.00 –4–49 14.5</td>
<td>7.94 (7.89) 6.00 –4–45 16.1</td>
</tr>
<tr>
<td>III Somatization</td>
<td>7.95 (7.92) 6.00 –4–41 14.3</td>
<td>7.99 (7.86) 7.00 –4–42 13.6</td>
</tr>
<tr>
<td>IV Interpersonal problems</td>
<td>7.90 (7.71) 6.00 –3–37 15.7</td>
<td>7.92 (7.65) 6.00 0–46 14.8</td>
</tr>
</tbody>
</table>
the questionnaire or the face-to-face interview format is recommended for future studies. In previous studies on mode differences, effects due to sampling differences have often been confounded with mode effects (Aquilino, 1998). In the present study, in which mode effects were examined within the same respondents, this potential source of bias was completely ruled out. Finally, weighted data were analysed using the SPSS statistical package, which may affect the computation of standard errors and might lead to erroneously significant findings (Houtzager and Liu, 1998). However, in this study, the adverse consequences of the use of SPSS with weighted data are probably limited. With depression data on the first and the second wave, the mode effect was also studied using Stata, a program that incorporates weighting effects in the estimation of standard errors. In the absence of the option of using multiple analysis of variance (MANOVA) for repeated measurements, a linear regression was applied. The results did not affect the outcome of this study.

In conclusion, this study adds new information to research on the measurement of depression. Until now, most studies on depression rely on data obtained by cross-section, whereas some have measured depression at two successive waves. However, in order to obtain a reliable picture of the development of depression over time, more frequent measurements are required. To limit costs, it is advisable to use a less costly mode, such as mail questionnaires, in addition to the more commonly used face-to-face interviews. However, in using both modes in one study, a mode effect occurs. Mail questionnaires yield systematically higher depression scores than face-to-face interviews in both younger old and older old males and females. It has been shown in this study that a 'sandwich formula' in which mail questionnaires are preceded and followed by a face-to-face interview can be successfully used in a longitudinal study, provided that a transformation is performed in order to deal with the mode effect.

Acknowledgements

This study was conducted as a part of the Longitudinal Aging Study Amsterdam (LASA), which is primarily funded by the Dutch Ministry of Health, Welfare and Sports and by the Netherlands Organization for Scientific Research (NWO), Programme Chronic Diseases - Neurological and Psychiatric Disorders.

References

Kohout FJ, Berkman LF, Evans DA, Cornoni-Huntley J, Two
The CES-D in a mixed-mode repeated measurements design


Correspondence to: S.W. Geerlings, Longitudinal Aging Study Amsterdam, De Boelelaan 1081 C (SCW-o5), 1081 HV Amsterdam, The Netherlands.
Tel. 020-4446769. Fax 020-4446775.
Email S.Geerlings @SCW.VU.NL.