Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Revue d'Épidémiologie et de Santé Publique 59 (2011) 327–340

Revue d'Épidémiologie et de Santé Publique
Epidemiology and Public Health

Original article

Psychometric properties of the Center for Epidemiologic Studies Depression Scale (CES-D) in French clinical and nonclinical adults

Propriétés psychométriques du Center for Epidemiologic Studies Depression Scale (CES-D) sur un échantillon français d'adultes cliniques et non-cliniques

A.J.S. Morin [a,*,b,1], G. Moullec [c,d,1], C. Maïano [e,1], L. Layet [f], J.-L. Just [g], G. Ninot [c]

[a] *University of Sherbrooke, Department of Psychology, Sherbrooke, Canada*
[b] *Educational Excellence and Equity (E3) Research Program, Center for Educational Research (University of Western Sydney), Sydney, Australia*
[c] *Laboratory Epsylon-Dynamics of Human Abilities & Health Behaviors, Departments of Medicine, Human & Sport Sciences, University of Montpellier, 4, boulevard Henri-IV, 34000 Montpellier, France*
[d] *Hôpital du Sacré-Coeur de Montréal, Research Center, 5400, boulevard Gouin-Ouest, H4J 1C5 Montréal, Canada*
[e] *UMR 6233, Étienne-Jules Marey, Institute of Movement Sciences, CNRS-University of Aix-Marseille-II, CP 910, avenue de Luminy, 13288 Marseille cedex 09, France*
[f] *Montfavet Psychiatric Hospital Center, 2, avenue de La-Pinède, 84143 Avignon, France*
[g] *Clinic "La Costière", 171, chemin de La-Costière, 06000 Nice, France*

Received 9 September 2010; accepted 29 March 2011
Available online 16 September 2011

Abstract

Background. – Previous research on the Center for Epidemiologic Studies Depression Scale (CES-D) has five main limitations. First, no study provided evidence of the factorial equivalence of this instrument across samples of depressive and community participants. Second, only one study included systematic tests of measurement invariance based on confirmatory factor analyses (CFA), and this study did not consider the higher-order factor structure of depression, although it is the CES-D global scale score that is most often used in the context of epidemiological studies. Third, few studies investigated the screening properties of the CES-D in non-English-language samples and their results were inconsistent. Fourth, although the French version of the CES-D has been used in several previous studies, it has never been systematically validated among community and/or depressed adults. Finally, very few studies have taken into account the ordered-categorical nature of the CES-D answer scale. The purpose of the study reported herein was therefore to examine the construct validity (i.e., factorial, reliability, measurement invariance, latent mean invariance, convergence, and screening properties) of the CES-D in a French sample of depressed patients and community adults.

Methods. – A total sample of 469 participants, comprising 163 clinically depressed patients and 306 community adults, was involved in this study. The factorial validity, and the measurement and latent mean invariance of the CES-D across gender and clinical status, were verified through CFAs based on ordered-categorical items. Correlation and receiver operator characteristic curves were also used to test the convergent validity and screening properties of the CES-D.

Results. – The present results: (i) provided support for the factor validity and reliability of a second-order measurement model of depression based on responses to the CES-D items; (ii) revealed the full measurement invariance of the first- and second-order measurement models across gender; (iii) showed the partial strict measurement invariance (four uniquenesses had to be freely estimated, but the factor variance–covariance matrix also proved fully invariant) of the first-order factor model and the complete measurement invariance of the second-order model across patients and community adults; (iv) revealed a lack of latent mean invariance across gender and across clinical and community subsamples (with women and patients reporting higher scores on all subscales and on the full scale); (v) confirmed the convergent validity of the CES-D with measures of depression, self-esteem, anxiety, and hopelessness; and (vi) demonstrated the efficacy of the screening properties of this instrument among clinical and nonclinical adults.

* Corresponding author.
E-mail address: alexandre.morin@usherbrooke.ca (A.J.S. Morin).

[1] Since all three contributed equally to the preparation of this paper, the order of appearance of the first, second, and third authors (A.J.S.M., G.M., and C.M.) was determined at random: they should all be considered first authors.

0398-7620/$ – see front matter
doi:10.1016/j.respe.2011.03.061

Conclusion. – This instrument may be useful for assessing depressive symptoms or for the screening of depressive disorders in the context of epidemiological studies targeting French patients and community men and women with a background similar to those from the present study.

Keywords: Cut-off scores; Confirmatory factor analyses; Ordered-categorical items; WLSMV; CES-D; Depression; Mood disorders; Diagnosis; Convergent validity

Résumé

Position du problème. – Les études antérieures sur le Center for Epidemiologic Studies Depression Scale (CES-D) comportent cinq principales limites. Premièrement, aucune étude n'est parvenue à mettre en évidence l'équivalence factorielle de cet instrument auprès d'adultes de la population générale et dépressifs. Deuxièmement, à notre connaissance, une seule étude a eu recours à des tests systématiques d'invariance en employant des analyses factorielles confirmatoires (AFC), et elle n'inclut pas la structure de second-ordre de la dépression, alors que le score global du CES-D est très souvent utilisé dans le contexte d'études épidémiologiques. Troisièmement, peu d'études ont étudié les propriétés de dépistage du CES-D auprès d'échantillons non-anglophones et leurs résultats sont inconsistants. Quatrièmement, bien que la version française du CES-D ait préalablement été utilisée dans plusieurs études, elle n'a jamais été systématiquement validée auprès d'adultes de la population générale et/ou dépressifs. Finalement, peu d'études antérieures ont considéré la nature catégorielle ordonnée des réponses au CES-D. L'objectif de cette étude est d'examiner la validité de construit (i.e. factorielle ; fidélité ; invariance de la mesure ; invariance de moyenne latente ; concomitante ; propriétés de dépistage) du CES-D français auprès d'un échantillon de patients dépressifs et d'adultes de la population générale.

Méthode. – Un échantillon total de 469 participants, comprenant 163 patients adultes dépressifs, et un échantillon de 306 adultes de la population générale, ont été inclus dans cette étude. La validité factorielle, ainsi que l'invariance de la mesure et de la moyenne latente du CES-D – selon le genre et le statut clinique – ont été vérifiées à l'aide d'AFC pour items catégoriels ordonnés. Les corrélations et les courbes caractéristiques de fonctionnement du récepteur ont été utilisées, afin de tester la validité concomitante et les propriétés discriminantes du CES-D.

Résultats. – Les résultats : (i) démontrent la validité factorielle et la fidélité du modèle de mesure de second ordre de la dépression sur la base des réponses aux items du CES-D ; (ii) révèlent l'invariance complète du modèle de mesure de premier et de second-ordre en fonction du genre et une absence d'invariance des moyennes latentes selon le genre (les femmes rapportent des scores significativement plus élevés sur l'ensemble des échelles) ; (iii) montrent une invariance partielle stricte du modèle de mesure de premier ordre (quatre résidus ont dus être librement estimés mais la matrice de variance-covariance factorielle s'est avérée complètement invariante) et l'invariance complète du modèle de mesure de second ordre entre les patients et les adultes de population générale ; (iv) révèlent l'absence d'invariance des moyennes latentes de premier et de second ordre en fonction du genre et du statut clinique des participants (les femmes et les patients présentant des scores plus élevés sur les sous-échelles et l'échelle globale du CES-D) ; (v) confirment la validité concomitante du CES-D avec des mesures de dépression, d'estime de soi, d'anxiété et de désespoir ; (vi) démontrent l'efficacité des propriétés de dépistage de cet instrument auprès d'adultes dépressifs et non-dépressifs.

Conclusion. – Cet instrument peut être utile pour évaluer les symptômes dépressifs, ou dépister les troubles dépressifs majeurs dans le contexte d'études épidémiologique ciblant des populations françaises d'hommes et de femmes dépressifs ou de la population générale présentant des caractéristiques semblables à l'échantillon de la présente étude.

Mots clés : Score seuil ; Analyse factorielle confirmatoire ; Items catégoriels ; WLSMV ; CES-D ; Dépression ; Troubles de l'humeur ; Diagnostic ; Validité concomitante

Developed by the National Institute of Mental Health Center for Epidemiologic Studies, the Center for Epidemiologic Studies-Depression Scale (CES-D) has been widely used to assess depressive symptoms in community and population-based epidemiological studies [1]. This instrument comprises 20 items that cover the main symptoms of depression. These items are grouped into four distinct subscales, which are proposed to converge on a single higher-order factor of depression: depressed affect (DA: blues, sad, etc.), positive affect (PA: hopeful, happy, etc.), somatic complaints (SC: bothered, appetite, etc.), and disturbed interpersonal relationship (IR: unfriendly, disliked, etc.). The participants answer each item on a four-point scale on which they indicate the frequency with which they experienced the corresponding symptom during the past week [0 = rarely or none of the time (less than 1 day); 1 = some or little of the time (1–2 days); 2 = occasionally or a moderate amount of the time (3–4 days); and 3 = most or all of the time (5–7 days)]. From these items, four are reversed-scored to break possible answering tendencies. The total score can vary from 0 to 60, with higher scores indicating a greater number of symptoms.

Radloff [2] conducted the first systematic evaluations of the CES-D psychometric properties on three separate community samples. Using principal component analyses, he found support for the four proposed subscales: DA, PA, SC, and IR. Additional analyses also demonstrated that the full scale presented: (i) acceptable internal consistency coefficients (α) ranging from .85 to .90 in the nonclinical and clinical samples; (ii) moderate test-retest reliability coefficients (r) ranging from .51 to .32 for time intervals varying between 2 weeks and 12 months; and (iii) moderate correlations with several convergent measures of depressive symptoms, general psychopathology, positive and negative affects, social desirability, medication, etc.

Following this initial study, the CES-D has been widely cross-culturally adapted, translated and/or validated in China [3], France [4], Germany [5], Greece [6], Italy [7], the Netherlands [8], Portugal [9], Russia [10], and Spain [11,12], as well as in additional English-speaking samples of community [13] and clinically depressed adults [14] and children or adolescents [15,16]. Although the preceding studies successfully replicated Radloff's results [2] regarding the satisfactory

psychometric properties of this instrument, few of these studies used confirmatory factor analyses (CFA), one gold standard for the evaluation of the construct validity of psychometric inventories. Indeed, in addition to being particularly well suited to the verification of the proposed higher-order factor structure of the CES-D, CFAs directly test theoretically grounded measurement models against observations and extract, latent variables that are net of item-specific measurement errors [17–21].

Fortunately, some studies attempted to replicate Radloff's [2] results on community or clinical samples of adults within a CFA framework [9,22–28], and most of these studies [9,22,24–26,28] also verified whether the four factors could themselves be represented by a single higher-order depression factor. The results from all these studies showed that (i) the a priori four-factor model and the second-order single-factor model fit their data well and better than alternative factor models and (ii) the second-order single-factor model proved slightly superior to the first-order four-factor model. These findings have recently been confirmed in the Shafer [1] meta-analysis of 28 studies published between 1977 and 2001.

Nevertheless, none of the preceding studies provided evidence of the measurement equivalence (i.e., invariance) of the CES-D across samples of depressive and community subjects. This is alarming given that the CES-D is specifically designed to identify *clinical* depression in epidemiological *community* samples. This requires a preliminary verification that the CES-D does indeed measure the same construct, in the same manner, notwithstanding the clinical (depressed versus nondepressed) status of the evaluated individuals [29]. In other words, measurement invariance tests allow one to verify if the higher scores on the instrument–which should be observed in depressed individuals–are really due to higher levels on the construct of interest (i.e., depression) rather than to the instrument measuring a different construct, or measuring it differently in depressed individuals [30]. Such measurement bias could be present when: (i) the items measure the construct with more or less error in the different subgroups (i.e., uniquenesses noninvariance), (ii) the items are scored systematically higher or lower in the various subgroups irrespective of the participant's level on the latent construct of interest (i.e., intercepts noninvariance), or (iii) the items are differently related to the construct of interest in the various subgroups (i.e., factor loadings noninvariance).

In addition, the observation that women present a depression rate twice as high as men (as well as higher average levels) has repeatedly been called one of the best-known facts of psychiatric epidemiology [31]. One possible explanation for gender-based differences in depressive symptoms is that they are not ''real'' and are rather the result of one or more artifacts [32]. Nevertheless, these artifactual explanations were not supported in empirical studies [33–38]. The hypothesis that the items commonly used in the CES-D could be gender-biased has also recently received increased attention from epidemiologists and psychologists. Indeed, since 1993, five studies have investigated potential gender biases in the CES-D [27,39–42]. Although they were based on different methodologies, these studies suggest that, given similar levels of depression, women were likely to score higher (intercepts noninvariance) than men on some items (item 17: ''*I had crying spells*'' [27,39–42]; item 10: ''*I felt fearful*'' [40,42]; Item 11: ''*My sleep was restless*'' [40,42]), while men were more likely to score higher on item 13 (''*I talked less than usual*'' [27]). However, only one of these studies used a CFA methodology [27]. It is interesting to note that this study found no evidence of noninvariance in additional model parameters (loadings and uniquenesses).

Finally, even though the CES-D was initially developed by Radloff [2] for the identification of clinical levels of depression in epidemiological studies, few studies have investigated the appropriateness of the proposed cut-off scores–limiting its use to the evaluation of depressive symptom intensity [9]. Original research based on receiver operator characteristic (ROC) curves designed to optimize sensitivity and specificity suggest a cut-off score of 16 for the total sample [14]. Additional studies using the same technique among English-speaking samples provided divergent cut-off scores ranging from 12 [43] to 27 [44]. In addition, until recently, few studies cross-culturally investigated the screening property of the CES-D; their results are also divergent. For example, in two Spanish studies these cut-off scores ranged from 16 [11] to 26 [12], whereas in Portuguese and Greek samples the cut-off scores ranged from 20 [9] to 23/24 [6].

1. The present study

The goal of the present study was to further investigate the reliability, validity, measurement invariance, and appropriate cut-off scores of the CES-D, based on a CFA approach. The main CFA model that will be tested hypothesized a priori that the answers to the CES-D could be explained by four first-order factors (DA, PA, SC, and IR), which in turn would load on a single second-order factor representing depression. This model will be compared to various alternative models previously reported in the literature [1,28,45] and will first be tested on a pooled sample of male and female community adults and clinically depressed patients. Then the measurement invariance of the CFA model will be verified on various subgroups (males and females, community and clinical). The criterion-related validity of the resulting factor model will also be estimated by comparing the subscales and total scale scores with results from another validated measurement of depression (the Beck Depression Inventory) as well as with measurements of various constructs known to be related to depression, such as anxiety, hopelessness, and self-esteem [46–49].

In the present study, the French version of the CES-D [4] was administered to a sample of adults. This represented an additional challenge for the present study, while contributing new data to the literature. Indeed, the current French version of the CES-D, although it has previously been used in several studies [50–52], has never been systematically validated among community and/or clinically depressed adults. Indeed, the Furher and Rouillon [4] only presented information regarding the translation of the questionnaire and suggested cut-off points for men (17) and women (23). Thus, the systematic validation

of the French CES-D is also an important contribution in its own right, especially given that French (i) is the official language in 32 countries and territories worldwide [53], (ii) is the main language in five European countries (France, Belgium, Switzerland, Monaco, and Luxembourg), (iii) is one of the official languages of European institutions and remains the most widely taught second language, (iv) is one of the United Nations' two official languages, and (v) is also one of Canada's two official languages.

2. Method

2.1. Participants and procedures

A total of 469 participants were involved in this study (65.7% females) with a mean age of 40.7 years (standard deviation [SD], 16.2; range, 18–89 years). This sample comprised a first subsample of 306 community adults (59.5% females) not currently suffering from a major depressive episode (MDE) or any mental disorder, with a mean age of 35.4 years (SD, 14.3; range, 18–82 years). The second subsample consisted of 163 patients (77.3% females) with a mean age of 50.6 years (SD, 15.1; range, 19–89 years) suffering from a MDE according to the DSM-IV [54] and ICD-10 [55] criteria. All participants gave written informed consent and the study protocol was carried out in accordance with the standards of the local ethics committee.

The first subsample comprised volunteer adults from southern France (Avignon, Montpellier, Nice, and Marseille) who were recruited from various university classes and student families. A brief interview with the volunteers was first conducted by a member of the research team and followed by the administration of sections of the Mini International Neuropsychiatric Interview (MINI) [56]. This procedure was used to confirm that all participants were physically healthy and did not suffer from an MDE or any other mental disorder. The volunteers who failed to meet these criteria were excluded from the study. The second subsample was recruited within an inpatient unit in a public psychiatric hospital (Hôpital de Montfavet) and two private clinics (la Costière and Saint-Luc) located in southern France. Clinical diagnosis was reached with the fifth French version of the MINI. Only patients with a diagnosis of MDE (single or recurrent) on the MINI were included in the study. Of the eligible patients, those with alcohol addiction and/or psychotic disorders according to DSM-IV and ICD-10 criteria were excluded from the study. All questionnaires used in this study as well as clinical interviews (MINI) were administered by members of the research team in a single one-on-one session. To ensure the uniform assessment of the clinical group, the same research assistant administered the questionnaires and the interviews to all patients.

2.2. Measurements

2.2.1. Clinical diagnosis

The presence of a MDE diagnosis was assessed with the fifth French version of the MINI [56]. This instrument is a short structured diagnostic interview that can be used as a tool to diagnose 16 axis I psychiatric disorders according to DSM-IV and ICD-10 criteria. Each of the MINI's 16 separate modules involves standardized close-ended questions. Interviewers read these questions verbatim to the interviewees. Psychiatric diagnosis and history in each specific module was made according to the number of affirmative replies to the questions. MINI ratings have been shown to possess acceptable rates of sensitivity (.94) and specificity (.79) for the diagnosis of a MDE and elevated rates of inter-rater reliability for all 16 diagnoses (kappa coefficients ranging from 0.88 to 1.00; for greater detail on the reliability and validity of the MINI and its convergence with both DSM and ICD diagnoses, see [57] and [58]).

2.2.2. Depression

Two instruments were used to assess the severity of depressive symptoms: the previously described French version [4] of the CES-D [2] and the French version [59,60] of the 13-item Beck Depression Inventory (BDI-13) [61]. The items from the French version of the CES-D are presented in Table 1.

The French BDI-13 comprises 13 items rated on a behaviorally anchored answer scale ranging from 0 (absence of symptoms) to 3 (most severe symptoms) to assess symptom severity during the past week including today. In previous studies, the French BDI-13 presented good internal consistency ($\alpha = .90$) and moderate 4-month test-retest correlations ($r = .62$) [59,60]. In this study, the internal consistency of the BDI was also satisfactory ($\alpha = .93$).

2.2.3. Anxiety

The French version [62] of the Beck Anxiety Inventory (BAI) [63] was used to assess the severity of participants' symptoms of anxiety. Respondents indicated the degree to which they had been bothered by each of the 21 symptoms during the "past week including today" on a severity scale ranging from 0 (not at all) to 3 (severely, I could barely stand it). It has been shown that the French BAI presents an excellent internal consistency with community adults (α ranging from .84 to .93) and a satisfactory 4-week test-retest correlation ($r = .63$) [62]. In this study, the internal consistency of the BAI was satisfactory ($\alpha = .93$).

2.2.4. Hopelessness

The French version [64] of the Beck Hopelessness Scale (BHS) [65] was used to measure negative attitudes about the future experienced by the respondents over the past week. This instrument consists of 20 true–false statements that are scored 0 or 1. In previous studies, the French version of BHS showed excellent internal consistency in clinically depressed ($\alpha = .89$) and community ($\alpha = .79$) samples, as well as a satisfactory test-retest correlation over 2 weeks ($r = .81$) [64]. In this study, the internal consistency of the BHS was satisfactory ($\alpha = .88$).

2.2.5. Self-esteem

The French version [66] of the Rosenberg Self-Esteem Inventory (RSEI) [67] was used to assess overall feelings of self-worth or self-acceptance. The 10 items from this

Table 1
Items of the French CES-D.

N°	Items	Scale
1	I was bothered by things that usually don't bother me (*J'ai été contrarié(e) par des choses qui d'habitude ne me dérangent pas*)	SC
2	I did not feel like eating; my appetite was poor (*Je n'ai pas eu envie de manger, j'ai manqué d'appétit*)	SC
3	I felt that I could not shake off the blues even with help from my family and friends (*J'ai eu l'impression que je ne pouvais pas sortir du cafard, même avec l'aide de ma famille et de mes ami(e)s*)	DA
4	I felt that I was just as good as other people *(J'ai eu le sentiment d'être aussi bien que les autres)*	PA[a]
5	I had trouble keeping my mind on what I was doing *(J'ai eu du mal à me concentrer sur ce que je faisais)*	SC
6	I felt depressed *(Je me suis senti(e) déprimée)*	DA
7	I felt that everything I did was an effort *(J'ai eu l'impression que toute action me demandait un effort)*	SC
8	I felt hopeful about the future *(J'ai été confiant(e) en l'avenir)*	PA[a]
9	I thought my life had been a failure *(J'ai pensé que ma vie était un échec)*	DA
10	Je me suis senti(e) craintif(ve). *(I felt fearful.)*	DA
11	My sleep was restless *(Mon sommeil n'a pas été bon)*	SC
12	I was happy *(J'ai été heureux(se))*	PA[a]
13	I talked less than usual *(J'ai parlé moins que d'habitude)*	SC
14	I felt lonely *(Je me suis senti(e) seul(e))*	DA
15	People were unfriendly *(Les autres ont été hostiles envers moi)*	IR
16	I enjoyed life *(J'ai profité de la vie)*	PA[a]
17	I had crying spells *(J'ai eu des crises de larmes)*	DA
18	I felt sad *(Je me suis senti(e) triste)*	DA
19	I felt that people disliked me *(J'ai eu l'impression que les gens ne m'aimaient pas)*	IR
20	I could not get "going" *(J'ai manqué d'entrain)*	SC

CES-D: Center for Epidemiologic Studies - Depression scale; DA: Depressed affect; PA: Positive affect; SC: Somatic complaints; IR: Disturbed interpersonal relationships.
[a] Reversed score.

instrument are rated on a 4-point Likert scale ranging from "strongly agree" (4) to "strongly disagree" (1). In previous studies, the French version of the RSEI showed acceptable internal consistency coefficients (α = .70–.90) and a satisfactory test-retest correlation over 3 weeks (r = .84) [66]. In this study, the internal consistency of the RSEI was in the acceptable range (α = .75).

2.3. Analyses

As the CES-D items are rated on a four-point ordered-categorical answer scale, the maximum likelihood (ML) estimation (through classical or robust ML estimators) was deemed inappropriate in light of recent simulation studies showing that a minimum of five answering categories are a prerequisite to the assumptions of continuity underlying ML estimation [68–71]. This conclusion is further reinforced by the significant and elevated non-normality of the data (normalized Mardia coefficients for kurtosis = 181.98). It is interesting to note that most of the previously reviewed CFA studies of the CES-D failed to take this characteristic of the CES-D into account and relied on ML estimation, thus potentially inducing systematic biases in their results (for exceptions, see [25,28]). Following recent recommendations and simulation study results [70,72–75], we decided to use the Mplus 6.1 [76] robust variance-adjusted weighted least squares estimator (WLSMV [75]), which estimates CFA models from polychoric correlation matrices. Assessment of model fit and comparison between models were based on [19,29,77–79]: the chi-square statistic (χ^2), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the root mean square error of approximation (RMSEA), and the 90% confidence interval of the RMSEA. Values greater than .90 for the CFI and TLI are considered to be indicative of adequate model fit, although values approaching .95 are preferable. Values lower than .08 or .06 for the RMSEA support acceptable and good model fit, respectively. Concerning the RMSEA 90% CI, values less than .05 for the lower bound (left side) and less than .08 for the upper bounds (right side) or containing 0 for the lower bound and less .05 for the upper bounds (right side) indicate acceptable and good model fit.

Measurement invariance tests across gender and clinical groups were performed in a sequential strategy devised through a combination of Meredith and Teresi's [30] recommendations for first-order factor models and Cheung's [80] recommendations for higher-order factor models. The measurement invariance of the first-order factor model was estimated first, without a second-order latent construct [80], in the following sequence that was adjusted to the ordered-categorical nature of the items [81,82]: (i) configural invariance, (ii) weak invariance (invariance of the factor loadings), (iii) strong invariance (invariance of the loadings and thresholds), (iv) strict invariance (invariance of the loadings, thresholds, and uniquenesses), (v) invariance of the variance–covariance matrix (invariance of the loadings, thresholds, uniquenesses, and variances–covariances), and (vi) latent means invariance (invariance of the loadings, thresholds, uniquenesses, variances–covariances, and latent means). Then the invariance of the second-order structure was verified in the following sequence, with the baseline specified according to the conclusions of steps (i)–(iv) of the preceding sequence: (i) second-order configural invariance, (ii) second-order weak (loadings) invariance, (iii) second-order strong (loadings and intercepts), (iv) second-order strict (loadings, intercepts, and disturbances) invariance, (v) second-order variance (loadings, intercepts, disturbances, and

variance) invariance, and (vi) second-order latent mean (loadings, intercepts, disturbances, variance, and mean) invariance. Details of model specification under WLSMV are reported in the online appendix.

Critical values for the tests of multigroup invariance across gender or clinical status were evaluated (using the preceding model in the invariance sequence as comparison) by χ^2 difference tests and changes in CFI and RMSEA [29,83,84]. It should be noted that with the WLSMV estimator, the chi-square values are not exact, but rather adjusted or ''estimated'' as the closest integer necessary to obtain a correct *p*-value. Thus, in practice, only the *p*-value should be interpreted. This is especially important for the chi-square difference tests, which cannot be computed by hand but need to be conducted using the Mplus DIFFTEST function ($MD\Delta\chi^2$) [85,86]. However, like the chi-square itself, $MD\Delta\chi^2$ tends to be oversensitive to sample size and to minor model misspecifications. In this regard and to take into account the overall number of $MD\Delta\chi^2$ tests used in this study, the significance level to identify noninvariance was set at .01 [17,82,87]. However, using additional indices to complement the chi-square difference test is also generally recommended [29,83,84]: a CFI decrease of .01 or less and a RMSEA increase of .015 or less between a model and the preceding model in the invariance hierarchy indicate that the measurement invariance hypothesis should not be rejected.

3. Results

3.1. Stage 1. Factor validity and reliability of the CES-D models

Six a priori CFA models from the available literature [1,28,45] were examined for the CES-D scores: (i) a one-factor model (Model 1); (ii) a two-factor model (Model 2, combining PA and IR in a single factor and combining DA and SC in a second factor); (iii) two different three-factor models (Model 3a, combining PA-DA in a single factor; Model 3b, combining DA and SC in a single factor); (iv) the a priori CES-D four-factor model (Model 4); and (v) the a priori CES-D four-factor model with a single higher-order factor (Model 5). Model 1 a priori hypothesized that: (i) answers to the CES-D could be explained by a single factor of depression; (ii) each item would have a non-zero loading on the depression factor; and (iii) uniquenesses would be uncorrelated. Models 2–5 a priori hypothesized: (i) answers to the CES-D could be explained by two to four first-order factors (see above); (ii) each item would have a non-zero loading on the CES-D factor it was designed to measure and zero loadings on all other factors; (iii) the first-order factors would be correlated (Models 2–4) or load on a single higher-order factor of depression (Model 5); and (iv) uniquenesses would be uncorrelated.

The goodness-of-fit statistics of these various CFA models are reported in Table 2. They show that although all models present satisfactory fit indices, Models 3b, 4, and 5 clearly present a higher level of fit to the data than models 1, 2, and 3a. Comparison of Models 3b and 4 shows almost identical goodness-of-fit indices (with the exception of the RMSEA,-which is slightly better for Model 4) but a significant $MD\Delta\chi^2$ (15.36, df = 3, $p \leq .01$), favoring the a priori Model 4. In addition, examination of the factor loadings of the combined DA-SC factor revealed that this factor is mostly defined by the DA items, with the vast majority of the SC items showing lower factor loadings. In accordance with the a priori hypotheses, Model 4 was retained. Then comparison of Model 4 with the higher-order factor Model 5 again showed almost identical goodness-of-fit indices and a nonsignificant $MD\Delta\chi^2$ (8.57, df = 3, $p \geq .01$). Since Model 5 is convergent with the theoretical framework underlying the CES-D and provides an equivalent degree of fit to data as Model 4, while being more parsimonious (replacing six latent factors correlations with four second-order factor loadings and thus freeing two degrees of freedom), this hierarchical model was retained for the following analyses (Table 2). The standardized factor loadings, reported in Fig. 1, are all significant and substantial. The second-order factor loadings associated with the DA and SC factors are very high. They refer to the degree to which the higher-order latent variable (i.e., depression) predicts the first-order factors. The amount of variance in the first-order factor remaining unexplained by the second-order factor is reflected by the first-order disturbances and is a direct function of the loadings (calculated as 1 minus the squared loading). This disturbance reflects the ''unique'' part of the first-order factor, i.e., independent of the higher-order depression factor, and thus reflects its specificity. The fact that some of the second-order loadings are quite high indicates that most of what is assessed by the DA and SC factors is determined by the underlying depression factor. On the contrary, the PA and IR factors incorporate more specificity. It is important to note here that higher-order factors are estimated from first-order factors that are already assessed without item-specific measurement error, which is absorbed by the items' uniquenesses. Thus, first-order factor disturbances reflect variance that is unrelated to depression but also unrelated to random measurement error. This unique variance has been called systematic error in the psychometric literature. The large size of the second-order loadings indicates a low level of systematic measurement errors in the first-order factors.

The factors' reliability was computed from the model's standardized parameters, using McDonald's [88] ω coefficient: $(\Sigma|\lambda i|)^2/([\Sigma|\lambda i|]^2 + \Sigma\delta ii)$ where λi are the factor loadings and δii the uniquenesses. The results revealed that the scales of this model presented for the pooled sample, acceptable ω coefficients equal to .96 for DA, .86 for PA, .91 for SC, .83 for IR, and .93 for the full scale.

3.2. Stages 2–3. Measurement and latent mean invariance across gender and clinical groups

In the second and third stages, the second-order CFA model was first estimated separately in gender-related (Models 6a and 6b) and clinical/nonclinical subsamples (Models 8a and 8b). Then measurement invariance tests across gender (Models 7a and 7b) and clinical groups (Models 9a and 9b) were conducted

Table 2
Goodness of Fit Indices of CES-D Models*.

Stages	Model	N°	Description	χ^2	*df*	CFI	TLI	RMSEA	RMSEA 90% CI	MD$\Delta\chi^2$(*df*)	ΔCFI	ΔTLI	ΔRMSEA
Stage 1	CFA	1	Single factor model	734.884*	170	.973	.970	.084	.078–.090				
		2	Two correlated factors: 1: PA + IR; 2: DA + SC	515.152*	169	.984	.982	.066	.060–.073				
		3a	Three correlated factors: 1: SC; 2: IR; 3: PA + DA	571.679*	167	.981	.978	.072	.065–.078				
		3b	Three correlated factors: 1: PA; 2: IR; 3: DA + SC	321.044*	167	.993	.992	.044	.037–.052				
		4	Four correlated factors: 1: DA; 2: SC; 3: PA; 4: IR	307.104*	164	.993	.992	.043	.036–.051				
		5	Four 1st-order factors and one 2nd-order factor	315.460*	166	.993	.992	.044	.036–.051				
Stage 2	CFA, 1st-order gender-invar.	6a	Men ($n = 161$)	194.662*	164	.995	.994	.034	.000–.051				
		6b	Women ($n = 308$)	275.960*	164	.992	.991	.047	.037–.057				
		7a	1-Configural invariance	471.741*	328	.993	.992	.043	.034–.052				
			2-Weak invariance (loadings)	493.419*	344	.993	.992	.043	.034–.051	26.965 (16)	.000	.000	.000
			3-Strong invariance (thresholds)	518.899*	380	.993	.993	.039	.031–.048	43.389 (36)	.000	+.001	−.004
			4-Strict invariance (uniquenesses)	543.057*	400	.993	.993	.039	.030–.047	39.091 (20)*	.000	.000	.000
			5-Variances-covariances invariance	504.746*	410	.993	.996	.031	.021–.040	13.799 (10)	.000	+.003	−.008
			6-Latent means invariance	709.378*	414	.985	.987	.055	.048–.062	54.397 (4)*	−.008	−.009	+.024
	CFA, 2nd-order gender-invar. (from 7a4)	7b	1-Configural invariance (from model 7a4)	549.646*	404	.993	.993	.039	.031–.047				
			2-Weak invariance (2nd-order loadings)	575.894*	407	.992	.992	.042	.034–.050	14.185 (3)*	−.001	−.001	+.003
			3-Strong invariance (2nd-order inter./1st order means)	570.070*	410	.992	.993	.041	.032–.049	1.504 (3)	.000	+.001	−.001
			4-Strict invariance (2nd-order uniq./1st order var.)	566.085*	414	.992	.993	.040	.031–.047	5.599 (4)	.000	.000	−.001
			5-Variance invariance of the 2nd-order factor	510.521*	415	.992	.996	.031	.021–.040	0.688 (1)	.000	+.003	−.009
			6-Latent mean invariance of the 2nd-order factor	713.973*	416	.985	.986	.055	.048–.062	26.424 (1)*	−.007	−.010	+.024
Stage 3	CFA, 1st-order clinical-invar.	8a	Community sample ($n = 306$)	254.493*	164	.977	.974	.042	.032–.052				
		8b	Depressed patients ($n = 163$)	225.503*	164	.973	.968	.048	.031–.063				
		9a	1-Configural invariance	523.978*	328	.969	.964	.050	.042–.058				
			2-Weak invariance (loadings)	514.236*	344	.969	.970	.046	.037–.054	7.532 (16)			
			3-Strong invariance (thresholds)	584.986*	380	.968	.968	.048	.040–.055	84.924 (36)*	−.001	−.002	+.002
			4-Strict invariance (uniquenesses)	732.541*	400	.947	.950	.060	.053–.066	108.272 (20)*	−.021	−.018	+.012
			4′-Partial strict invariance (items 1, 2, 11, 15 free)	663.887*	396	.959	.959	.054	.047–.061	68.092 (16)*	−.009	−.009	+.006
			5-Variances-covariances invariance (from 4′)	659.119*	406	.959	.963	.052	.044–.059	27.067 (10)*	.000	+.004	−.002
			6-Latent means invariance (from 4′)	3258.110*	410	.550	.583	.172	.167–.178	724.368 (4)*	−.409	−.380	+.120
	CFA, 2nd-order clinical-invar. (from 9a4′)	9b	1-Configural invariance (from model 9a4′)	668.270*	400	.958	.960	.053	.046–.061				
			2-Weak invariance (2nd-order loadings)	675.454*	403	.957	.959	.054	.047–.061	9.976 (3)	−.001	−.001	+.001
			3-Strong invariance (2nd-order inter./1st-order means)	709.534*	406	.952	.955	.056	.050–.063	20.131 (3)*	−.005	−.004	+.002
			4-Strict invariance (2nd-order uniq./1st-order var.)	750.284*	410	.946	.950	.059	.053–.066	29.550 (4)*	−.006	−.005	+.003
			5-Variance invariance of the 2nd-order factor	672.722*	411	.946	.962	.052	.045–.059	1.428 (1)	.000	+.012	−.007
			6-Latent mean invariance of the 2nd-order factor	3263.363*	412	.549	.584	.172	.166–.177	381.893 (1)*	−.397	−.378	+.120

CFA: Confirmatory factor analytic model; χ^2 (B-S): Bollen-Stine chi-square; *df*: Degrees of freedom; CFI: Comparative fit index; TLI: Tucker-Lewis index; RMSEA: Root mean square error of approximation; RMSEA 90% CI: 90% Confidence interval for the RMSEA point estimate; DA: Depressed affect; PA: Positive affect; SC: Somatic complaints; IR: Disturbed interpersonal relationships; MD$\Delta\chi^2$: Change in χ^2 relative to the preceding model calculated from Mplus DIFFTEST function; ΔCFI: Change in comparative fit index relative to the preceding model; ΔTLI: Change in Tucker-Lewis index relative to the preceding model; ΔRMSEA: Change in root mean square error of approximation relative to the preceding model; *$p < .01$.

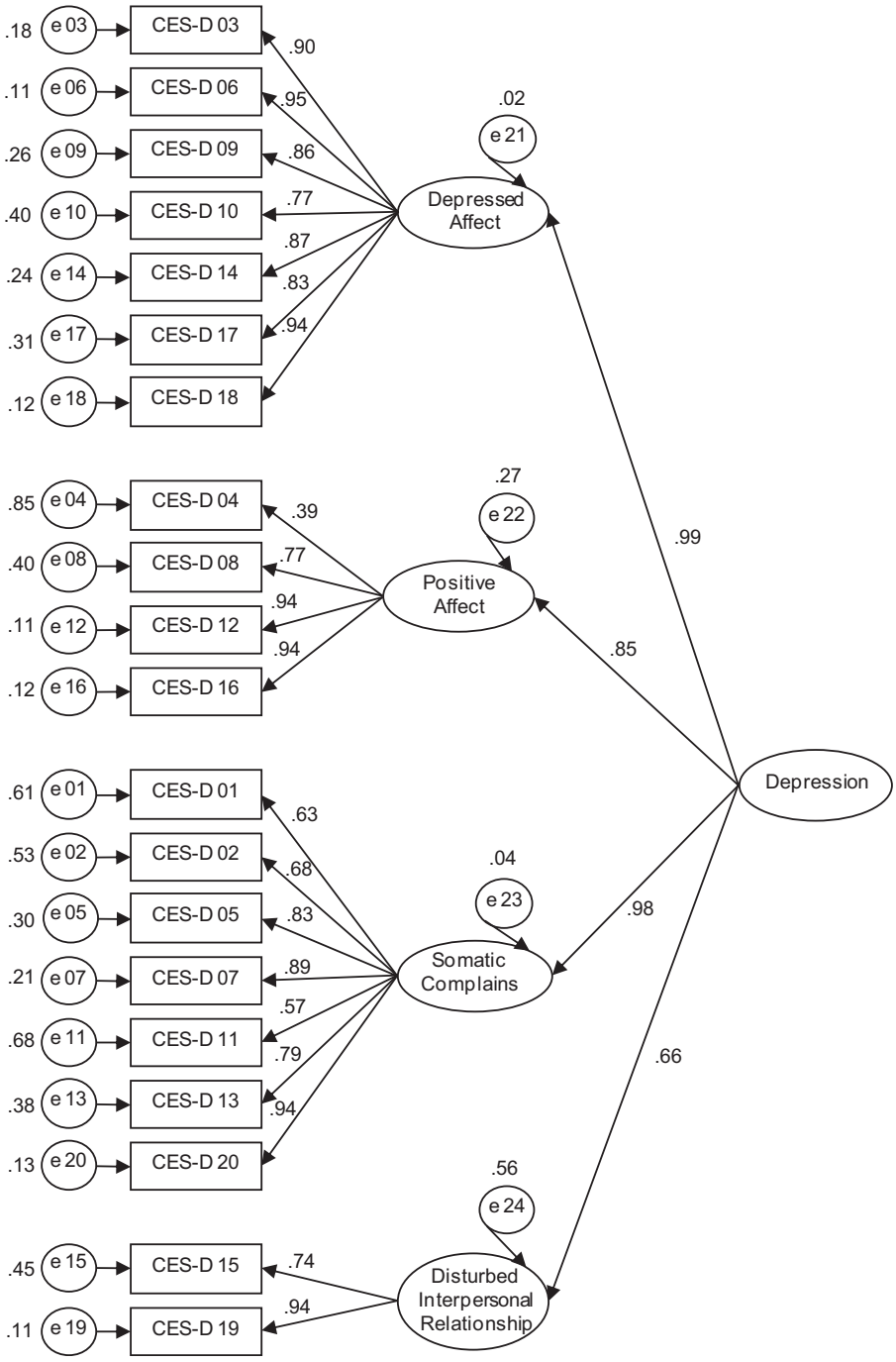

Fig. 1. Estimated standardized uniquenesses, disturbances, and loadings for Model 5. All loadings are significant at $p < .001$.

in the above-described sequential strategy. The results from these models are reported in Table 2 and show that the a priori higher-order factor model provided a satisfactory degree of fit to the data in the specific gender (Models 6a and 6b) and clinical/nonclinical subsamples (Models 8a and 8b).

The results from the gender-based measurement and latent mean invariance for the first-order structure (Model 7a) revealed that the first three steps of invariance testing (i.e., Hypotheses 1–3) resulted in a significant χ^2, acceptable goodness-of-fit indices, and equivalent fit indices (nonsignificant $MD\Delta\chi^2$, $\Delta CFIs \leq .01$, $\Delta RMSEAs \leq .015$). The fourth level of measurement invariance (Hypothesis 4) added equality constraints on the items' uniquenesses. Although this model resulted in a significant $MD\Delta\chi^2$ when compared to the preceding model, the goodness of fit showed absolutely no decrease, suggesting that the χ^2 may be overreacting to minor misspecifications, a hypothesis confirmed by examination of the model's modification indices. Thus, these results confirmed the strict invariance of the first-order measurement model. The next model (Hypothesis 5) tested the invariance of the variance–covariance matrix. This model resulted in a significant χ^2 and acceptable goodness-of-fit indices that show no

decrease compared to the previous model, supporting the full invariance of the variance–covariance matrix. The last model (Hypothesis 6) tested the invariance of the latent factor means and resulted in a significant MD$\Delta\chi^2$, a ΔRMSEA exceeding the .015 criterion, and ΔCFI and ΔTLI approaching the .01 criterion. These results show that the first-order latent factor means are not invariant across gender. Examination of the estimated latent factor means from the preceding model (Hypothesis 5) revealed that women's levels of depression tended to be significantly higher (DA = .595; PA = .496; SC = .510; IR = .364; all $p \leq .01$) than men's levels (latent means set at zero). The results from the subsequent CFAs, in which the gender-based measurement and latent mean invariance of the second-order structure (Model 7b) was verified, support the full (i.e., Hypotheses 1–5) measurement invariance of the higher-order CFA model but indicate the presence of a significant (Hypothesis 6) gender-based latent mean difference on the higher-order depression factor (women = .570 with men's latent mean set at 0, $p \leq .01$). This result is very interesting in that it shows that all of the first-order gender-based latent means differences observed above are fully represented by differences in the higher-order depression factor and thus do not differ across first-order factors (i.e., once the higher-order factor has been included in the model, no significant gender-based differences are observed on the higher-order intercepts of the DA, PA, SC, and IR factors).

The results from the clinical status tests of measurement and latent mean invariance for the first-order structure (Model 9a) revealed that the first three steps of invariance testing (i.e., Hypotheses 1–3) resulted in a significant χ^2, acceptable goodness-of-fit indices, and equivalent fit indices, supporting the strong measurement invariance of the CES-D across clinical status. However, the fourth level of measurement invariance (Hypothesis 4) resulted in a highly significant MD$\Delta\chi^2$, a ΔRMSEA approaching the .015 criterion, and ΔCFI and ΔTLI exceeding the .01 criterion. These results show that the strict invariance hypothesis should be rejected. Inspection of the model's modification indices revealed that this result was specifically due to the noninvariance of the uniquenesses associated with items 1, 2, 11, and 15. When the invariance constraints were relaxed on these specific items (Hypothesis 4′), the results support the strict invariance of the DA and PA factors and the partial strict invariance of the SC and IR factors due to a higher level of item-specific measurement errors on items 1, 2, 11, and 15 in the clinical group, which would be consistent with the attention difficulties inherent to depressive disorders. The last two steps (Hypotheses 5 and 6) confirmed the invariance of the variance–covariance matrix (nonsignificant MD$\Delta\chi^2$, ΔCFIs $\leq .01$, ΔRMSEAs $\leq .015$) across clinical status and quite clearly showed the noninvariance of the first-order latent factor means. Examination of the estimated latent factor means from the preceding model (Hypothesis 5) revealed that clinical participants' level of depression tended to be significantly higher (DA = 2.187; PA = 1.720; SC = 2.003; IR = 1.027; all $p \leq .01$) than nonclinical participants' level (latent means set at zero). The results from the subsequent CFAs, in which the measurement and latent mean invariance of the second-order structure (Model 9b) was verified across clinical/nonclinical status, supported the full (i.e., Hypotheses 1–5) measurement invariance of the higher-order CFA model but indicated the presence of a significant (Hypothesis 6) latent mean difference on the higher-order depression factor (clinical = 2.205 with nonclinical latent mean set at 0, $p \leq .01$). Once again, this result reveals that the first-order latent means differences observed above are fully reflected by differences in the higher-order depression factor.

3.3. Stage 4: criterion-related validity

In the third stage, the criterion-related validity of the CES-D was examined with another measure of depression (BDI-13) and with measures of self-esteem (RSEI), hopelessness (BHS), and anxiety (BAI). In order to minimize Type I error rate inflation, a Bonferroni correction was applied: the alpha error was thus set at .01 (.05/5). The results from these correlation analyses are reported in Table 3 and show that the CES-D overall and subscale-specific scores were significantly and negatively correlated with the RSEI and significantly and positively correlated with the BDI-13, the BAI, and the BHS. As positive and significant relations were expected between these instruments and the CES-D, these results support the criterion-related convergent validity of the CES-D. However, it should also be noted that the correlations between the CES-D subscales and the full scale with the BDI-13 and BAI were nearly the same magnitude, whereas it was expected that the CES-D would correlate more strongly with the BDI-13 than with the BAI as proof of its criterion-related divergent validity. Given the known overlap between measures of depression and anxiety, given the fact that the BDI-13 and BAI were specifically developed as complementary instruments, and given the known comorbidity between depression and anxiety, the correlations between the CES-D with both the BDI-13 and the BAI were also computed while partialling out the remaining instrument. More precisely, the correlation between the CES-D and the BDI was computed while partialling out the BAI scores and the correlation between the CES-D and the BAI was computed while partialling out the BDI-13 scores. These adjusted correlations confirmed that the associations between the CES-D and the BDI-13 were higher than the correlations

Table 3
Concurrent Validity of the CES-D.

Scales	BDI-13	BAI	RSEI	BHS
DA	.87* (.65*)[a]	.80* (.39*)[b]	−.67*	.63*
PA	.69* (.47*)	.58* (.09)	−.60*	.61*
SC	.84* (.59*)	.79* (.41*)	−.64*	.60*
IR	.54* (.25*)	.51* (.18*)	−.42*	.42*
Full	.89* (.71*)	.82* (.42*)	−.70*	.67*

DA: Depressed affect; PA: Positive affect; SC: Somatic complaints; IR: Disturbed interpersonal relationships; BDI-13: Beck depression inventory with 13 items; RSEI: Rosenberg self-esteem inventory; BAI: Beck anxiety inventory; BHS: Beck hopelessness scale; * $p < .001$.

[a] Zero-order correlation controlling for (BAI).

[b] Zero-order correlation controlling for (BDI-13).

between the CES-D and the BAI, thus supporting the criterion-related divergent validity of the French CES-D.

3.4. Stage 5: determination of the cut-off points

During the fourth stage, the sensitivity, specificity, true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN) rates were computed to determine appropriate cut-off points for the pooled sample. These rates were calculated for a variety of cut-off scores by comparing them with depression diagnoses obtained from the MINI. The possible gender difference in the sensitivity and specificity of various cut-off points was also verified. Furthermore, an ROC curve was created to represent the relationship between TP (sensitivity) and FP (1–specificity) ratios as a function of various cut-off levels. The area under the curve (AUC) was also calculated in all samples as a measure of the overall accuracy of the scale.

The sensitivity and specificity of the full-scale CES-D at various cut-off levels for the pooled sample and the gender subsamples are reported in Table 4. In the pooled sample, the curve is substantially above the random ROC (AUC = .933; 95%CI, .910–.957) and the optimal cut-off point (i.e., the highest sum of sensitivity and specificity and the lowest difference between the two) for the full scale of the CES-D appeared to correspond to a score of 19. This cut-off point, which provided a sensitivity of .853 and a specificity of .859, resulted in the accurate classification of 263 community adults and 139 patients and in the erroneous classification of 43 community adults and 24 patients. The possible gender differences in sensitivity and specificity rates were also tested at various cut-off points. In the men's sample, the curve is substantially above the random ROC (AUC = 929; 95%CI, .875–.984) and the optimal cut-off point also appeared to correspond to a score of 16. This cut-off point, which provided a sensitivity of .865 and a specificity of .871, resulted in the accurate classification of 108 community adults and 32 patients and in the erroneous classification of 16 community adults and five patients. Finally, in the women's sample, the curve is substantially above the random ROC (AUC = 927; 95%CI .898–.955) and the optimal cut-off point appeared to correspond to a score of 20. This cut-off point, which provided a sensitivity of .841 and a specificity of .852, resulted in the accurate classification of 155 community adults and 106 patients and in the erroneous classification of 27 community adults and 20 patients.

4. Discussion

These findings demonstrate that, in the total sample, the hypothesized second-order factor model satisfactorily fit the data, providing a better fit than the alternative models. These results confirm those from previous studies [9,13,22,24,26–28]. Further analyses also confirmed that the various CES-D subscales had adequate internal consistency coefficients (ω = .83–.96).

In the gender-based comparisons, the results show that the CES-D measurement model was fully invariant, up to the level of the second-order factor variance–covariance matrix, across men and women. These results contradict those from previous studies in which a significant lack of gender-based invariance was observed for many of the CES-D items [27,39–42]. This may be due to biases induced in these previous studies that neglected to specifically consider the non-normal ordered-categorical nature of the CES-D items. Indeed, preliminary analyses of the present data based on traditional ML estimation tend to confirm this hypothesis (not reported here but available upon request from the lead author). Moreover, the first-order and second-order latent means were found to differ across gender in the expected direction, with women showing higher levels of depression than men [31]. Interestingly, our preliminary ML-based analyses failed to find such gender-based differences, suggesting that previous studies in which a lack of gender differences was also observed [89–95] might also have been biased by the arbitrary application of continuous-variable methodologies to ordered-categorical items. However, these results clearly underline that future studies need to pay greater attention to measurement biases in the instruments designed to measure depression and to the effects of using more or less appropriate methodologies. One of the most interesting aspects of the current results is the observation that gender-based differences in first-order DA, PA, SC, and IR factors disappear once the second-order depression factor is taken into account, showing that gender-based

Table 4
Sensitivity and Specificity of the CES-D at various cut-off levels for the pooled and gender subsamples.

Pooled (n = 469)							Men (n = 161)						Women (n = 308)					
Cutoff score	TP	TN	FP	FN	Se	Sp	TP	TN	FP	FN	Se	Sp	TP	TN	FP	FN	Se	Sp
15	149	230	76	14	.914	.752	33	103	21	4	.892	.831	116	127	55	10	.921	.698
16	148	243	63	15	.908	.794	**32**	**108**	**16**	**5**	**.865**	**.871**	116	135	47	10	.921	.742
17	145	251	55	18	.890	.820	31	109	15	6	.838	.879	114	142	40	12	.905	.780
18	142	256	50	21	.871	.837	31	112	12	6	.838	.903	111	144	38	15	.881	.791
19	**139**	**263**	**43**	**24**	**.853**	**.859**	30	114	10	7	.811	.919	109	149	33	17	.865	.819
20	135	270	36	28	.828	.882	29	115	9	8	.784	.927	**106**	**155**	**27**	**20**	**.841**	**.852**
21	134	277	29	29	.822	.905	29	118	6	8	.784	.952	105	159	23	21	.833	.874
22	128	281	25	35	.785	.918	28	118	6	9	.757	.952	100	163	19	26	.794	.896

Se: sensitivity; Sp: Specificity; TP: True positive; FP: False positive; TN: True negative; FN: False negative; the text in bold corresponds to the best cut-off scores in each subgroup.

differences clearly lie at the level of the depression higher-order construct and do not vary across more specific components of depression.

The results also confirm that the CES-D's first- and second-order measurement model was reasonably invariant across the clinical and community subgroups, the only exception being related to the measurement errors associated with four of the 20 items, which were slightly higher in the clinical subgroup, consistent with the attention problems inherent to depression. This partial noninvariance of the items' uniquenesses underlines the importance of relying on latent variable methodologies in depression research since these methods are the only way to control for these biases. When these slight biases were taken into account, the results also showed clear latent mean differences, completely explained by differences in the higher-order latent factor, which confirmed that participants from the clinical subgroup presented higher levels of depression than the community participants. To our knowledge, this is the first time the presence of possible measurement biases has been investigated across clinical and nonclinical subgroups in depression research. If the present results can be replicated, they would clearly support the purported ability of the CES-D to identify clinical depression in community epidemiological samples.

The results show that the subscale and full-scale scores of the CES-D were moderately (RSEI, BHS) or highly (BAI, BDI-13) correlated with measures of depression, self-esteem, anxiety, and hopelessness, which concur with results from previous studies [46–49,96] and support the criterion-related convergent validity of the CES-D. However, the CES-D appeared to correlate highly and equivalently with both the BDI and the BAI. Fortunately, when these correlations were computed while partialling out the variance due to the overlap between these clinical states in order to obtain ''purer'' criterion measures of depression and anxiety, the results confirmed the criterion-related divergent validity of the CES-D, which was found to be more highly correlated to the BDI-13 than to the BAI [97,98].

These results also indicate that the CES-D can be effectively used to detect the possible presence of depressive disorders in clinical and nonclinical settings. For this purpose, the use of a cut-off point of 19 seems optimal, because it accurately classified 85% of the depressed patients and 86% of the community adults. This value is higher than the original score of 16 [14] but using a lower cut-off point than 19 would increase the specificity rate significantly (and thus result in the exclusion of too many depressed participants). On the contrary, the use of a cut-off value higher than 19 would tend to excessively decrease the sensitivity rate and result in the inclusion of too many nondepressed participants. Additional results also demonstrate that the gender of the participants slightly affected the recommended cut-off scores. Indeed, it may be preferable to use: (i) a lower cut-off point (i.e., 16) for men for a similar classification accuracy (87% of the depressed and community men were correctly classified with this cut-off point) and (ii) a higher cut-off point (i.e., 20) for women for a similar classification accuracy (84% of the depressed women and 85% of the community women were correctly classified with this cut-off point). Moreover, it should be noted that these gender-based cut-off points are slightly lower than those recommended by Führer and Rouillon [4] (men: 17; women: 23) with the French translation of the CES-D. Following an anonymous reviewer's suggestion, we complemented this analysis with a newly developed method allowing for the direct incorporation of covariates in ROC analyses and that allows for the estimation of the effects of these covariates on the estimated cut-off scores [99]. In the present study, the results remained unchanged, possibly because a single covariate (gender) was used, for which specific cut-off scores needed to be calculated. However, this method should be seriously considered in the context of future studies in which the effects of multiple covariates, and their interactions, would need to be considered.

Several limitations should be kept in mind when interpreting the findings. First, this study was based on a single sample of adults. Therefore, whether the validity, reliability, and measurement invariance factor of the French CES-D across the overall sample and specific subgroups (i.e., gender, clinical/nonclinical) can be replicated to other samples of adults or to younger or older populations remains an open question. This is especially true for the tests of invariance that needed to be conducted on relatively small samples of men and clinical participants. Although the sample size in these subgroups was deemed sufficient for the present study, it clearly limits the generalizability of the findings and underlines the need for replication efforts, especially among individuals differing from those who participated in the present study. To ensure that this instrument can be used among adults, its validity, reliability, and measurement invariance factor in such populations must first be demonstrated in an independent sample. Finally, the community group was rather homogeneous in terms of age and social profile and consequently cannot be considered representative of the general population. Therefore, replicating these results on a larger clinical sample and a more heterogeneous community sample should be a future research priority.

In conclusion, the psychometric properties of the higher-order depression structure of the French CES-D were found to be adequate. This instrument may be profitably used in research either for assessing depression symptoms or screening depressive disorders, in French patients and community men and women with a background similar to those from the present study.

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

Acknowledgments

The authors thank Valery Balestra-Depont, Cécile Mantrant, Sylvie Gossery, and Muriel Wybo for their contribution to the organization of the study, as well as Pablo Sillon for his assistance. Completion of this study was also made possible by

the dedication of the personnel of each participating centre: Clinic la Costière (Gérard Marro, Georges Benichou, Jean-Pierre Dany, Patrick Ferrer, Richard Moriano, Alain Salimpour, Jean-Charles Struelens), Clinic Saint-Luc (Claude Capdeville, Sophie Vincent, Christian Bessi, Pierre Bonhomme), and Hôpital de Montfavet (Marie-Noëlle Petit, Lisa Rosello and Charles Dahan).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.respe.2011.03.061.

References

[1] Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol 2006;62:123–46.
[2] Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Applied Psychol Measure 1977;1:385–401.
[3] Cheung CK, Bagley C. Validating an American scale in Hong Kong: the Center for Epidemiological Studies Depression Scale (CES-D). J Psychol 1998;132:169–86.
[4] Führer R, Rouillon F. The French version of CES-D: Description and translation of the self-report scale. Psychiatrie Psychobiol 1989;4:163–6.
[5] Hautzinger M. The CES-D scale: A depression-rating scale for research in the general population. Diagnostica 1988;34:167–73.
[6] Fountoulakis K, Iacovides A, Kleanthous S, Samolis S, Kaprinis SG, Sitzoglou K, et al. Reliability, validity and psychometric properties of the Greek translation of the Center for Epidemiological Studies-Depression (CES-D) Scale. BMC Psychiatry 2001;1:3.
[7] Fava GA. Assessing depressive symptoms across cultures: Italian validation of the CES-D self-rating scale. J Clin Psychol 1983;39:249–51.
[8] Beekman AT, van Limbeek J, Deeg DJ, Wouters L, van Tilburg W. A screening tool for depression in the elderly in the general population: the usefulness of Center for Epidemiological Studies Depression Scale (CES-D). Tijdschr Gerontol Geriatr 1994;25:95–103.
[9] Gonçalves B, Fagulha T. The Portuguese Version of the Center for Epidemiologic Studies Depression Scale (CES-D). Eur J Psychol Assess 2004;20:339–48.
[10] Dershem LD, Patsiorkovski VV, O'Brien DJ. The use of the CES-D for measuring symptoms of depression in three rural Russian villages. Soc Indic Res 1996;39:89–108.
[11] Soler J, Perez-Sola V, Puigdemont D, Perez-Blanco J, Figueres M, Alvarez E. Validation study of the Center for Epidemiological Studies-Depression of a Spanish population of patients with affective disorders. Actas Luso Esp Neurol Psiquiatr Cienc Afines 1997;25:243–9.
[12] Vazquez FL, Blanco V, Lopez M. An adaptation of the Center for Epidemiologic Studies Depression Scale for use in non-psychiatric Spanish populations. Psychiatry Res 2007;149:247–52.
[13] Knight RG, Williams S, McGee R, Olaman S. Psychometric properties of the Centre for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life. Behav Res Ther 1997;35:373–80.
[14] Weissman MM, Sholomskas D, Pottenger M, Prusoff BA, Locke BZ. Assessing depressive symptoms in five psychiatric populations: a validation study. Am J Epidemiol 1977;106:203–14.
[15] Faulstich ME, Carey MP, Ruggiero L, Enyart P, Gresham F. Assessment of depression in childhood and adolescence: an evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). Am J Psychiatry 1986;143:1024–7.
[16] Roberts RE, Andrews JA, Lewinsohn PM, Hops H. Assessment of depression in adolescents using the Center for Epidemiologic Depression Scale. Psychol Assess 1990;2:122–8.
[17] Bollen K. Structural Equations with Latent Variables. New York (NY): John Wiley & Sons; 1989.
[18] Borsboom D. The attack of the psychometricians. Psychometrika 2006;71:425–40.
[19] Byrne BM. Factor analytic models: viewing the structure of an assessment instrument from three perspectives. J Pers Assess 2005;85:17–32.
[20] Kahn JH. Factor Analysis in Counseling Psychology Research, Training, and Practice: Principles, Advances, and Applications. Counsel Psychol 2006;34:684–718.
[21] Hagger MS, Orbell S. A confirmatory factor analysis of the revised illness perception questionnaire (IPQ-R) in a cervical screening context. Psychol Health 2005;20:161–73.
[22] Boisvert JA, McCreary DR, Wright KD, Asmundson GJ. Factorial validity of the center for epidemiologic studies-depression (CES-D) scale in military peacekeepers. Depress Anxiety 2003;17:19–25.
[23] Golding JM, Aneshensel CS. Factor structure of the Center for Epidemiologic Studies Depression Scale among Mexican Americans and non-Hispanic Whites. Psychol Assess 1989;1:163–8.
[24] Hertzog C, Alstine JV, Usala PD, Hultsh DF, Dixon R. Measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D) in older populations. Psychol Assess 1990;2:64–72.
[25] Nguyen HT, Kitner-Triolo M, Evans MK, Zonderman AB. Factorial invariance of the CES-D in low socioeconomic status African Americans compared with a nationally representative sample. Psychiatry Res 2004;126:177–87.
[26] Rhee SH, Petroski GF, Parker JC, Smarr KL, Wright GE, Multon KD, et al. A confirmatory factor analysis of the Center for Epidemiologic Studies Depression Scale in rheumatoid arthritis patients: additional evidence for a four-factor model. Arthritis Care Res 1999;12:392–400.
[27] Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R. Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Res 1993;49:239–50.
[28] Sheehan TJ, Fifield J, Reisine S, Tennen H. The measurement structure of the Center for Epidemiologic Studies Depression Scale. J Pers Assess 1995;64:507–21.
[29] Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organ Res Methods 2000;3:4–70.
[30] Meredith W, Teresi JA. An essay on measurement and factorial invariance. Med Care 2006;44:S69–77.
[31] Angold A, Worthman CW. Puberty onset of gender differences in rates of depression: a developmental, epidemiologic and neuroendocrine perspective. J Affect Disord 1993;29:145–58.
[32] Leach LS, Christensen H, Mackinnon AJ. Gender differences in the endorsement of symptoms for depression and anxiety: are gender-biased items responsible? J Nerv Ment Dis 2008;196:128–35.
[33] Bebbington PE. The origins of sex differences in depressive disorder: bridging the gap. Int Rev Psychiatry 1996;8:295–332.
[34] Hankin BL, Abramson LY. Development of gender differences in depression: description and possible explanations. Ann Med 1999;31:372–9.
[35] Hyde JS, Mezulis AH, Abramson LY. The ABCs of depression: integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression. Psychol Rev 2008;115:291–313.
[36] Piccinelli M, Wilkinson G. Gender differences in depression. Critical review. Br J Psychiatry 2000;177:486–92.
[37] Sprock J, Yoder CY. Women and Depression: an update on the report of the APA Task Force. Sex Roles 1997;36:269–303.
[38] Wolk S, Weissman M. Women and depression: an update. Rev Psychiatry 1995;14:227–59.
[39] Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale. experience from the New Haven EPESE study. J Clin Epidemiol 2000;53:285–9.
[40] Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. J Gerontol B Psychol Sci Soc Sci 2000;55:273–82.
[41] Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is cored: An illustration with the Center for Epidemiologic Studies Depression Scale. Educ Psychol Meas 2003; 63:65–74.

[42] Yang FM, Jones RN. Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. J Clin Epidemiol 2007;60:1195–200.
[43] Lewinsohn PM, Seeley JR, Roberts RE, Allen NB. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychol Aging 1997; 12:277–87.
[44] Costello E, Devins G. Screening for depression among women attending their family physicians. Can J Behav Sci 1989;21:434–51.
[45] Perreira K, Deeb-Sossa N, Harris K, Bollen K. What are we measuring? An evaluation of The CES-D across Race/Ethnicity and immigrant generation. Social Forces 2005;83:1567–602.
[46] Abramson LY, Alloy LB, Hankin BL, Haeffel GJ, MacCoon DG, Gibb BE. Cognitive vulnerability-stress models of depression in a self-regulatory and psychobiological context. In: Gotlib IH, Hammens CL, editors. Handbook of Depression. London (UK): Guilford; 2002. p. 268–94.
[47] Angold A, Costello EJ, Erkanli A. Comorbidity. J Child Psychol Psychiatry 1999;40:57–87.
[48] Gotlib IH, Hammens CL. Handbook of depression. New-York, NY: Guilford; 2002.
[49] Roberts JE, Monroe SM. Vulnerable self-esteem and social processes in depression: toward an interpersonal model of self-esteem regulation. In: Joiner Jr TE, Coyne JC, editors. The interactional nature of depression: advances in interpersonal approaches. Washington (DC): American Psychological Association; 1999. p. 149–87.
[50] Cathebras P, Mosnier C, Levy M, Bouchou K, Rousset H. Screening for depression in patients with medical hospitalization. Comparison of two self-evaluation scales and clinical assessment with a structured questionnaire. Encephale 1994;20:311–7.
[51] Chabrol H, Montovany A, Chouicha K, Duconge E. Study of the CES-D on a sample of 1,953 adolescent students. Encephale 2002;28:429–32.
[52] Verdier-Taillefer MH, Gourlet V, Fuhrer R, Alperovitch A. Psychometric properties of the Center for Epidemiologic Studies-Depression scale in multiple sclerosis. Neuroepidemiology 2001;20:262–7.
[53] International Francophone Organisation. http://www.francophonie.org.
[54] American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 4th ed, Washington (DC): American Psychiatric Association; 1994.
[55] World Health Organization. International Statistical Classification of Diseases and Related Health Problems. Geneva (CH): WHO Press; 1994.
[56] Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59(Suppl. 20):22–33 [quiz 4-57].
[57] Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K, et al. The MINI International Neuropsychiatric Interview (M.I.N.I.) A Short Diagnostic Structured Interview: Reliability and Validity According to the CIDI. Eur Psychiatry 1997;12:224–31.
[58] Sheehan DV, Lecrubier Y, Harnett Sheehan K, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry 1997;12:232–41.
[59] Bourque P, Beaudette D. Psychometric study of the Beck Depression Inventory among a sample of French student. Can J Behav Sci 1982; 14:211–8.
[60] Collet L, Cottraux J. The shortened Beck depression inventory (13 items). Study of the concurrent validity with the Hamilton scale and Widlocher's retardation scale. Encephale 1986;12:77–9.
[61] Beck AT, Beck WR. Screening depressed patients in family practice: a rapid technique. Postgrad Med 1972;52:81–5.
[62] Freeston MH, Ladouceur R, Thibodeau N, Gagnon F, Rheaume J. The Beck Anxiety Inventory. Psychometric properties of a French translation. Encephale 1994;20:47–55.
[63] Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988; 56:893–7.
[64] Bouvard M, Charles S, Guerin J, Aimard G, Cottraux J. Study of Beck's hopelessness scale. Validation and factor analysis. Encephale 1992; 18:237–40.
[65] Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol 1974;42:861–5.
[66] Vallières EF, Vallérand R. [French-Canadian translation and validation of the Rosenberg's Self-Esteem Scale]. Int J Psychol 1990;25:305–16.
[67] Rosenberg M. Society and the adolescent self-image. Princeton (NJ): Princeton University Press; 1965.
[68] DiStefano C. The impact of categorization with confirmatory factor analysis. Structural Equation Model 2002;9:327–46.
[69] Dolan CV. Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. Br J Math Stat Psychol 1994;47:309–26.
[70] Finney SJ, DiStefano C. Non-normal and categorical data in structural equation modeling. In: Hancock GR, Mueller, RO (Eds.). Structural Equation Modeling: A Second Course. Greenwich, Connecticut: IAP, 2006:269–314.
[71] Lubke GH, Muthen BO. Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. Structural Equation Modeling 2004; 11:514–34.
[72] Beauducel A, Herzberg PY. On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. Structural Equation Modeling 2006;13:186–203.
[73] Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods 2004;9:466–91.
[74] Forero CG, Maydeu-Olivares A, Gallardo-Pujol D. Factor analysis with ordinal indicators: a Monte Carlo study comparing DWLS and ULS estimation. Structural Equation Modeling 2009;16:625–41.
[75] Muthén BO, du Toit SHC, Spisic D. Robust inference using Weighted Least Squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Mplus Technical Report; 1997; Available from: http://gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf.
[76] Muthén LK, Muthén BO. Mplus user's guide. Los Angeles, CA: Muthén & Muthén; 2010.
[77] Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structure Equation Model 1999;6:1–55.
[78] MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1:130–49.
[79] Yu CY. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Los Angeles: University of California; 2002.
[80] Cheung GW. Testing equivalence in the structure, means, and variances of higher-order constructs with structural equation modeling. Organ Res Methods 2008;11:593–613.
[81] Millsap RE, Yun-Tein J. Assessing factorial invariance in ordered-categorical measures. Multivariate Behav Res 2004;39:479–515.
[82] Morin AJS, Madore I, Morizot J, Boudrias J-S, Tremblay M. The workplace affective commitment multidimensional questionnaire: factor structure and measurement Invariance. Int J Psychol Res 2009;4:307–44.
[83] Chen FF. Sensitivity of goodness of fit indexes to lack of measurement. Structure Equation Model 2007;14:464–504.
[84] Cheung GW, Rensvold RB. Evaluating goodness-of fit indexes for testing measurement invariance. Structure Equation Model 2002;9:233–55.
[85] Muthén BO. Mplus Technical Appendices. Los Angeles, CA: Muthén and Muthén; 2004; Available from: http://www.statmodel.com/techappen.shtml.
[86] Asparouhov T, Muthén BO. Robust chi-square difference testing with mean and variance adjusted test statistics. Los Angeles, CA: Muthén and Muthén 2006; Available from: http://www.statmodel.com/examples/webnote.shtml#web10.
[87] Rensvold RB, Cheung GW. Testing measurement model for factorial invariance: a systematic approach. Educ Psychol Measure 1998;58: 1017–34.

[88] McDonald RP. The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. Br J Math Stat Psychol 1970;23:1–21.
[89] Nolen-Hoeksema S. Sex differences in unipolar depression: evidence and theory. Psychol Bull 1987;101:259–82.
[90] Scheibe S, Preuschhof C, Cristi C, Bagby RM. Are there gender differences in major depression and its response to antidepressants? J Affect Disord 2003;75:223–35.
[91] Christensen H, Jorm AF, Mackinnon AJ, Korten AE, Jacomb PA, Henderson AS, et al. Age differences in depression and anxiety symptoms: a structural equation modelling analysis of data from a general population sample. Psychol Med 1999;29:325–39.
[92] Morin AJS, Janosz M, Larivée S. The Montreal Adolescent Depression Development Project (MADDP): School life and depression following high school transition. Psychiatry Res J 2009;3.
[93] Salokangas RK, Vaahtera K, Pacriev S, Sohlman B, Lehtinen V. Gender differences in depressive symptoms. An artefact caused by measurement instruments? J Affect Disord 2002;68:215–20.
[94] Steer RA, Beck AT, Brown G. Sex differences on the revised Beck Depression Inventory for outpatients with affective disorders. J Pers Assess 1989;53:693–702.
[95] Wenzel A, Steer RA, Beck AT. Are there any gender differences in frequency of self-reported somatic symptoms of depression? J Affect Disord 2005;89:177–81.
[96] Gladstone TR, Kaslow NJ. Depression and attributions in children and adolescents: a meta-analytic review. J Abnorm Child Psychol 1995;23:597–606.
[97] Beck AT, Steer RA. Manual for the Beck Depression Inventory. San Antonio, TX: The Psychological Corporation Harcourt Brace and Company; 1993.
[98] Beck AT, Steer RA. Beck Anxiety Inventory Manual. San Antonio, TX: The Psychological Corporation Harcourt Brace and Company; 1993.
[99] Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol 2008;168:89–97.