BMC Pediatrics

RESEARCH

Open Access

Test-retest reliability and factor structure of the Danish 6-item version of the Hopkins Symptom Checklist-core depression (SCL-6) in adolescents

Christine Leonhard Birk Sørensen[1*], Therese Koops Grønborg[2] and Karin Biering[1]

Abstract

Background Short efficient questionnaires to detect depressive symptoms in adolescents are sparsely evaluated. We aimed to examine the test–retest reliability and structural and convergent validity of the Danish version of the 6-item version of the Hopkins Symptom Checklist-core depression (SCL-6) in adolescents.

Methods Our study population consisted of 122 adolescents. Ninety-one adolescents completed SCL-6 (test sample) in the first round, 82 adolescents completed SCL-6 in the second round (retest sample), with 66 completing the questionnaire both rounds (test–retest sample). Reliability was evaluated by intraclass correlation (ICC) and standard error of measurement (SEM), and structural validity was evaluated by confirmatory factor analysis (CFA) and Mokken analyses. Convergent validity was assessed using the 4-item version of the Centre for Epidemiological Studies Depression Scale (CES-DC4).

Results The mean sum score in the test sample was higher than in the retest sample (0.74 (95% CI:0.06;1.43)). The limits of agreement (LoA) were broad (-4.73;6.21). The intraclass correlation (ICC(1,1)) was 0.79 (95% CI:0.67;0.87)) showed good reliability, while the SEM (2.03 (95% CI: 1.68; 2.37)) was large considering the range of the scale. Cronbach's alpha showed good internal consistency at both test (0.80) and retest (0.81). The inter-item correlations and Mokken analyses supported the conclusion of a unidimensional scale. The CFA did not find an acceptable fit of a one factor solution. The convergent validity showed a moderate correlation with the CES-DC4 (0.48).

Conclusion SCL-6 showed acceptable internal consistency and test–retest reliability. The results from the CFA were inconclusive in terms of demonstrating unidimensionality, but Mokken analyses supported unidimensionality like previous studies. We find the scale usable for population-based research on depressive symptoms in adolescents, but do not recommend it as screening tool on individual level.

Keywords Psychometrics, Depression, Adolescents, Reliability, Validity

Background

Depression is one of the leading causes of disability worldwide according to the World Health Organization (WHO). The WHO stresses that depression can have severe consequences including suicide. Suicide is the fourth leading cause of death in 15–29-year-olds [1]. Therefore, epidemiological research that investigates the

*Correspondence:
Christine Leonhard Birk Sørensen
chleso@rm.dk
[1] Department of Occupational and Environmental Medicine, University Research Clinic, Danish Ramazzini Centre, Goedstrup Hospital, Herning 7400, Denmark
[2] Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark

course, risk factors and consequences of depression is crucial to inform decision makers in applying qualified preventive strategies.

Questionnaires as tools for investigating depression in large populations are quick, cheap, and easy to administer. The Hopkins Symptoms Checklist (SCL-90) is a widely used questionnaire measuring a broad spectrum of psychopathology. The SCL-90 contains 13 items specifically measuring depressive symptoms [2]. However, a competing 90-item version of the scale exists as Derogatis changed two items of the original SCL-90 and developed the SCL-90R [3]. Since then, several subscales have been derived from the SCL-90 and the SCL-90R to specifically measure depressive symptoms. The psychometric properties of the subscales to measure depressive symptoms have been widely tested in the SCL-25 [4–8], the SCL-13 [9], the SCL-10 [10–12], and the SCL-6 [13–16]. The SCL measures depressive symptoms that are consequences of the disease "depression", and therefore the scale follows a reflective model [17].

The SCL-90, the SCL-90R, the SCL-10, and the SCL-6 have all been validated in a Danish context [10, 14, 15, 18]. The SCL have been tested in adolescents [10, 14], adults [15, 18], in the general population [18], in primary care setting [10, 14], and in hospital setting [15]. A study examining the SCL-90 and SCL-90R reported unidimensionality (Mokken analysis, coefficient of homogeneity = 0.52) of the 13-item depression subscale and a Rasch analysis found the subscale robust in the general population in Denmark [18]. The SCL-10 was tested in primary care in 14–16-year-old adolescents. The adolescents were recruited from the patient lists of the included general practitioners and thereby represented the general population of 14–16-year-olds. The study concluded good unidimensionality (Mokken analysis, coefficient of homogeneity = 0.49 and Cronbach's alpha = 0.88) and great criterion validity of the SCL-10 compared with the Composite International Diagnostic Interview (CIDI), analyzed with the receiver operating characteristic (ROC) curve (area under the curve (AUC) = 0.90 (95% confidence interval (CI) 0.83–0.96)). Using a cut-off point of 16 for SCL-10 and CIDI as the gold standard [10], the sensitivity was 87.5% for both girls and boys and the specificity was 72.4% for girls and 87.9% for boys. Moreover, the short depressive subscale, the SCL-6, has been tested in a sample of 14–16-year-olds recruited from the patient lists of randomly selected general practitioners and showed great criterion validity when compared with CIDI (ROC curve, AUC = 0.84 (0.73–0.95)). Christensen et al. found a sensitivity of 0.85 (95% CI: 0.70–0.94) and a specificity of 0.79 (95% CI: 0.74–0.84) when using a cut-point of 9 [14]. Bech et al. tested the psychometric properties of both the full SCL-90 and several short scales in the general population, including the SCL-6. According to their Mokken analysis, the SCL-6 was unidimensional, and the scale discriminated significantly between diagnostic groups of depression [15]. The SCL-6 also showed great psychometric properties in a Swedish study, where Hanson et al. found high degree of unidimensionality (Mokken analysis, coefficient of homogeneity = 0.70). With a cut-point for depression of 17, the specificity was 0.98 and the sensitivity was 0.68 when using an score of ≥ 26 on the Major Depression Inventory (MDI) as gold-standard [13]. In conclusion, studies have found great criterion, discriminative, and structural validity of the SCL-6 and suggest that the SCL-6 can be a useful screening tool for depression.

However, studies of the reliability of the scale are nonexistent, and the structural validity of the scale has only been evaluated with Mokken analyses. Reliability is a prerequisite for validity [17], and it is therefore of great importance to detect whether the SCL-6 is sufficiently reliable. This study aimed to evaluate the SCL-6 in terms of the test–retest reliability and structural and convergent validity using confirmatory factor analysis in Danish adolescents (15–17-years-olds) and compare the results with psychometrics of the 4-item version of the Centre for Epidemiological Studies Depression Scale (CES-DC4), which has been evaluated on the same sample.

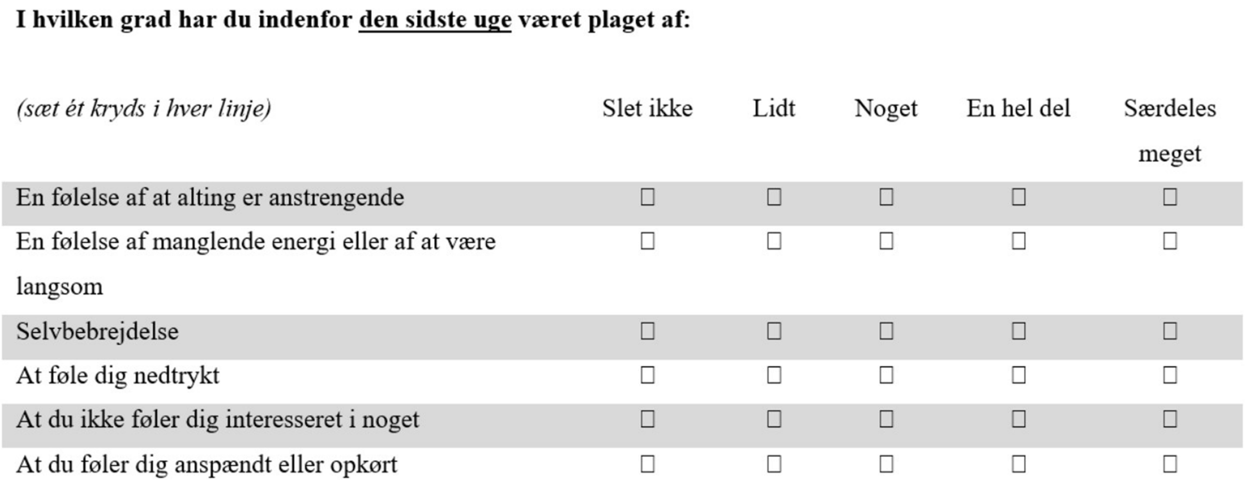
Method

Population

Seventy-one public and private schools with 9th grades in Aarhus, Favrskov, Silkeborg, and Skanderborg municipalities were invited to participate in this study. Schools for pupils with special needs were not invited. Fifty-eight schools did not answer the request, 10 declined to participate and 3 schools agreed to participate. This resulted in five 9th grades participating in the study. Adolescents were included in the study if they were 15 years old and were able to read and understand Danish. The aim was to include at least 50 adolescents after dropout, to fulfill the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criteria of adequate sample size [19, 20].

Data variables

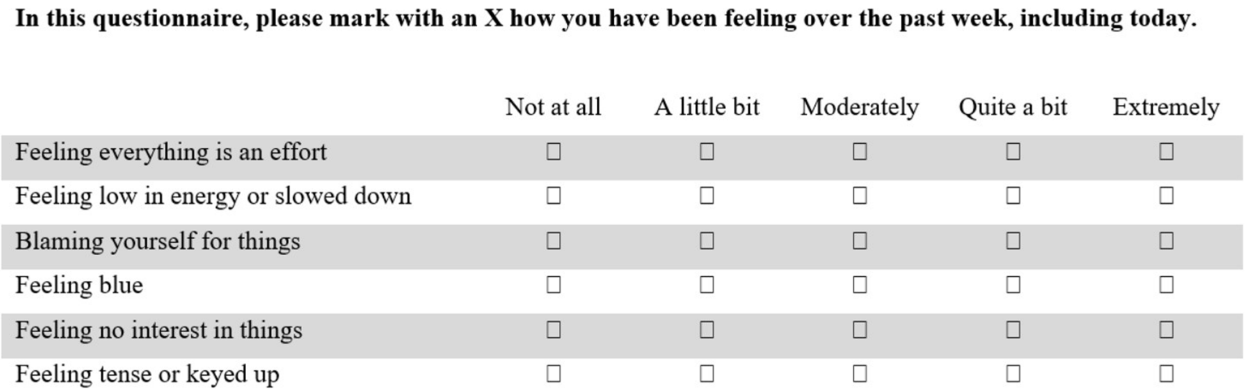
Data collection was in the form of a printed questionnaire with questions about age, sex, the Danish CES-DC4 (Supplementary Fig. 1) and the Danish SCL-6 (Fig. 1). The survey was initiated to evaluate the psychometric properties of the SCL-6 and the CES-DC4 in adolescents [21, 22]. The original English version of the SCL-6 is shown in Fig. 2, and the original English version of the CES-DC4 is shown in Supplementary Fig. 2. From the original scale, the items of the SCL-6 have the numbers



I hvilken grad har du indenfor den sidste uge været plaget af:

(sæt ét kryds i hver linje)	Slet ikke	Lidt	Noget	En hel del	Særdeles meget
En følelse af at alting er anstrengende	□	□	□	□	□
En følelse af manglende energi eller af at være langsom	□	□	□	□	□
Selvbebrejdelse	□	□	□	□	□
At føle dig nedtrykt	□	□	□	□	□
At du ikke føler dig interesseret i noget	□	□	□	□	□
At du føler dig anspændt eller opkørt	□	□	□	□	□

Fig. 1 Danish version of the SCL-6



In this questionnaire, please mark with an X how you have been feeling over the past week, including today.

	Not at all	A little bit	Moderately	Quite a bit	Extremely
Feeling everything is an effort	□	□	□	□	□
Feeling low in energy or slowed down	□	□	□	□	□
Blaming yourself for things	□	□	□	□	□
Feeling blue	□	□	□	□	□
Feeling no interest in things	□	□	□	□	□
Feeling tense or keyed up	□	□	□	□	□

Fig. 2 English version of the SCL-6

71, 14, 26, 30, 32, and 57. The intensity of the depressive symptoms is scored on a 5-point Likert scale resulting in a sum score of 0–24. A higher score represents more severe depressive symptoms [23].

The CES-DC4 measures the same construct, depressive symptoms, as the SCL-6 does. The score of each item in the CES-DC4 ranges from 0–4 and thereby the sum score ranges from 0–16. We expect a strong positive correlation between SCL-6 and CES-DC4, since the scales measure the same construct in the same population at the same time.

Study design

This study rated adequate or good on all parameters in the COSMIN checklist for the design of studies on measurement properties, structural validity, internal consistency, measurement error and reliability, and hypotheses testing for construct validity [24]. The study examines the reliability of the SCL-6 through test–retest with a two-week interval. The two-week interval was chosen because it was assumed that the adolescents had stable depressive symptoms in this time interval, and that two weeks would minimize recall bias from test to retest [17].

Data collection

The data collection was conducted from January 2020 to March 2020. During the data collection period, the national winter vacation took place from the 10th of February until the 16th of February. Supplementary Fig. 3 shows the timing of the test and retest for the classes compared with the winter vacation. Both the test and the retest took place in the classrooms while the adolescents were placed in their seats. The tests were introduced, which included a presentation of the research, information about voluntary participation and anonymity, and information about the practical execution of fulfilling of the questionnaires. To secure anonymity, every student got an emoji on their questionnaire and a token with the emoji on. At the retest they showed the token and got a questionnaire with the same emoji. Hereby, the questionnaires from the test and the retest could be paired, while the authors were unable to connect the questionnaire answers to a specific student. The same introduction was conducted at the test and the retest, apart from a small change of the information of practical execution since at the test the adolescents were assigned an emoji, and at the retest, they were asked to remember their emoji.

Statistical analysis

The characteristics of the included adolescents were described with total numbers and percentages for age and sex. An analysis of the descriptive data was performed between the adolescents who completed both questionnaires and the adolescents who were either not present at one of the tests or had missing items in one of the tests.

The internal consistency was evaluated at test and retest by Cronbach's α, item-rest correlations, and inter-item correlation [17, 25]. Spearman's ρ method was used to estimate the inter-item correlations. Confirmatory factor analysis (CFA) and Mokken analyses were conducted to assess the structural validity of the scale at both the test and retest scores. A variance-adjusted weighted least-squares method (WLSMV) estimator was used to treat the items of the scale categorical. The scale is constructed as a unidimensional scale, so a one-factor model was expected to fit the scale. Model fit was evaluated by goodness-of-fit and badness-of-fit indices. The Tucker-Lewis Index (TLI) adjusts for the number of model parameters and ranges from 0–1. The Comparative Fit Index (CFI) assesses fit relative to a null model and ranges from 0 to 1. Both TLI and CFI indicate an acceptable fit when values are above 0.9. The lack of fit per degree of freedom in the model was expressed by the Root Mean Square Error of Approximation (RMSEA). The average of the differences between observed and predicted correlations was described with the standardized root mean square residual (SRMR). Both RMSEA and SRMR indicate a good fit with values below 0.08 [17, 26]. RMSEA and TLI are prone to false model rejections when the sample size is not adequate [25]. The sum score in the three samples was summarized using mean, SD, and range. Bootstrap methods were used to obtain estimates of the 95% CI of Cronbach's alpha, inter-item correlations, item-rest correlations, and SEM.

The reliability and agreement between tests were evaluated among the adolescents who completed both questionnaires. A Bland Altman plot with mean difference, confidence interval, and limits of agreement was generated to evaluate the extent of random and systematic error. The strength of the association between test and retest was assessed from the correlation between the test and retest. To assess the reliability of the scale, intraclass correlation (ICC(1,1)) and standard error of measurement (SEM) were estimated. The ICC[1, 1] assessed absolute agreement by a one-way random effect model for a single measurement, and the calculation of SEM was based on the same model. Agreement was estimated as the mean difference between the test and retest scores.

The convergent validity was tested with the CES-DC4 using Spearman's ρ on the sum scores of the scales and bootstrap methods to estimate the 95% CI. Spearman's ρ will only be based on adolescents who have answered all items in both CES-DC4 and SCL-6.

To investigate the effect of the timing of the tests in relation to the winter vacation, mean differences between the test and retest were estimated for each class separately as a sensitivity analysis. Furthermore, all reliability and agreement results were repeated on a subsample consisting only of the classes not expected to be influenced by the winter vacation.

Stata 17.0 software was used to perform most statistical analyses [27], while the CFA was performed in R version 1.2.5019 [28] and the R package lavaan [29].

Results

Participants

The total sample consisted of 122 adolescents in the five 9th grades. Adolescents were excluded from the study if age was less than 15 years, or the adolescent could not read, nor write Danish. Double indication or no indication in the questionnaire was recorded as a missing answer. Three samples were described; the "test sample" consisting of all the adolescents who completed the test questionnaire ($n=91$), the "retest sample" consisting of all the adolescents who completed the retest questionnaire ($n=82$), and the "test–retest sample" consisting of all the adolescents who completed both the test and the retest questionnaires ($n=66$) (Fig. 3).

Analyses of structural validity and internal consistency were both conducted on 91 adolescents (74.6%) in the test sample and 82 adolescents (67.2%) in the retest sample. Analyses of agreement and reliability were conducted on 66 adolescents (54.1%) in the test–retest sample (Fig. 3). The time interval between the two tests ranged from 14 to 18 days.

The descriptive information about the adolescents is presented in terms of age and sex (Table 1).

Analysis of incomplete sample

The analysis of the incomplete sample ($n=46$) showed no differences between adolescents excluded from the test–retest and adolescents included in the test–retest in terms of age, sex, and the individual scores for 5 out of 6 items in test and retest (data not shown). A difference in the answers for item 32 was observed between the incomplete sample ($n = 22$) and the test–retest sample ($n = 66$) in the retest ($p = 0.03$) The incomplete sample had a higher prevalence of the answer "Not at all" (86%) than the complete sample (64%) and a lower prevalence of "A little bit" (5%) than the complete sample (24%).

The mean sum score in the test, retest, test–retest, and incomplete sample were estimated (Table 2). The adolescents' sum score ranged from 0–18 out of the possible 0–24.

Internal consistency

The estimated Cronbach's α was within the recommended interval (0.7–0.9) with 0.80 (95% CI: 0.73; 0.86) at the test and 0.81 (95% CI: 0.73; 0.90) at the retest [17]. The inter-item correlation ranged from 0.26 (95% CI: 0.06; 0.46) to 0.52 (95% CI: 0.33; 0.65) at the test and 0.24 (95% CI: −0.01; 0.43) to 0.64 (95% CI: 0.46; 0.77) at the retest, and thereby supported the conclusion of a unidimensional scale. The item-rest correlation ranged between 0.47 (95% CI: 0.26; 0.68) and 0.64 (95% CI: 0.50; 0.78) at the test and between 0.46 (95% CI: 0.26; 0.67)

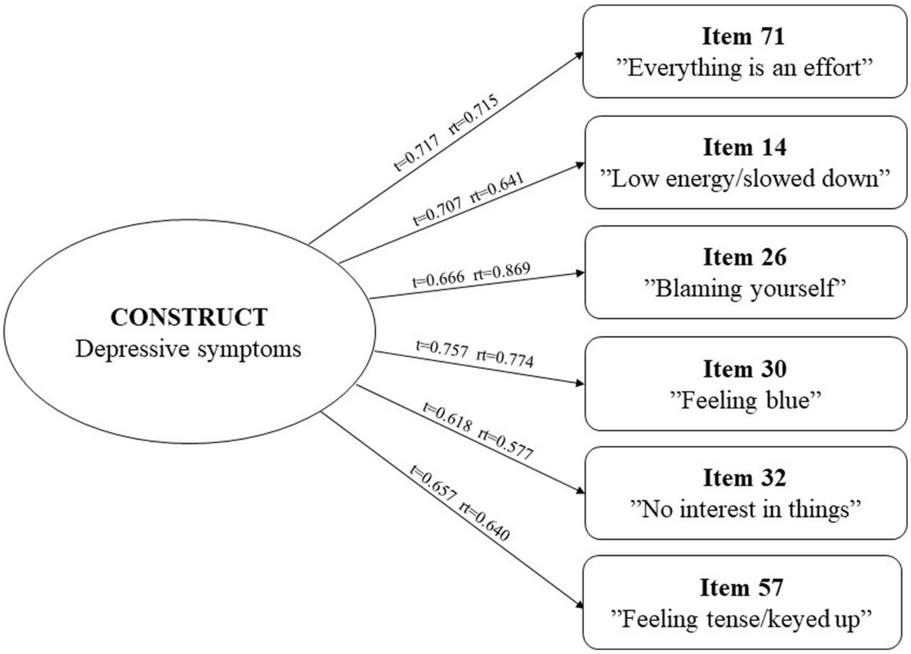


Fig. 3 Flowchart of participants

Table 1 Characteristics of the samples in the SCL-6 study

	Test ($n = 91$)	Retest ($n = 82$)	Test–retest ($n = 66$)	Incomplete ($n = 46$)
Age, 15/16–17 (%)	74/17 (81.3/18.7)	63/19 (76.8/23.2)	54/12 (81.8/18.2)	38/8 (82.6/17.4)
Sex, m/f (%)	49/41 (54.4/45.6)	41/40 (50.6/49.4)	36/30 (54.6/45.5)	21/23 (47.7/52.3)

Table 2 Characteristics of the samples in the SCL-6 study

	Test ($n = 91$)	Retest ($n = 82$)	Test–retest ($n = 66$)	Incomplete ($n = 46$)
Sum score mean test (SD) [range]	6.29 (4.52) [0–18]	N/A	6.41 (4.64) [0–18]	5.96 (4.26) [0–16][a]
Sum score mean retest (SD) [range]	N/A	5.57 (4.26) [0–17]	5.67 (4.13) [0–17]	5.66 (4.89) [0–17][b]

[a] $n = 25$

[b] $n = 16$

and 0.67 (95% CI: 0.54; 0.80) at the retest, indicating that every item contributed to a distinction between adolescents with low and high scores on the rest of the items. Analysis showed that deletion of items would result in a lower Cronbach's α, both at the test and the retest.

Structural validity

The fit indices are shown in Table 3. Model 1 had RMSEA above the recommended limit at both test and retest and had TLI and SRMR outside the recommended limit at retest.

The highest factor loadings were item 30 (0.76) at test and item 26 (0.87) at retest. The lowest loadings were for item 32 both at test (0.62) and retest (0.58) at retest. Thereby, item 30 contributed the most to the factor at test and item 26 contributed the most to the factor at retest, while item 32 contributed the least at both test and retest (Fig. 4). Mokken analyses showed good fit of the one factor solution with a coefficient of homogeneity above the recommended limit of 0.4 with 0.45 at test and 0.48 at retest [30].

Reliability

The estimated SEM was 2.03 (95% CI: 1.68; 2.37) and ICC[1, 1] was 0.79 (95% CI: 0.68; 0.86).

Agreement

Analysis of the mean showed a higher mean at the test than at the retest (mean difference = 0.74 (95% CI: 0.06; 1.43)). Sensitivity analysis showed that class 1 (*n* = 15) was the class with the biggest difference between the test and the retest (2.07 (0.63; 3.51)) and that class 4 and class

Table 3 Fit indices for confirmatory factor analysis models of SCL-6 at test and retest

Model fit indices	Comparative fit index (CFI)	Tucker-Lewis Index (TLI)	Root mean square error of approximation (RMSEA)	Standardized root mean square residual (SRMR)
One-factor model				
Test (*n* = 91)	0.97	0.96	0.10 (95% CI: 0.00; 0.17)	0.06
Retest (*n* = 82)	0.93	0.89	0.17 (95% CI: 0.11; 0.24)	0.09

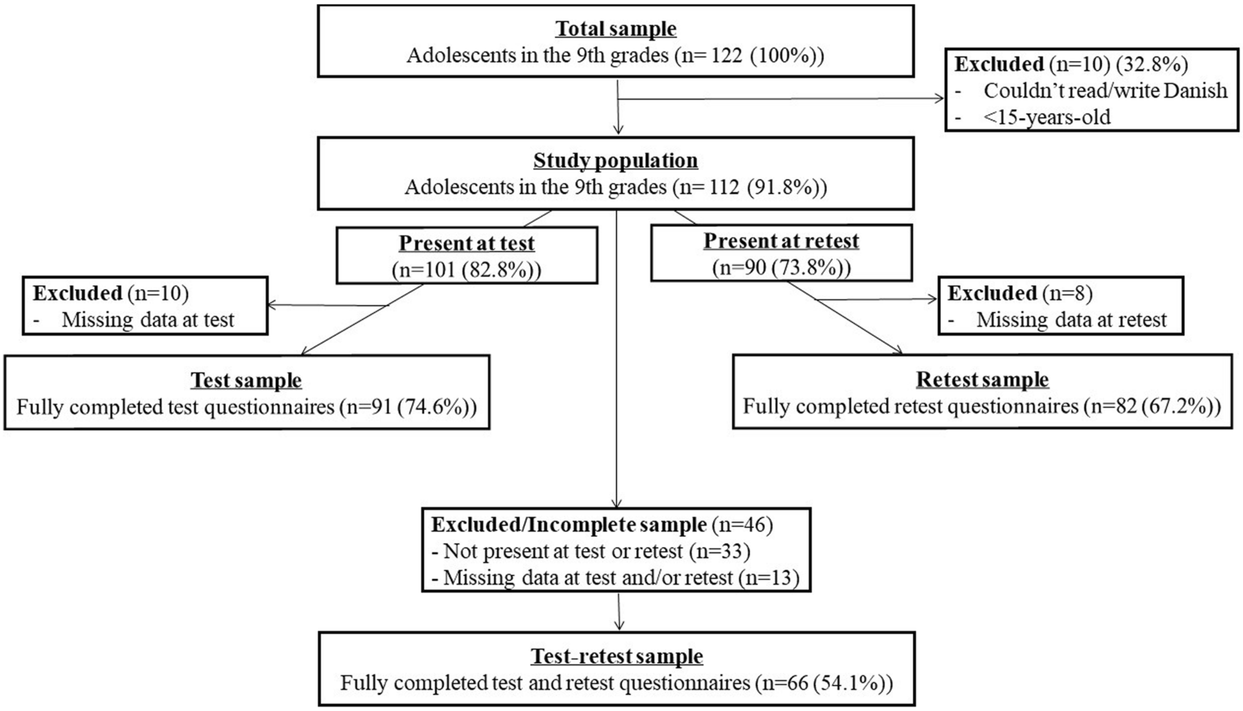

Total sample
Adolescents in the 9th grades (n= 122 (100%))
Excluded (n=10) (32.8%)
- Couldn't read/write Danish
- <15-years-old
Study population
Adolescents in the 9th grades (n= 112 (91.8%))
Present at test
(n=101 (82.8%))
Present at retest
(n=90 (73.8%))
Excluded (n=10)
- Missing data at test
Excluded (n=8)
- Missing data at retest
Test sample
Fully completed test questionnaires (n=91 (74.6%))
Retest sample
Fully completed retest questionnaires (n=82 (67.2%))
Excluded/Incomplete sample (n=46)
- Not present at test or retest (n=33)
- Missing data at test and/or retest (n=13)
Test-retest sample
Fully completed test and retest questionnaires (n=66 (54.1%))

Fig. 4 Factor structure and factor loadings for test (t) and retest (rt) of the SCL-6

5 had a lower mean at the test than at the retest (Supplementary Table 1).

The Bland Altman plot showed constant variation of the mean difference, no linear tendency, and showed an outlier from class 1 with a difference of 9 between the sum score at the test and the retest (Fig. 5). The LoA were −4.73 (95% CI: −5.89; −3.57) to 6.21 (95% CI: 5.06; 7.37), and the estimated correlation between the test and the retest was 0.80 (95% CI: 0.70; 0.88).

The horizontal solid dark green line is the mean difference in sum scores and the dashed dark green lines on each side are the 95% CI of the mean difference. The solid black lines are the Limits of Agreement.

Convergent validity

104 adolescents had completed all items in both SCL-6 and CES-DC4 and were included in the convergent validity analyses. The sum scores of the scales were not normally distributed. Therefore, Spearman's ρ method was used to evaluate the correlation between the two scales. The Spearman's ρ was 0.39 (95% CI: 0.20; 0.58) between sum scores at test ($p < 0.05$) and 0.54 (0.39; 0.69) between sum scores at retest ($p < 0.05$).

Discussion

This study is the first to evaluate the reliability, and structural and convergent validity of the Danish SCL-6 in adolescents. We found acceptable internal consistency and test–retest reliability of the SCL-6 in 15–17-year-old adolescents. The CFA did not find an acceptable fit of a one factor solution, while the Mokken analyses showed a good fit. The convergent validity with the CES-DC4 was moderate.

Measurement property evidence

The internal consistency of the scale was good. The Cronbach's α was good at both test (α = 0.80 (95% CI: 0.73; 0.86)) and retest (α = 0.81 (95% CI: 0.73; 0.90)), indicating that the items reflected the same construct. This mirrors the result in a study on the SCL-6 in the US (Cronbach's α = 0.88)[16]. The inter-item correlations supported the conclusion of a unidimensional scale, and thereby support findings from previously studies of a unidimensional scale [13, 15]. The present study was the first to report the item-rest correlations, which showed that all items contributed to distinct between adolescents who have high or low scores on the rest of the items. In the same population, the CES-DC4 showed poorer internal consistency with a Cronbach's α on 0.61 and lower inter-item and item-rest correlations compared with the SCL-6 [21].

Concerning the structural validity of the scale, RMSEA had values above the recommended limit of 0.08 at both test (0.10) and retest (0.17), indicating lack of fit per degree of freedom. At retest, both the TLI (0.89) and the CFI (0.09) were below the recommended limit of 0.9, overall indicating a poor fit to a unidimensional scale. That being said, all of the estimates are close to the recommended limits, and the acceptable limits for a good fit of a model are heavily discussed [31]. Furthermore, our sample size is in the smaller end of recommended sample sizes for CFA. Sample sizes of 100 and 200 are often recommended as the minimum acceptable sizes, but an even larger sample size would be needed when the items are categorical as in our study [32]. However, the Mokken analyses showed good structural validity in line with previous research using Mokken analyses [13, 14, 18].

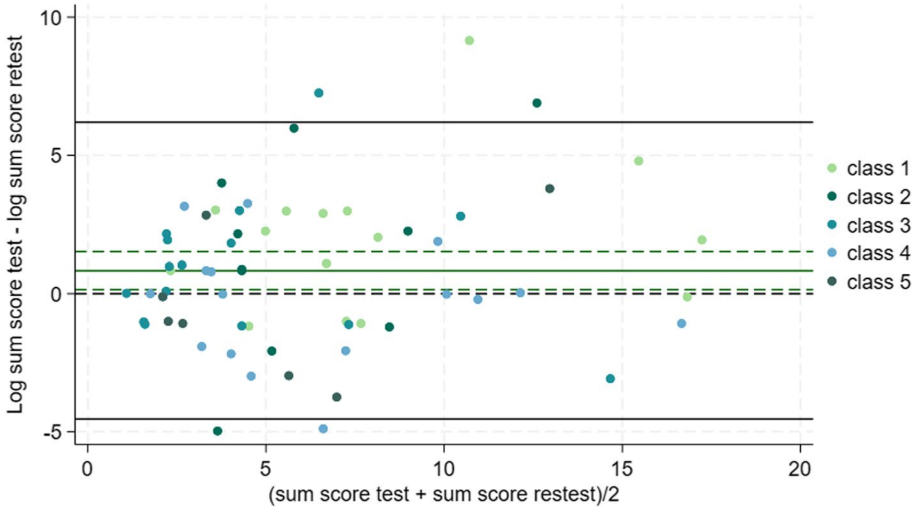


Fig. 5 Bland Altman plot

We found good reliability of the SCL-6 with an ICC (0.79) larger than the limit for "acceptable" of 0.7[17]. The CES-DC4 found a lower ICC (0.60) in the same population [21].

The mean of the sum score at the test was higher than the mean of the sum score at the retest (mean = 0.74 (95% CI: 0.06; 1.43)) implying a small systematic error. This was, however, driven by a large mean difference in class 1 (2.07 (95% CI: 0.63; 3.51)). Reasons for such an effect could be that something happened before the test that affected the answers negatively, or something happened before the retest that affected the answers positively. Class 1's retest was performed just before the start of winter vacation (Supplementary Fig. 3), and so the adolescents could have had their spirits heightened solely from the outlook of the winter vacation (Supplementary Table 1). Excluding class 1 improves the mean difference (0.35 (95% CI: −0.42; 1.13) and the SEM (1.94 (95% CI: 1.56; 2.32), while the ICC (0.77 (95% CI: 0.63; 0.86) is comparable with results on the full sample, and LoA (lower = −5.03 (−6.33; −3.74), upper = 5.74 (4.44; 7.04)) are more symmetrical around zero (Supplementary Table 2). Thus, the difference in the mean test scores is not expected to be a systematic error of the SCL-6 but more an artifact of the timing of the tests. In the same population, the CES-DC4 found a mean difference between test and retest on 0.39 (95% CI: −0.02; 0.80), which also indicated a slight tendency to a higher sum score in the test compared with the retest [21]. Future studies should account for vacation periods when planning data collection, ensuring that both the test and retest occur at least one week before or after any vacation period.

Both SCL-6 and CES-DC4 are designed to measure the same construct, so we expected a strong correlation between the scores from the two instruments but found only a moderate correlation ($\rho = 0.39–0.54$). This can be due to the poor reliability and high SEM of the CES-DC4 [21]. Both questionnaires measure the construct of depressive symptoms, however, the CES-DC4 has a question about the social context ("I felt like kids I know were not friendly or that they didn't want to be with me."), an aspect not covered by the SCL-6. The SCL-6 has more elaborate questions about the three core symptoms of depression: mood, reduced energy, and reduced desire. The CES-DC4 has two questions regarding mood, of which one of them has positive wording ("I was happy") that can lead to different interpretations of the two questionnaires from the recipient, since all questions in SCL-6 are phrased in negative wording. While the SCL-6 has two questions about energy ("Feeling everything is an effort" & "Feeling low in energy or slowed down"), the question about energy in the CES-DC4 has previously shown problems, as adolescents describe this to be a normal part of being a teenager ("It was hard to get started doing things.") [21]. The authors believe that the moderate correlation can be explained by the poor reliability of the CES-DC4 and differences in the constructs of the questionnaires, with the CES-DC4 placing less emphasis on reduced energy and reduced desire.

Practical relevance

The LoA were broad (−4.73 to 6.21) and the SEM (2.03 (95% CI: 1.68; 2.37)) was relatively high considering the range of sum scores used from 0–18. Therefore, we do not recommend using the scale to measure changes in the depressive symptoms at an individual level as the sum score of the scale is attached with a lot of uncertainty. Consistent with this, we do not recommend using the scale as a screening tool for individual-level clinical diagnostics. We do not believe the observed mean difference between test and retest scores has practical implications for the use of SCL-6 in adolescents. However, it is important to consider that adolescence is a developmental stage characterized by emotional tumult [33]. Daily mood fluctuations, typically for this age group, may influence how adolescents respond to questionnaire items about depressive symptoms. Generally, the psychometric properties were better in the SCL-6 than in the CES-DC4 in this population and therefore, the SCL-6 is preferable over the CES-DC4 to measure depressive symptoms in Danish adolescents in a population-based setting.

Strengths and limitations

Since the adolescents were anonymous, misclassification is not suspected. A total of 54.1% of the entire sample was included in the test–retest sample. The excluded fraction is expected to be missing at random since the analysis of incomplete showed no differences between excluded and included adolescents in most cases (data not shown). One difference between the incomplete sample and the test–retest sample was observed in item 32 score in the retest ($p = 0.03$), but we expect the difference to be due to chance because of small sample size in the retest in the incomplete sample ($n = 22$). Most adolescents were excluded because they were not present at the test or the retest (32.0%). Adolescents excluded due to missing answers (10.7%) were not related to a specific item. The risk of selection bias is therefore considered to be minimal.

The adolescents were answering both questionnaires at the same setting; in their classroom sitting at their seats. The authors encouraged the teachers not to change the seats of the adolescents between the test and retest. Moreover, the authors strived for placing the test and retest at the same weekday and at the same time of the day. Thereby, the setting has a minimal influence on the adolescents' answers of the questionnaires.

The adolescents could change token with another student without our notice and thereby the connection of test and retest could be biased. However, the adolescents would have no incentive to do so, and we therefore do not expect this bias.

Generalizability

Only three out of the 71 contacted schools agreed to participate in the study. In Denmark, schools are often contacted for different investigations, and therefore, the schools must decline many inquiries. For this reason, we expected a low number of participating schools but also the participation to be random. Moreover, the three participating schools consisted of 9th grades in both a larger city (> 350.000 inhabitants) and a smaller town (< 7.000 inhabitants) which improves the representativeness. The results are therefore expected to be transferable to regular 9th grades in the rest of Denmark.

To achieve higher participation rates in future studies, involving a senior researcher in outreach efforts, rather than a master's student, could increase the likelihood of schools agreeing to participate. Although schools were offered a lecture on mental health for their students, none took advantage of this opportunity, but other compensations could be offered. In questionnaire studies, financial compensation has been shown to positively impact participation rates [34]. However, offering an economic incentive to schools is not legal in Denmark and therefore not an option. Some schools indicated difficulty finding time to participate due to scheduling all lectures for the entire academic year in advance. Our contact with schools occurred in December and January; initiating outreach earlier, before the school year's schedule is finalized, may improve participation rates.

Expanding outreach to include a broader and more diverse range of schools, particularly those located farther from universities who are potentially less frequently contacted, may also increase the number of participants. A randomized selection of schools could be used to ensure the generalizability of results. However, it is essential to ensure physical presence during questionnaire completion to maintain procedural accuracy. Implementing such an approach would require additional resources.

Instrument changes

During data collection, the author was met with questions about the meaning of the wording in item 26 (in Danish: "selvbebrejdelse", in English: "blaming yourself for things"). Therefore, the content validity and the cultural adaption of the scale, and especially item 26, in this target group should be reevaluated.

Further, some measurement properties still need to be investigated. Adolescence can be an age with a lot of emotional tumult, and therefore it would be relevant to assess the construct validity to determine if the scale captures the construct of depressive symptoms rather than emotional tumult [33]. The result from class 1 also suggests that the test scores are sensitive to fluctuations in everyday life such as an upcoming vacation.

Future research

The SCL-6 shows promising reliability results. Further studies are warranted to both address the timing of the test and retest and ensure minimal influence from unrelated events, but also larger studies to thoroughly investigate the structural validity. Lastly, the cultural adaption of the scale to adolescents should be revisited.

Conclusion

This study found acceptable internal consistency and test–retest reliability of the SCL-6 in 15–17-year-old adolescents. The mean test score was higher than the mean retest score, but after further analyses of the mean difference, systematic error of the scale is not suspected. The scale is constructed as and considered unidimensional. The results from the CFA were ambiguous in terms of demonstrating unidimensionality, and the structural validity should be further investigating in a larger study population. The authors find the scale usable for future population-based research on depressive symptoms in adolescents. Taken the broad LoA into account, we do not recommend using the scale as screening tool for depression in individuals.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12887-025-05510-1.

Supplementary Material 1.

Acknowledgements

The author expresses gratitude to the teachers who dedicated part of their lessons to support the study and to the adolescents who took part in it.

Clinical trial number

Not applicable.

Authors' contributions

All authors contributed to conceptualization and methodology of the study. CLBS was responsible for software, formal analysis, investigation, resources, data curation, visualization, project administration, and writing the draft. KB and TKG contributed with validation, writing and editing of the article and supervision. All authors have read and agreed to the published version of the manuscript.

Funding
There was no external funding for this manuscript, as it was conducted in continuations of a Master's thesis.

Data availability
Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate
The study adhered to the principles outlined in the Declaration of Helsinki. In Denmark, Ethics Committee approval is not required for questionnaire-based research, as stipulated by Danish law. Detailed guidelines can be found in § 14,2 of the "Act on Research Ethics Review of Health Research Projects," available on the National Committee on Health Research Ethics' website (https://researchethics.dk/). According to Danish law, parental consent is mandatory only for clinical trials, not for questionnaire-based studies like this one unless the adolescent is below age 15. Therefore, adolescents below age 15 was excluded from the study. Verbal informed consent was obtained from all adolescent participants [35].

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 October 2024 Accepted: 12 February 2025
Published online: 17 March 2025

References
1. World Health Organization. Depression: World Health Organization; 2021 [updated 13/09/2021. Available from: https://www.who.int/news-room/fact-sheets/detail/depression.
2. Sartorius N, Ban TA. Assessment of depression. Berlin: Springer; 1986. p. 376.
3. Bech P. Clinical psychometrics. 1. ed. Chichester, West Sussex: Wiley-Blackwell; 2012. xi, p. 202.
4. Fröjdh K, Håkansson A, Karlsson I. The Hopkins Symptom Checklist-25 is a sensitive case-finder of clinically important depressive states in elderly people in primary care. Int J Geriatr Psychiatry. 2004;19(4):386–90.
5. Nabbe P, Le Reste J, Guillou-Landreat M, Assenova R, Kasuba Lazic D, Czachowski S, et al. Nine forward-backward translations of the Hopkins Symptom Checklist-25 with cultural checks. Front Psych. 2021;12:688154.
6. Nabbe P, Le Reste J, Guillou-Landreat M, Gatineau F, Le Floch B, Montier T, et al. The French version of the HSCL-25 has now been validated for use in primary care. PLoS ONE. 2019;14(4): e0214804.
7. Glaesmer H, Braehler E, Grande G, Hinz A, Petermann F, Romppel M. The German Version of the Hopkins Symptoms Checklist-25 (HSCL-25) –factorial structure, psychometric properties, and population-based norms. Compr Psychiatry. 2014;55(2):396–403.
8. Ashaba S, Kakuhikire B, Vořechovská D, Perkins J, Cooper-Vince C, Maling S, et al. Reliability, validity, and factor structure of the Hopkins Symptom Checklist-25: population-based study of persons living with HIV in Rural Uganda. AIDS Behav. 2018;22(5):1467–74.
9. Lundin A, Hallgren M, Forsell Y. The validity of the symptom checklist depression and anxiety subscales: a general population study in Sweden. J Affect Disord. 2015;183:247–52.
10. Haavet O, Sirpal M, Haugen W, Christensen K. Diagnosis of depressed young people in primary health care–a validation of HSCL-10. Fam Pract. 2011;28(2):233–7.
11. Syed H, Zachrisson H, Dalgard O, Dalen I, Ahlberg N. Concordance between Hopkins Symptom Checklist (HSCL-10) and Pakistan Anxiety and Depression Questionnaire (PADQ), in a rural self-motivated population in Pakistan. BMC Psychiatry. 2008;8:59.
12. Kleppang A, Hagquist C. The psychometric properties of the Hopkins Symptom Checklist-10: a Rasch analysis based on adolescent data from Norway. Fam Pract. 2016;33(6):740–5.
13. Magnusson Hanson L, Westerlund H, Leineweber C, Rugulies R, Osika W, Theorell T, et al. The Symptom Checklist-core depression (SCL-CD6) scale: psychometric properties of a brief six item scale for the assessment of depression. Scandinavian journal of public health. 2014;42(1):82–8.
14. Christensen K, Haugen W, Sirpal M, Haavet O. Diagnosis of depressed young people–criterion validity of WHO-5 and HSCL-6 in Denmark and Norway. Fam Pract. 2015;32(3):359–63.
15. Bech P, Bille J, Møller S, Hellström L, Østergaard S. Psychometric validation of the Hopkins Symptom Checklist (SCL-90) subscales for depression, anxiety, and interpersonal sensitivity. J Affect Disord. 2014;160:98–103.
16. Alvir J, Schooler N, Borenstein M, Woerner M, Kane J. The reliability of a shortened version of the SCL-90. Psychopharmacol Bull. 1988;24(2):242–6.
17. Vet HCWd. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011. x, p. 338 s.
18. Olsen L, Mortensen E, Bech P. The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. Acta Psychiatr Scand. 2004;110(3):225–9.
19. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539–49.
20. Mokkink LB, Prinsen CA, Patrick DL, Alonso J, Bouter LM, de Vet HC, et al. COSMIN Study Design checklist for Patient-reported outcome measurement instruments. location VUmc: Department of Epidemiology and Biostatistics Amsterdam Public Health research institute Amsterdam University Medical Centers. 2019. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf.
21. Sørensen CLB, Grønborg TK, Biering K. Reliability and structural validity of the Danish Short 4-item version of the Center for Epidemiological Studies Depression Scale for Children (CES-DC4) in adolescents. BMC Pediatr. 2022;22(1):388.
22. Sørensen CLB, Grønborg T. K., Biering K. The Danish short 4-item version of the Center for Epidemiological Studies Depression Scale for Children (CES-DC4): a test-retest reliability study of 15–17-year-olds adolescents (under review). Inquiry. 2021.
23. Lipman R, Covi L, Shapiro A. The Hopkins Symptom Checklist (HSCL)–factors derived from the HSCL-90. J Affect Disord. 1979;1(1):9–24.
24. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.
25. Kirkwood BR, Sterne JAC. Essential medical statistics. 2. edition, reprinted ed. Malden, Mass.: Blackwell Science; 2003. x, p. 501 sider.
26. Lt Hu, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model: Multidiscip J. 1999;6(1):1–55.
27. StataCorp. Stata statistical software: release 16. College Station. TX: StataCorp LLC; 2019.
28. The R Foundation. The R project for statistical computing: the R foundation; 2021. Available from: https://www.r-project.org/.
29. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48(2):1–36.
30. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks: SAGE Publications; 2002.
31. Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods. Behav Res Methods. 2019;51(1):409–28.
32. Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. Educ Psychol Meas. 2013;76(6):913–34.
33. Olsson G, von Knorring AL. Depression among Swedish adolescents measured by the self-rating scale Center for Epidemiology

Studies-Depression Child (CES-DC). Eur Child Adolesc Psychiatry. 1997;6(2):81–7.

34. Smith MG, Witte M, Rocha S, Basner M. Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. BMC Med Res Methodol. 2019;19(1):230.
35. Nationalt Center for Etik. Vejledning om det informerede og stedfortrædende samtykke i sundhedsvidenskabelige forskningsprojekter [Guidance on informed and surrogate consent in health science research projects]. 2024. Available from: https://videnskabsetik.dk/vejledninger/vejledning-om-det-informerede-og-stedfortraedende-samtykke-i-sundhedsvidenskabelige-forskningsprojekter.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.