Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility

Emily Stockings, Louisa Degenhardt, Yong Yi Lee, Cathrine Mihalopoulos, Angus Liu, Megan Hobbs, George Patton

PII: S0165-0327(14)00785-X
DOI: http://dx.doi.org/10.1016/j.jad.2014.11.061
Reference: JAD7152

To appear in: Journal of Affective Disorders

Received date: 26 August 2014
Revised date: 28 November 2014
Accepted date: 29 November 2014

Cite this article as: Emily Stockings, Louisa Degenhardt, Yong Yi Lee, Cathrine Mihalopoulos, Angus Liu, Megan Hobbs, George Patton, Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility, Journal of Affective Disorders, http://dx.doi.org/10.1016/j.jad.2014.11.061

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility

Emily Stockings*1 PhD, Louisa Degenhardt1 PhD, Yong Yi Lee2 MHEcon, Cathrine Mihalopoulos3 PhD, Angus Liu1 MPH, Megan Hobbs4 PhD, George Patton5 MD.
1National Drug and Alcohol Research Centre, University of New South Wales, 22-32 King Street, Randwick, New South Wales, Australia 2031.
2School of Population Health, University of Queensland, Herston Road, Herston, Queensland, Australia, 4006.
3Population Health Strategic Research Centre, Deakin University, 211 Burwood Highway, Burwood, Victoria, Australia, 3125.
4Clinical Research Unit for Anxiety and Depression (CRUfAD), University of Newcastle, St Vincent’s Hospital, 390 Victoria Street, Darlinghurst, 2010.
5Centre for Adolescent Health, The Royal Children’s Hospital, 50 Flemington Road, Parkville, Victoria, 3025.

*Corresponding author: Emily Stockings, Email: e.stockings@unsw.edu.au Phone: (02)9385 1062 Address: National Drug and Alcohol Research Centre (NDARC), Building R3, 22-32 King Street, Randwick, New South Wales, Australia, 2031.

Abstract

Background: Depression symptom screening scales are often used to determine a clinical diagnosis of major depressive disorder (MDD) in prevention research. The aim of this review is to systematically examine the reliability, validity and diagnostic utility of commonly used screening scales in depression prevention research among children and adolescents.

Methods: We conducted a systematic review of the electronic databases PsycINFO, PsycEXTRA and Medline examining the reliability, validity and diagnostic utility of four commonly used depression symptom rating scales among children and adolescents: the Children's Depression Inventory (CDI), Beck Depression Inventory (BDI), Center for Epidemiologic Studies - Depression Scale (CES-D) and the Reynolds Adolescent Depression Scale (RADS). We used univariate and bivariate random effects models to pool data and conducted metaregression to identify and explain causes of heterogeneity.

Results: We identified 54 studies (66 data points, 34,542 participants). Across the four scales, internal reliability was 'good' (pooled estimate: 0.89, 95% Confidence Interval (CI):
0.86 to 0.92). Sensitivity and specificity were 'moderate' (sensitivity: 0.80, 95% CI: 0.76 to 0.84; specificity: 0.78, 95% CI: 0.74 to 0.83). For studies that used a diagnostic interview to determine a diagnosis of MDD, positive predictive power for identifying true cases was mostly poor. Psychometric properties did not differ on the basis of study quality, sample type (clinical vs. nonclinical) or sample age (child vs. adolescent).

Limitations: Some analyses may have been underpowered to identify conditions in which test performance may vary, due to low numbers of studies with adequate data.

Conclusions: Commonly used depression symptom rating scales are reliable measures of depressive symptoms among adolescents; however, using cutoff scores to indicate clinical levels of depression may result in many false positives.

Keywords: Depression, Children, & Adolescents, Psychometrics, Validity, Psychiatric Symptom Rating Scales, Prevention.

Abbreviations:
ADIS-C: Anxiety Disorders Interview Schedule for Children
AUC: Area under the curve
BDI: Beck Depression Inventory
BSI: Brief Symptom Inventory
BYI: Beck Youth Inventories
CBCL: the Child Behaviour Checklist
CDI: Children’s Depression Inventory
CDRS-R: Children’s Depression Rating Scale – Revised
CES-D: Center for Epidemiologic Studies- Depression Scale
CES-DC: Center for Epidemiologic Studies – Depression Scale for Children
CIDI: Composite International Diagnostic Interview
DAWBA: The Development and Wellbeing Assessment
DEPS-10: Depression Scale - Version 10
DICA-IV: Diagnostic Interview for Children and Adolescents- Version 4
DISC: The National Institute of Mental Health Diagnostic Interview Schedule for Children
DSM: Diagnostic and Statistical Manual
DSRS: Depression Symptom Rating Scale (DSRS)
HADS: Hospital Anxiety and Depression Scale
ICD: International Classification of Diseases
KID-SCID: Structured Clinical Interview for DSM-IV disorders – Child version
Kinder DIPS: Diagnostisches Interview bei psychischen Störungen im Kindes- und Jugendalter [German]
K-SADS: Kiddie-Schedule for Affective Disorders and Schizophrenia
MDD: Major depressive disorder
MDI-C: Multi-score Depression Inventory-Children
MINI: Mini-International Neuropsychiatric Interview
MINI-KID: The Mini-International Neuropsychiatric Interview for Children
NPV: Negative predictive value
PPV: Positive predictive value
PRIME-MD: The Primary Care Evaluation of Mental Disorders
RADS: Reynolds Adolescent Depression Inventory
RCDAS: Revised Child Anxiety and Depression Scale
ROC: Receiver operator characteristic
SBB-DES: The Self-Report Questionnaire—Depression [German]
SCAN: Schedules for Clinical Assessment in Neuropsychiatry
SCID: Structured Clinical Interview for Depression for DSM disorders
SCID-1: Structured Clinical Interview for DSM-IV Axis I disorders
SDQ: Strengths and Difficulties Questionnaire
SMFQ: Short Mood and Feelings Questionnaire
YSR: Youth Self-Report
Introduction

Major depressive disorder (MDD) is the leading global cause of disability among young people aged 10-24 years, accounting for 8.2% of the global non-fatal disease burden (Gore et al., 2011). Approximately 3% of children and 6% of adolescents suffer current or recent depression (Costello et al., 2006). MDD in young people is associated with poor academic performance, substance abuse, attempted and completed suicide and an increased risk of suffering depression during adulthood (Birmaher et al., 1996, Brent et al., 1986). Despite the significant health burden associated with MDD, studies have suggested that less than 50% of youths seek mental health treatment for the condition (Reavley et al., 2010, Leaf et al., 1996). Valid and accurate screening tools for depression may assist clinicians in identifying MDD in youths, and may subsequently increase the rates of appropriate treatment and referral (Hosman et al., 2005, Andrews et al., 2002).

Given the significant health burden associated with MDD, there has been growing recognition of the need to develop programs aiming to prevent the onset of MDD during childhood and adolescence (Hosman et al., 2005, Andrews et al., 2002). Promisingly, the number of studies that have examined the efficacy of preventative interventions for MDD among children and adolescents more than doubled between 2004 and 2010 (Merry et al., 2004, Merry et al., 2011). Such interventions have typically been delivered in the school setting during regular classes by teachers or trained external facilitators (Merry et al., 2011). Given that routinely administering structured or semi-structured diagnostic interviews in schools can be costly and time consuming, many trials have used categorical thresholds on MDD symptom screening scales, such as the Center for Epidemiologic Studies – Depression Scale (CES-D) (Radloff, 1977) as a proxy for a diagnosis of MDD (Merry et al., 2011). Evaluating the efficacy of preventative interventions using symptom screening scales is problematic in two ways. Firstly, symptom screening scales that impose a categorical
threshold over a larger number of symptoms that are included in the DSM and the ICD can increase the number of cases that are identified compared to the cases where DSM or ICD diagnoses were applied (Ferrari et al., 2013).

Secondly, while the reliability, validity and utility of identifying cases of MDD using the CES-D and other symptom screening scales have been established in adult populations (Radloff, 1977), the same clinical thresholds have been applied among childhood and adolescent samples. No review has systematically examined whether the clinical thresholds identified in adult samples are reliable, valid or useful for identifying cases of MDD among children and adolescents. There is also no review that has quantitatively synthesized the traditional psychometric properties of symptom screeners for MDD among children and adolescents. Indeed, previous reviews examining the traditional psychometric characteristics of depression symptom screening scales among children and adolescents have been non-systematic and qualitative in nature, the most comprehensive of which were conducted more than a decade ago (Myers and Winters, 2002b, Brooks and Kutcher, 2001). The absence of robust data in support of applying the adult-derived thresholds for MDD when assessing for MDD among children and adolescents means that it is difficult to interpret the efficacy, effectiveness and efficiency of early-life preventative interventions for MDD meaningfully.

The purpose of this review is to systematically: 1) identify symptom screening scales that are commonly used in childhood and adolescent preventative interventions for MDD; and 2) identify evidence of reliability, validity and diagnostic utility of these symptom screening scales.
Methods

Search 1: Identify symptom screening scales that are commonly used in childhood and adolescent preventative interventions for MDD.

a) Search strategy

Given the large number of randomized controlled trials examining the efficacy of preventive interventions for MDD among children and adolescents, and the existence of multiple reviews synthesizing these findings (e.g. (Merry et al., 2004, Merry et al., 2011)), we conducted a systematic review of reviews, and updated recent reviews with any studies which may have been published subsequent to their publication date. The review of reviews was conducted in August, 2013 using the online databases PubMed, Medline, PSYCINFO and the Cochrane Library of Systematic Reviews, in consultation with a librarian and in accordance with the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) statement (Liberati et al., 2009). Databases were searched using a combination of MeSH terms and text words pertaining to depression and dysthymia (Depressive Disorder, Major Depression, Dysthymic Disorder, “dysthymia.mp.”), prevention (Primary Prevention, Preventative Psychiatry, “prevention.mp.”) and intervention trials (Intervention Studies, “intervention.mp.”). An additional search of empirical studies dated from August 2010-present was conducted in January 2014 to identify recently published randomised controlled trials not included in the existing reviews. This additional search was conducted in the electronic databases PubMed, Medline and PSYCINFO using the search string (((depress* OR dysthymi*)) AND (child* OR adolescen*)) AND (prevent* OR early intervention* OR risk OR at-risk OR vulnerab*)) AND (randomised controlled trial OR controlled trial [Publication Type]).
b) Inclusion criteria

i) Reviews: Reviews were eligible for inclusion if: 1) they were published between 2003 and August 2013 in the English language; 2) the authors employed systematic methods of reviewing the literature; and 3) the data were reported in a usable form, and:

ii) Empirical studies: Individual studies identified within the reviews, and the updated search of empirical studies were eligible for inclusion if: 4) the participants were aged between 5 and 18 years; 5) participants were randomly assigned to an intervention or a control group; 6) the control group received no intervention, placebo or usual care; and 7) the intervention aimed to prevent of the onset of depression or dysthymia.

c) Data extraction

Data were extracted from each empirical study by co-author ES and were cleaned and double checked by co-author YL and included: 1) sociodemographics of the sample; 2) details of the intervention and delivery; 3) timing of follow-up assessments; 4) the methods used to measure the prevalence, or symptoms of depression or dysthymia, including depression symptom screening scales, or structured diagnostic interviews; and 5) outcome data. Symptom screening scales were defined as any type of measure that provided an assessment of MDD, producing a numerical score that has guidelines for interpretation, whether completed by the person of interest (self report) or someone else (e.g. parent, guardian, teacher), with any type of response format (Myers and Winters, 2002a). For the purposes of this review, depression symptom screening scales were defined as ‘common’ if they were used in more than two studies. Scales used in less than two studies are mentioned only briefly here.
Search 2: Identify evidence of reliability, validity and diagnostic utility of commonly used depression symptom screening scales used in childhood and adolescent preventative interventions for MDD.

a) Search strategy

For each symptom screening scale for MDD that was identified as commonly used in Search 1, evidence for each scale’s reliability, validity and diagnostic utility in any child or adolescent sample (not limited to samples where preventive interventions were applied) was obtained through a series of additional systematic searches of the literature using PsycINFO, PsycEXTRA and Medline and a combination of MeSH terms and text words pertaining to depression (exp Major Depression), children (children OR adolesc~) symptom screening scales (exp Rating scales/exp Screening Tests/exp Psychological Tests/exp Psychiatric Status Rating scales), and psychometric properties (exp Psychometrics/exp Test Validity/exp Test Reliability/exp Diagnostic Utility), and the name of each individual scale as text words (e.g “children’s depression inventory”, “center for epidemiologic studies depression scale”, “beck depression inventory”). Other sources of literature included existing reviews of depression scales (Brooks and Kutcher, 2001, Brooks and Kutcher, 2003, Pavuluri and Birmaher, 2004, Suzuki, 2011, Collett et al., 2003), and hand searches of articles identified as relevant.

b) Eligibility criteria

Articles reporting on the scales’ reliability, validity and diagnostic utility were eligible for inclusion in the review if the study was: 1) published between 1980 (following publication of the initial validation studies) and August 2013 in the English language; 2) the participants were aged between 5 and 18 years; 3) the scale was identified in Search 1; and 4) the study reported the traditional psychometric properties of the scale including measures of reliability.
and/or validity. We excluded any revisions or short versions of the scales if they were not analogous to the original scales (i.e. differed in terms of number of items, scoring method, etc.). Studies attempting to translate an existing scale into another language without reporting psychometric properties of the scale itself were excluded.

c) Data extraction

Data from each included study were extracted by co-author AL and were cleaned and double checked by co-author ES and included: 1) characteristics of the sample (including sample size, age, gender, country, language and setting); 2) the names of the scales assessed and the number of items; 3) the test-retest and inter-rater reliability of the scale; 4) the internal validity of the scale (e.g. Cronbach’s alpha); 5) details about the external validity of the scale including the categorical thresholds used to indicate cases of MDD, the type of clinical interview used to determine cases of MDD (i.e. the ‘criterion standard’), and the associated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and area under the curve (AUC) results derived from receiver operating characteristics (ROC) analyses (described below). Sensitivity, specificity, PPV and NPV were calculated manually when data were partially reported, and could be imputed based on the presentation of other data.

Primary outcomes: Traditional psychometric properties and diagnostic utility

a) Data synthesis

Data were synthesised via meta analysis using the statistical software program Stata/SE version 13.1 (StataCorp, 2013). We used univariate random effects models to pool data
across all studies for the primary outcomes internal reliability and AUC. Bivariate random effects models (using the ‘mvmeta’ command) were used to pool data for the primary outcomes sensitivity and specificity in order to take into account the interdependency of these values. Values for PPV and NPV were not meta analysed and were synthesised in a qualitative manner only, as these values are highly dependent on the prevalence of depression in each sample, which was likely to vary substantially between each study included in this review (Leeflang et al., 2012). In order to identify and explain any causes of heterogeneity, random effects metaregression (using the ‘metareg’ command) was performed on the primary outcomes (with the exception of PPV and NPV) for each of the four scales, and across the four scales overall. The effects of the sample setting (clinical vs. nonclinical), age (children vs. adolescents), risk of bias quality score, the cutoff score used, and the scale (CDI, BDI, CES-D, RADS) were evaluated by individually adding them as covariates in the regression models. Due to the small number of studies, these analyses were primarily exploratory.

Additionally, the $I^2$ index was employed to quantify any heterogeneity in the pooled estimates, and was described as low, moderate or high according to an $I^2$ value of 25, 50 and 75%, respectively (Higgins et al., 2003). Statistical significance for all analyses was set at $p < .05$.

b) Reliability

The internal consistency of each scale was assessed using Cronbach’s alpha [$\alpha$], and test retest and interrater reliability were assessed using the Kappa [$\kappa$] statistic. These reliability coefficients were classified as ‘excellent’ if $\alpha \geq .9$, ‘good’ if $.85 \leq \alpha < .9$, ‘moderate’ if $.80 \leq \alpha < .85$, ‘fair’ if $.75 \leq \alpha < .80$, or ‘unsatisfactory’ if $\alpha < .75$ (Ponterotto and Ruckdeschel, 2007).
c) Validity

To examine the validity of each scale, we assessed five key measures of discriminative validity: 1) sensitivity (the proportion of true cases correctly identified, or “true positive rate”); 2) specificity (proportion of non-cases correctly identified, or “true negative rate”); 3) PPV (probability that subjects identified as cases are true cases, or “precision”); and 4) NPV (probability that the subjects identified as not being a case are not cases) (Murphy and Davidshofer, 1994). For inclusion of studies in the meta analysis, we assigned a minimum quality assessment for the criterion standard as diagnoses made using either a standardised (e.g. the Composite International Diagnostic Interview (CIDI)) or structured (e.g. Kiddie-Schedule for Affective Disorders and Schizophrenia (K-SADS)) diagnostic interview schedule based on DSM or ICD-10 diagnostic criteria. Clinician rated diagnoses without reference to a specific diagnostic instrument were excluded from the analyses. Based on expert statistical recommendation, we assigned a value of ‘excellent’ for the scales’ sensitivity, specificity, PPV and NPV if the pooled value of coefficient across studies was ≥ .9, a value of ‘good’ if .8 ≤ coefficient < .9, a value of ‘moderate’ if 0.6 ≤ coefficient < 0.8, and a value of ‘low’ coefficient < .6 (Andrews et al., 1993). We additionally assessed the scale’s diagnostic accuracy using the area under the curve (AUC) receiver operating characteristic (ROC) analysis. ROC curves plot the tradeoff between sensitivity and specificity values for all possible cutoff scores, and the AUC is a measure of the scales’ overall ability to correctly classify individuals as cases or noncases (Brooks and Kutcher, 2001). This is calculated as the sum of true positives and the sum of true negatives divided by the total population (not taking into account false positive or negatives). AUC values range from 0.5 (random classification) to 1 (perfect classification). Henderson (Henderson, 1993)
recommended the following interpretation with respect to the scales’ diagnostic accuracy: AUC > 9 ‘high’; 0.7 < AUC ≤ 0.9 ‘moderate’; AUC ≤ 0.7 ‘low’.

Risk of bias

Risk of bias in the included studies was examined using the risk assessment tool developed by the Cochrane Collaboration Diagnostic Test Accuracy Working Group (Reitsma et al., 2009), derived from the Revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Whiting et al., 2011b). Risk was determined across four domains: patient selection, index test, reference standard, and flow and timing, and was categorised and quantified as either low (3), unclear (2) or high (1) in order to derive a total quality score for use in the metaregression analyses.

Results

Search 1: Identify commonly used depression symptom screening scales in childhood and adolescent preventative interventions for MDD.

Search 1 identified 17 systematic reviews of childhood and adolescent preventative interventions for MDD. These 17 reviews contained a total of 394 individual trials, of which 288 were duplicates (in instances where the same trial appeared in more than one review) and 106 were unique.

Of the 106 unique trials, 103 assessed depressive symptomatology as an outcome using depression symptom screening scales, and 32 assessed the prevalence of MDD using either cutoff scores on depression symptom screening scales (n = 15) or by administering structured or semistructured clinical interviews (n = 17). Across the 103 trials, a total of 17 different depression symptom screening scales, and five structured or semistructured clinical interviews were used. The most commonly used screening scale for assessing depressive
symptomatology was the Children’s Depression Inventory (CDI; 45 studies), followed by the Beck Depression Inventory (BDI; 16 studies), the Center for Epidemiologic Studies – Depression Scale (CES-D; 15 studies), and the Reynolds Adolescent Depression Scale (RADS; 5 studies). The following scales were used as outcome measures in two studies or fewer: the Brief Symptom Inventory (BSI), the Child Behaviour Checklist (CBCL), the Children’s Depression Rating Scale – Revised (CDRS-R), the Depression Symptom Rating Scale (DSRS), Beck Youth Inventories (BYI), the Depression Scale version 10 [Finnish] (DEPS-10), the Hospital Anxiety and Depression Scale (HADS), the Multi-score Depression Inventory – Children (MDI-C), the Revised Child Anxiety and Depression Scale (RCADS), Youth Self-Report (YSR), and The Self-Report Questionnaire—Depression [German] (Selbstbeurteilungsbogen—Depressive Störungen; SBB–DES). Two studies were found to have used undefined composite constructs. The most commonly used measure to determine the presence of MDD was the Kiddie-Schedule for Affective Disorders and Schizophrenia (K-SADS; 7 studies), followed by the Anxiety Disorders Interview Schedule for Children (ADIS-C; 5 studies), the Diagnostic Interview for Children and Adolescents version 4, (DICA-IV; 1 study), and the Structured Clinical Interview for Depression for DSM disorders (SCID; 1 study).

*Search 2: Identify evidence of reliability, validity and diagnostic utility of commonly used depression symptom screening scales used in childhood and adolescent preventative interventions for MDD.*

Figure 1 shows the study selection process. The search for the reliability, validity and diagnostic utility of depression symptom screening scales for children and adolescents yielded a total of 2017 results, of which 1798 were unique. Of these, 1626 were excluded as they were not relevant to the topic. Of the remaining 172 publications, 77 were excluded as
the age of the sample was outside the specified age range (5-18), 30 conducted unrelated psychometric tests, 10 assessed the validity of translated versions only and 3 assessed short or alternate versions of the target scale. We identified the remaining 52 papers (comprising 54 studies) as relevant, of which 21 assessed the reliability, validity and diagnostic utility of the CDI, 17 examined the BDI (one study additionally examined the CES-D, and study additionally examined the RADS), 10 examined the CES-D and 6 examined the RADS (Figure 1). The 54 studies comprised a total of 66 individual data points and 34,542 participants. Studies were conducted predominantly in the United States, with a minority conducted in Europe (Spain, Germany, Belgium, Denmark, the Netherlands, Greece, Sweden and Switzerland), Africa (Nigeria and Rwanda), India and Taiwan. Half of the articles included in this review (n = 26 articles, 32 studies; 50%) were published since the last major review of symptom screening scales for MDD among children and adolescents in 2002 (Myers and Winters, 2002b).

Risk of bias in included studies

Of the 52 individual papers (based on 54 studies) included in the review, risk of bias was mostly unable to be determined, or determined to be high (Figure 1a, online supplement). Only 10 studies employed representative samples. Four studies pre-specified the cutoff scores on the scales to determine ‘caseness’, with most studies calculating these values post-hoc (n = 39). Of the 33 studies that used a diagnostic interview to determine the discriminative validity of the symptom screening scale, 24 used structured or semistructured clinical interviews based on ICD-10 or DSM criteria, however only 14 of these stated that the interviewers were blinded to the scale scores when administering the interview. The 9 studies that used undefined clinical interviews that were not based on ICD-10 or DSM diagnostic criteria were excluded from the analyses. Appropriate lag time between the index test (symptom screening
scale) and reference test (diagnostic interview) was deemed to be within one week (Whiting et al., 2011a), and 11 studies fulfilled this criterion.

**Commonly used depression scales for children and adolescents**

*Children’s Depression Inventory (CDI)*

The CDI is the child version of Beck Depression Inventory (BDI), designed for use with children aged 7-18 years (Kovacs, 1984). The scale is self report, comprises 27 items with three-point answers and takes approximately 10-20 minutes to complete and score (Kovacs, 1984). The period of assessment is the prior two weeks, and the range of possible scores is 0 to 54. The CDI was specifically designed for and tested among children and adolescents and includes a corresponding parent version. It has a good track record in depression research as it is one of the most widely used and studied scale of youth depression, with normative data available for children and adolescents (Kovacs, 1984). The CDI purports to possess a five factor structure, however one of the factors relates to externalising disorders (‘acting out’) and one to anxiety, resulting in uncertainty around the validity of depression construct (Myers and Winters, 2002b).

Reliability, validity and diagnostic utility coefficients for the CDI among clinical and nonclinical samples identified in our review are summarised in Table 1, and Figure 2. The 21 identified studies comprised a total of 25 data points (n = 8371). Of these, 14 studies (18 data points) provided details of the scales’ internal reliability (Allgaier et al., 2012, Carey et al., 1987, Crowley and Emerson, 1996, Crowley et al., 1994, Figueras Masip et al., 2010, Fundudis et al., 1991, Giannakopoulos et al., 2009, Nelson III et al., 1987, Reynolds et al., 1985, Saylor et al., 1984, Smucker et al., 1986,
Sorensen et al., 2005, Thompson et al., 2012, Weiss et al., 1991). The pooled estimate for internal reliability was 0.86 (18 data points, \( n = 7372 \), 95% confidence interval (CI): 0.84 to 0.88) and was classified as ‘good’ (Figure 2), however heterogeneity was high (\( I^2 = 76\% \)). Metaregression showed that sample type (clinical vs. nonclinical; \( t(1, 15) = 0.50, p = 0.62, \) adjusted \( R^2 < 0 \)), sample age (child vs adolescent; \( t(1,15) = 0.31, p = 0.76, R^2 < 0 \)) and risk of bias quality score (\( t(1, 15) = 1.61, p = 0.13, R^2 = 67.38\% \)) did not significantly affect the internal reliability values on the CDI.

Nine studies (10 data points) examined the discriminative validity of the CDI (Allgaier et al., 2012, Figueras Masip et al., 2010, Fristad et al., 1988, Fundudis et al., 1991, Roelofs et al., 2010, Shemesh et al., 2005, Timbremont et al., 2004, Sorensen et al., 2005, Craighead et al., 1995) with cutoff scores ranging from 11 to 19 (\( M = 14.5, SD = 2.67 \)). Of these, seven used structured or semi structured clinical interviews to determine presence of MDD. All seven studies examined the specificity and sensitivity of the CDI. Bivariate meta analyses revealed that sensitivity was ‘good’ overall (7 data points, \( n = 1432 \), pooled estimate: 0.83, 95% CI: 0.77 to 0.89), as was specificity (7 data points, \( n = 1432 \), pooled estimate: 0.84, 95% CI: 0.77 to 0.92), however heterogeneity was high (\( I^2 = 91\% \) and 96% respectively). Metaregression showed that the sample type (clinical vs. nonclinical; \( t(1, 6) = 0.59, p = 0.58, R^2 < 0 \)), sample age (child vs adolescent; \( t(1, 6) = 1.05, p = 0.33, R^2 < 0 \)), cutpoint used (\( t(1, 6) = 2.21, p = 0.06, R^2 = 38.68\% \)) and risk of bias quality score (\( t(1, 6) = 0.56, p = 0.59, R^2 < 0 \)) did not significantly affect sensitivity values on the CDI. Six studies examined the PPV of the CDI, all of which were conducted among clinical samples, and most were classified as ‘low’, with values of 0.28 (Allgaier et al., 2012), 0.35 (Roelofs et al., 2010), 0.38 (Shemesh et al., 2005), 0.21 (Sorensen et al., 2005), 0.63 (Timbremont et al., 2004) and 0.90 (Fristad et al., 1988). Five studies examined the NPV of the CDI, and as per the PPV values, all were conducted among clinical samples. NPV values were mostly classified as ‘excellent’, with values of 1.0
(Roelofs et al., 2010), 0.98 (Allgaier et al., 2012), 0.98 (Timbremont et al., 2004), 0.94 (Shemesh et al., 2005) and 0.63 (Sorensen et al., 2005). Four studies examined the diagnostic accuracy of the CDI using AUC analyses, all of which were conducted among clinical samples (Allgaier et al., 2012, Roelofs et al., 2010, Sorensen et al., 2005, Timbremont et al., 2004), and overall accuracy was classified as ‘high’ (4 data points, \( n = 1146 \), pooled estimate: 0.90, 95% CI: 0.79 to 0.98; Figure 2), however heterogeneity was high \( I^2=95\% \). There was insufficient data to conduct a metaregression on AUC values for the CDI.

**Beck Depression Inventory (BDI)**

The BDI was originally developed in 1961 as a depression symptom rating scale for the adult population (Beck and Alford, 2009). The original inventory comprises 21 items, with four (or five) statement responses representing how the respondent has been feeling during the past week and current day. The statements are presented in order of increasing severity and are scored from 0 to 3 (alternative statements sharing the same score are sometimes provided). Six of the items refer to vegetative symptoms; the remaining 15 items refer to either affective or cognitive symptoms. The range of possible scores is 0 to 63 (Beck et al., 1961). The scale is widely used among both adults and adolescents and has a strong track record in depression research (Myers and Winters, 2002b). The scale lacks items pertaining to school and does not offer parallel parent or teacher rating forms, and as such, its use among younger children may not be suitable (Kovacs and Beck, 1977). There are several alternate versions of the BDI. The BDI-1A is a revision of the original BDI (Beck et al., 1996), also containing 21 items but with refined response formats and a defined assessment period of the prior two weeks. The 21 item BDI-II is a further revision of the BDI, developed in 1996 to align with the DSM-IV criteria for depression (Beck et al., 1996).
Reliability, validity and diagnostic utility coefficients for the BDI among clinical and nonclinical samples of children and adolescents are summarised in Table 2 and Figure 3. The 17 identified studies comprised a total of 20 data points (n = 5464). Of these, 11 studies (12 data points) examined the scale’s internal reliability (Adewuya et al., 2007, Ambrosini et al., 1991, Barrera and Garrison-Jones, 1988, Dolle et al., 2012, Jolly et al., 1994, Krefetz et al., 2002, Kumar et al., 2002, Roberts et al., 1991, Strober et al., 1981, Teri, 1982, Kashani et al., 1990). The pooled estimate for internal reliability was 0.86 (12 data points, n = 4152, 95% CI: 0.81 to 0.90) and was classified as ‘good’, and heterogeneity was moderate (I² = 72%) (Figure 3). Metaregression showed that sample type (clinical vs. nonclinical; t (1, 9) = 1.86, p = 0.09, R² = 32.17%), sample age (child vs adolescent; t (1, 9) = 0.88, p = 0.40, R² = 73.52%) and risk of bias quality score (t (1, 9) = 0.70, p = 0.50, R² < 0) did not significantly affect the internal reliability values on the BDI.

Fifteen studies (18 data points) examined the discriminative validity of the BDI (Adewuya et al., 2007, Ambrosini et al., 1991, Barrera and Garrison-Jones, 1988, Bennett et al., 1997, Blom et al., 2010, Canals et al., 2001, Dolle et al., 2012, Kashani et al., 1990, Krefetz et al., 2002, Kumar et al., 2002, Marton et al., 1991, Roberts et al., 1991, Russell et al., 2012, Strober et al., 1981, Whitaker et al., 1990), with cutoff scores ranging from 11 to 24. Of the 13 studies that used a structured or semi-structured clinical interview, 12 studies (13 data points) examined the scales’ sensitivity and specificity (Adewuya et al., 2007, Ambrosini et al., 1991, Barrera and Garrison-Jones, 1988, Bennett et al., 1997, Canals et al., 2001, Dolle et al., 2012, Kashani et al., 1990, Marton et al., 1991, Roberts et al., 1991, Russell et al., 2012, Strober et al., 1981, Whitaker et al., 1990). Bivariate meta analyses revealed that sensitivity was ‘good’ overall (13 data points, n = 4597, pooled estimate: 0.81, 95% CI: 0.74 to 0.87), as was specificity (13 data points, n = 4597, pooled estimate: 0.81, 95% CI: 0.75 to 0.88; Figure 3), however heterogeneity was high (I² = 94% and 97% respectively). Metaregression showed
that the sample type (clinical vs. nonclinical; $t(1, 12) = 0.73, p = 0.48, R^2 < 0$), sample age (child vs adolescent; $t(1, 12) = 0.85, p = 0.41, R^2 < 0$), cutpoint used ($t(1, 12) = 0.13, p = 0.90, R^2 < 0$) and risk of bias quality score ($t(1, 12) = 1.32, p = 0.37, R^2 = 28.33\%$) did not significantly affect sensitivity values on the BDI. Similar results were found for specificity, with no impact of sample type ($t(1, 12) = 0.86, p = 0.40, R^2 < 0$), age ($t(1, 12) = 0.83, p = 0.42, R^2 < 0$), cutpoint used ($t(1, 12) = 0.94, p = 0.37, R^2 < 0$) or risk of bias quality score ($t(1, 12) = 0.39, p = 0.70, R^2 < 0$).

Eight studies examined the PPV of the BDI, half of which were conducted with clinical samples and half in non-clinical samples, with significant heterogeneity between studies. PPV values were higher overall in clinical samples, with values of 0.79 (Marton et al., 1991), 0.81 (Bennett et al., 1997), 0.81 (Dolle et al., 2012), and 0.93 (Ambrosini et al., 1991) and were lower, however more varied in nonclinical samples, with values of 0.10 (Roberts et al., 1991), 0.14 (Russell et al., 2012), 0.47 (Canals et al., 2001) and 0.88 (Adewuya et al., 2007) (Figure 3). Five studies examined the NPV of the BDI, with values of 0.80 (Bennett et al., 1997), 0.95 (Dolle et al., 2012), 0.97 (Russell et al., 2012), 0.98 (Adewuya et al., 2007), 0.99 (Canals et al., 2001) and 0.99 (Roberts et al., 1991) (Figure 3). Six studies (7 data points) used AUC analyses to examine the diagnostic accuracy of the BDI (Adewuya et al., 2007, Blom et al., 2010, Dolle et al., 2012, Kashani et al., 1990, Roberts et al., 1991, Russell et al., 2012), and overall accuracy was classified as ‘high’ (7 data points, $n = 3249$, pooled estimate: 0.92, 95% CI: 0.81 to 1.00). Metaregression showed that the sample type (clinical vs. nonclinical; $t(1, 6) = 0.39, p = 0.71, R^2 < 0$), sample age (child vs adolescent; $t(1, 6) = 0.09, p = 0.93, R^2 < 0$), cutpoint used ($t(1, 6) = 0.17, p = 0.87, R^2 < 0$) and risk of bias quality score ($t(1, 6) = 1.44, p = 0.19, R^2 = 14.16\%$) did not significantly affect AUC values on the BDI.
Center for Epidemiologic Studies Depression Scale (CES-D)

The CES-D is a 20-item self-report questionnaire designed to detect depression in the general population. The items address six symptom areas of depression including depressed mood, feelings of guilt/worthlessness, helplessness, psychomotor retardation, loss of appetite and sleep disturbance (Radloff, 1977). Each of the 20 items is rated on a scale from 0 to 3, with total scores ranging from 0-60. The period of assessment is the past week. The CES-DC is the child version of the original CES-D and retains the same number of items and response format; however language has been adjusted to suit a child reading level (Weissman et al., 1980).

Reliability, validity and diagnostic utility coefficients for the CES-D among clinical and nonclinical samples of children and adolescents are summarised in Table 3 and Figure 4. Nine studies (10 data points) examined the internal reliability of the original full scale CES-D (Aebi et al., 2009, Betancourt et al., 2012, Cuijpers et al., 2008, Fendrich et al., 1990, Garrison et al., 1991, Logsdon and Myers, 2010, Roberts et al., 1991, Thrane et al., 2004). The pooled estimate for internal reliability was 0.88 (10 data points, \(n = 9006\), 95% CI: 0.84 to 0.92, Figure 4) was classified as ‘good’, however heterogeneity was high (\(I^2 = 92\%\)).

Metaregression showed that sample type (clinical vs. nonclinical; \(t(1, 7) = 0.87, p = 0.41, R^2 < 0\)), sample age (child vs adolescent; \(t(1, 7) = 0.79, p = 0.45, R^2 = 2.75\%\)) and risk of bias quality score (\(t(1, 7) = 0.44, p = 0.67, R^2 < 0\)) did not significantly affect the internal reliability values on the CES-D.

2004), with cutoff scores ranging from 12 to 24. Of these, eight studies used structured or semi-structured clinical interviews to determine a diagnosis of depression. Bivariate meta analysis revealed that sensitivity coefficients were ‘moderate’ overall (9 data points, n = 9209, pooled estimate: 0.76, 95% CI: 0.67 to 0.84) as was specificity (9 data points, n = 9209, pooled estimate: 0.71, 95% CI: 0.63 to 0.79). Metaregression showed that sample type (clinical vs. nonclinical; t(1, 7) = 0.08, p = 0.93, R^2 < 0), sample age (child vs adolescent; t(1, 7) = 0.80, p = 0.45, R^2 < 0) risk of bias quality score (t(1, 7) = 1.71, p = 0.13, R^2 = 18.04%) and cutpoint used (t (1, 7) = 0.70, p = 0.51, R^2 < 0) did not significantly affect the sensitivity values on the CES-D. Similar results were found for specificity, with no impact of sample type (t(1, 7) = 0.12, p = 0.91, R^2 < 0), age (t(1, 7) = 1.36, p = 0.21, R^2 = 12.69%), risk of bias quality score (t(1, 7) = 0.77, p = 0.46, R^2 < 0) or cutpoint used (t(1, 7) = 2.12, p = 0.07, R^2 = 32.62%).

Five studies (six data points) examined the PPV and NPV of the CES-D, with only one study conducted among a clinical sample (Logsdon and Myers, 2010), and the remainder in nonclinical samples (Fendrich et al., 1990, Garrison et al., 1991, Logsdon and Myers, 2010, Prescott et al., 1998, Roberts et al., 1991). PPV was low overall, with values of 0.08 (Roberts et al., 1991), 0.15 (Fendrich et al., 1990), 0.16 (Garrison et al., 1991), 0.24 (Prescott et al., 1998), 0.25 (Logsdon and Myers, 2010) and 0.32 (Garrison et al., 1991). NPV was good overall (with the exception of the one study conducted among a clinical sample where NPV = 0.12; (Logsdon and Myers, 2010)), and values were 0.96 (Fendrich et al., 1990, Prescott et al., 1998), 0.98 (Garrison et al., 1991) and 0.99 (Roberts et al., 1991). Seven studies (nine data points) examined diagnostic utility of the CES-D using AUC analyses (Betancourt et al., 2012, Cuijpers et al., 2008, Garrison et al., 1991, Logsdon and Myers, 2010, Prescott et al., 1998, Roberts et al., 1991, Yang et al., 2004), which was classified as ‘moderate’ overall (9 data points, n = 8989, pooled estimate: 0.83, 95% CI: 0.74 to 0.90, I^2 = 99%). Metaregression
revealed that studies using higher cutpoints on the CES-D had higher diagnostic accuracy ($t(1, 8) = 5.1, p = 0.01, \beta = 1.01, R^2 = 53.22\%$), with no significant effect found for sample type (clinical vs. nonclinical; $t(1, 8) = 1.54, p = 0.17, R^2 = 12.88\%$), sample age (child vs. adolescent; $t(1, 8) = 1.12, p = 0.30, R^2 = 7.93\%$), and the risk of bias quality score ($t(1, 8) = 0.69, p = 0.51, R^2 < 0$).

Reynolds Adolescent Depression Scale (RADS)

The Reynolds Adolescent Depression Scale is a 30-item self report instrument developed in 1986 to assess depression symptom severity in adolescents (13-18 years) (Reynolds, 1986). Responses are made on a four point scale indicating frequency (almost never, hardly ever, sometimes, most of the time) of symptoms characteristics of depression. The scale assesses cognitive, somatic, psychomotor, and interpersonal symptoms derived from the DSM-III, with an assessment period of the past two weeks. The RADS takes approximately 10 minutes to complete, and scores range from 30 to 120 (Walker et al., 2005a). The RADS-2 is an updated version of the original RADS. It also contains 30 items, and addresses several psychometric issues in the original RADS such as re-examining estimates of internal consistency, criterion related validity, and known groups validity (Osman et al., 2010).

Scores across the reliability, validity and diagnostic utility for the RADS among clinical and nonclinical samples of children and adolescents are summarised in Table 4. Six studies (7 data points) examined the internal reliability of the RADS (Boyd and Gullone, 1997, Krefetz et al., 2002, Osman et al., 2010, Reynolds and Miller, 1985, Walker et al., 2005b, Weber and Terhorst, 2010), which was classified as ‘excellent’ overall (7 data points, $n = 11095$, pooled estimate: 0.93, 95% CI: 0.86 to 0.99), however heterogeneity was high ($I^2 = 92\%$).

Metaregression showed that sample type (clinical vs. nonclinical; $t(1, 4) = 0.52, p = 0.62, R^2$
< 0), sample age (child vs adolescent; $t(1, 4) = 1.10, p = 0.32, R^2 = 5.56\%$) and risk of bias quality score ($t(1, 4) = 0.66, p = 0.54, R^2 < 0$) did not significantly affect the internal reliability values on the RADS.

Only two studies examined the discriminative validity of the full scale RADS (Krefetz et al., 2002, Osman et al., 2010) using cutoff scores of 67 and 70, however neither study used a structured or semi-structured clinical interview based on DSM or ICD-10 criteria to determine a diagnosis of MDD, and thus meta analyses and metaregression for sensitivity, specificity and AUC values were not conducted.

**Overall judgement**

Comparisons across scales using metaregression revealed no differences in the four key measures of reliability and validity. Internal reliability was classified as ‘good’ overall (47 data points, $n = 31593$, pooled estimate: 0.89, 95% CI: 0.86 to 0.92, $I^2 = 93\%$), with metaregression revealing no impact of scale type (CDI, BDI, CES-D, RADS; $t(1, 44) = 2.54, p = 0.07, R^2 = 15.31\%$), sample (clinical vs. nonclinical; $t(1, 44) = 0.21, p = 0.84, R^2 < 0$), age (child vs. adolescent; $t(1, 44) = 1.30, p = 0.20, R^2 = 3.33\%$) or risk of bias quality score ($t(1, 44) = 0.05, p = 0.96, R^2 < 0$). Sensitivity and specificity coefficients were ‘moderate’ (sensitivity: 29 data points, $n = 15238$, pooled estimate: 0.80, 95% CI: 0.76 to 0.84; specificity: 29 data points, $n = 15238$, pooled estimate: 0.78, 95% CI: 0.74 to 0.83), with metaregression revealing no impact of scale type ($t(1, 29) = 1.36, p = 0.18, R^2 = 3.06\%$), sample type ($t(1, 29 = 1.24, p = 0.22, R^2 = 1.53\%$) age ($t(1, 29) = 1.25, p = 0.22, R^2 = 1.60\%$) cutpoint ($t(1, 29) = 0.55, p = 0.59, R^2 < 0$) or risk of bias quality score ($t(1, 29) = 1.79, p = 0.08, R^2 = 6.54\%$) for sensitivity coefficients. Similar results were found for specificity, with no difference on the basis of scale type ($t(1, 29) = 1.65, p = 0.12, R^2 = 16.12\%$) sample ($t(1,$
Overall diagnostic accuracy was 'moderate' (20 data points, $n = 13384$, pooled estimate: 0.86, 95% CI: 0.79 to 0.92, $I^2 = 98\%$), with metaregression revealing no impact of scale type ($t(1, 20) = 1.52, p = 0.14, R^2 = 6.19\%$), sample ($t(1, 20) = 0.84, p = 0.81, R^2 < 0$), age ($t(1, 20) = 0.41, p = 0.68, R^2 < 0$) cutpoint ($t(1, 20) = 0.25, p = 0.80, R^2 < 0$) or risk of bias quality score ($t(1, 20) = 0.97, p = 0.34, R^2 < 0$).

**Discussion**

This is the first systematic review and meta analysis to provide an objective judgement of the reliability, validity and diagnostic utility of symptom screening scales used in prevention research for MDD among children and adolescents. The review identified 54 studies (published in 52 papers), of which 32 studies were published since the last major narrative review was conducted in 2002 (Myers and Winters, 2002b). The outcomes of this review indicate that commonly used symptom screening scales for MDD including the CDI, BDI, CES-D and the RADS have good internal consistency when used among children and adolescents; and metaregression revealed that these properties did not differ when a range of factors were considered, including sample type (clinical vs. nonclinical), age (child vs. adolescent) and the risk of bias quality score. While the ability of each scale to correctly identify positive and negative cases (i.e. diagnostic utility) was moderate overall; positive predictive power was poor across most scales, suggesting that using cutoff scores on these scales to determine clinical levels of MDD may result in high misclassification rates, particularly when used in nonclinical settings (such as schools). Importantly, cutoff scores to determine ‘caseness’ varied widely, and no single score was identified to be applicable in both clinical and community settings, for any scale. However, metaregression revealed that...
studies using higher cutoff scores on the CES-D had higher diagnostic accuracy (AUC) than those with lower scores. Regardless, if researchers or clinicians choose to employ cutoff scores on commonly used symptom screening scales for MDD to determine clinical ‘caseness’, they will most likely need to be adjusted to suit the sample under investigation.

Limitations

There are several limitations to consider when interpreting the outcomes of this review. Firstly, risk of bias for the included studies was largely unable to be determined, or determined to be high due to insufficient reporting. Secondly, many studies adjusted the cutoff score to determine clinical caseness post hoc in order to maximise sensitivity and specificity values, which may have artificially increased the diagnostic accuracy results using AUC analyses, as these values are based on a trade off of sensitivity and specificity. Consequently, the results regarding diagnostic utility identified in this review may differ in real world settings where specific cutpoints are typically chosen a priori to classify participants as having clinically significant depressive symptoms. Further, such cutpoints are likely to vary substantially on a study by study basis and depending on the context in which the scale is used (e.g. schools versus clinics). Thirdly, positive and negative predictive values are dependent on the prevalence of the disorder in the sample (Trikalinos et al., 2012), and as such, the low positive predictive values identified in this review may reflect a low prevalence of depression among some samples; however the small number of studies and significant heterogeneity between them precluded the use of metaregression to determine factors which may have influenced predictive power in this review. Where metaregression was conducted, it is possible that some analyses were underpowered due to low numbers of studies with adequate data, as evidenced by low adjusted \( R^2 \) values for some outcomes. A greater number
of high quality studies examining the psychometric properties of these scales across a range of settings and ages is needed in order to identify circumstances in which performance of these scales is likely to differ. Further, while our review considered internal reliability of the scales, other measures of reliability that may impact scale utility, including test retest and inter rater reliability were less consistently reported, precluding the conduct of meta analyses and metaregression on these outcomes. Finally, it is also important to consider that while the criterion standards employed throughout this review (i.e. diagnostic interview schedules based on DSM or ICD-10 diagnostic criteria such as the K-SADs) are currently the gold standard in detecting the presence of a depressive disorder (Carlisle and McClellan, 2009), they are not perfect measures, and are also open to the same issues of reliability and validity as the screening scales themselves (Hodges, 1993).

There has been recent criticism of binary classification systems such as the DSM, which take the assumption that people should be classified as either having or not having a particular psychopathological disorder (Hankin et al., 2005), with suggestions that depression may occur on a continuum (Solomon et al., 2001), particularly so in youth (Hankin et al., 2005). In light of this, other outcomes in depression treatment that take into account the disability caused by the disorder have been considered, such as functional impairment (McKnight and Kashdan, 2009). Recent research suggests that indicators of disability or functional impairment, in the absence of a formal diagnosis of a mental disorder may be a useful measure of need for mental health services and treatment. An analysis of the 2007 Australian Survey of Mental Health and Wellbeing found that a significant proportion of those who had used mental health services in the past 12 months did not have a formally diagnosed mental disorder, but had other indicators of possible need, including maintenance treatment following a previous episode or significant psychological distress in the absence of a diagnosis (Harris et al., 2014). Thus, in addition to using classification systems such as ICD
and DSM to determine mental disorder diagnoses and symptom screening scales to measure symptom severity, researchers examining the efficacy of depression prevention and treatment interventions should also consider including items assessing functional impairment, as this may be a useful indicator of disability and need for treatment, even if a formal mental diagnosis is not given (Üstün and Kennedy, 2009).

While there has been some suggestion that routine screening of MDD symptoms may improve the detection of, and initiate timely treatment for MDD among children and adolescents at a population level (Thombs et al., 2012), the impact of routine screening for MDD on health outcomes for children and adolescents is unclear (Williams et al., 2009), and there is some concern that screening may have the potential to cause harm to those who may be diagnosed and treated inappropriately (MacMillan et al., 2005). Reviews conducted by the Canadian Task Force on Preventive Health Care, and the United Kingdom’s National Institute of Health and Clinical Excellence both concluded that there was insufficient evidence regarding health outcomes to support the routine screening of MDD among children and adolescents in primary healthcare settings (MacMillan et al., 2005, National Collaborating Centre for Mental Health, 2005). Unfortunately there have been few large prospective cohort studies that have examined the ability of MDD screening scales in predicting the onset of depression among adolescents in order to inform their routine use in healthcare settings (van Lang et al., 2007). Interestingly such studies have indicated that recurrent screening of depressive symptoms does not improve detection of MDD, but that cognitive and physical symptoms (such as sleeping problems) may be better predictors. (McKenzie et al., 2011, van Lang et al., 2007). For example, McKenzie et al, 2011 identified that particular items of the Short Mood and Feelings Questionnaire (SMFQ) that are not included in formal diagnostic criteria for MDD, such as “I hated myself” were most predictive of high depressive symptoms 12 months later. Such single items may have utility
for being incorporated into short screening measures, or existing population level surveys -
many of which exclude mental disorder symptomatology (Baxter et al., 2013), however
further research using large cohort studies is needed to support this.

Conclusions

Symptom screening scales for MDD are reliable measures of MDD symptomatology among
adolescent samples. While we found that the psychometric properties of the scales did not
differ when used in clinical and nonclinical settings, across a range of subject ages and with
variable study quality, limited data was available, and further high quality studies are needed
to identify conditions in which test performance may vary. Cutoff scores on symptom
screening scales for MDD may be useful for identifying 'risk status', however, the cutoffs
used will likely vary depending on the context in which the scale is applied, and
misclassification is likely to be high, particularly in nonclinical samples where disorder
prevalence is low (e.g. school settings). Further research examining the predictive ability of
depression screening scales in cohort studies of children and adolescents are needed.
References


Harris, M. G., Diminic, S., Burgess, P. M., Carstensen, G., Stewart, G., Pirkis, J. & Whiteford, H. A. 2014. Understanding service demand for mental health among Australians aged 16 to 64 years according to their possible need for treatment. Aust NZ J Psychiat.


StataCorp, 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp, LP.


for the quality assessment of diagnostic accuracy studies. Annals of internal medicine, 155, 529-36.
Acknowledgements

The authors would like to acknowledge Professor Gavin Andrews (UNSW Australia) and Dr Mary Lou Chatterton (Deakin University) for their assistance in developing ideas and refining the concept of the manuscript.
Conflict of interest

All authors declare that they have no conflicts of interest.
Contributors

EAS, LD and CM conceived the study. EAS, YL and AL conducted the reviews and extracted the data, and EAS and AL synthesised the data, with MH and GP contributing to interpretation. All authors contributed to, and approved the final manuscript.
Role of funding source

The work was funded by the National Health and Medical Research Council of Australia (NH&MRC Grant No. APP1041131).
Table 1. Validation evidence for the Children’s Depression Inventory (CDI) in child and adolescent samples

<table>
<thead>
<tr>
<th>Source</th>
<th>N</th>
<th>Age &amp; Gender (%m, %f)</th>
<th>Sample (location)</th>
<th>Scale Name (no. of items)</th>
<th>Reliability</th>
<th>Criterion Cut off</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AU C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allgaier et al. 2012</td>
<td>40</td>
<td>9-12 (56, 44)</td>
<td>Child hospital patients (Germany)</td>
<td>CDI (27)</td>
<td>α = 0.84</td>
<td>Kinder-DIPS</td>
<td>≥ 12</td>
<td>0.83</td>
<td>0.83</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Figueras Masip et al. 2010a</td>
<td>10</td>
<td>10-18 (41, 59)</td>
<td>Child and adolescent clinical sample (Spain)</td>
<td>CDI (27)</td>
<td>α = 0.85</td>
<td>Clinical interview</td>
<td>19</td>
<td>0.95</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Roelofs et al. 2010</td>
<td>51</td>
<td>7-18 (52, 48)</td>
<td>Child and adolescent clinical sample (Belgium)</td>
<td>CDI (27)</td>
<td>-</td>
<td>KID-SCID</td>
<td>16</td>
<td>0.92</td>
<td>0.95</td>
<td>0.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Shemesh et al. 2005</td>
<td>81</td>
<td>8-19 (42, 58)</td>
<td>Child and adolescent hospital patients (USA)</td>
<td>CDI (27)</td>
<td>-</td>
<td>K-SADS-PL</td>
<td>11</td>
<td>0.80</td>
<td>0.70</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>Sorensen et al. 2005</td>
<td>14</td>
<td>8-13 (54, 46)</td>
<td>Child and adolescent psychiatric patients (Denmark)</td>
<td>CDI (27)</td>
<td>α = 0.86</td>
<td>K-SADS-PL</td>
<td>12</td>
<td>0.63</td>
<td>0.64</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>Timbremont et al. 2004</td>
<td>80</td>
<td>8-18 (42, 58)</td>
<td>Child and adolescent psychiatric inpatients/outpatients, (Belgium, Netherlands)</td>
<td>CDI (27)</td>
<td>κ = 0.76</td>
<td>KID-SCID</td>
<td>16</td>
<td>0.84</td>
<td>0.94</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>Craighead et al. 1995</td>
<td>10</td>
<td>12-18 (54, 46)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>CDI (27)</td>
<td>-</td>
<td>K-SADS-III-R</td>
<td>17</td>
<td>0.81</td>
<td>0.84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weiss et al. 1991a</td>
<td>51</td>
<td>8-12 (73, 27)</td>
<td>Child in- and out-patients</td>
<td>CDI (27)</td>
<td>α = 0.86</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weiss et al. 1991b</td>
<td>76</td>
<td>13-16 (52, 48)</td>
<td>Adolescent in-and out-patients</td>
<td>CDI (27)</td>
<td>α = 0.88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nelson III and Politano 1990</td>
<td>96</td>
<td>6-15 (61, 39)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>CDI (27)</td>
<td>Test-retest: 0.62</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fristad et al. 1988</td>
<td>98</td>
<td>6-13 (54, 46)</td>
<td>Child and adolescent depressed inpatients, psychiatric inpatient controls,</td>
<td>CDI (27)</td>
<td>-</td>
<td>DICA ge15</td>
<td>0.89</td>
<td>0.65</td>
<td>0.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Study</td>
<td>Year</td>
<td>Age Range</td>
<td>Sample Description</td>
<td>CDI</td>
<td>α</td>
<td>Test-Retest</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------------------------------------------------------</td>
<td>-------</td>
<td>-----------</td>
<td>-------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carey et al. 1987a</td>
<td>1987</td>
<td>9-17 (75, 25)</td>
<td>Child and adolescent inpatients (USA)</td>
<td>CDI (27)</td>
<td>α = 0.83</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nelson III et al. 1987</td>
<td>1987</td>
<td>6-17 (68, 32)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>CDI (27)</td>
<td>α = 0.86</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saylor et al. 1984a</td>
<td>1984</td>
<td>7-16 (69, 31)</td>
<td>Child and adolescent inpatients (USA)</td>
<td>CDI (27)</td>
<td>Kuder-Richardson coefficient= 0.80</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thompson et al. 2012</td>
<td>2012</td>
<td>11-17 (53, 47)</td>
<td>Child and adolescent sample with inflammatory bowel disease (USA)</td>
<td>CDI (27)</td>
<td>α = 0.88</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Figueras et al. 2010b</td>
<td>2010</td>
<td>10-18 (46, 54)</td>
<td>Child and adolescent community sample (Spain)</td>
<td>CDI (27)</td>
<td>α = 0.82</td>
<td>Test-retest = 0.81</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giannakopoulos et al. 2009</td>
<td>2009</td>
<td>8-12 (47, 53)</td>
<td>Child nationwide school-based sample (Greece)</td>
<td>CDI (27)</td>
<td>α = 0.80; test-retest reliability ICC; 0.82 for girls and 0.62 for boys</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowley and Emerson 1996</td>
<td>1996</td>
<td>8-12 (51, 49)</td>
<td>Child school students (USA)</td>
<td>CDI (27)</td>
<td>α = 0.89</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowley et al. 1994</td>
<td>1994</td>
<td>11-16 (49, 51)</td>
<td>Child and adolescent community sample (USA)</td>
<td>CDI (27)</td>
<td>α = 0.86</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fundudis et al. 1991</td>
<td>1991</td>
<td>8-16 (50, 50)</td>
<td>Children and adolescents of staff of a university child psychiatry department (USA)</td>
<td>CDI (27)</td>
<td>α = 0.88</td>
<td>Standardised Psychiatric Interview</td>
<td>15 0.77 0.77 - - -</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carey et al. 1987b</td>
<td>1987</td>
<td>9-17 (74, 26)</td>
<td>Child and adolescent non-referred sample</td>
<td>CDI (27)</td>
<td>α = 0.75</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finch et al. 1987</td>
<td>1987</td>
<td>7-12 (51, 49)</td>
<td>Child sample from public schools (USA)</td>
<td>CDI (27)</td>
<td>Test-retest: 0.82</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Age Range</td>
<td>Sample Description</td>
<td>Measure</td>
<td>CDI</td>
<td>α</td>
<td>PPV</td>
<td>NPV</td>
<td>AUC</td>
<td>Note</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>--------------------</td>
<td>---------</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>Smucker et al. 1986</td>
<td>12</td>
<td>8-16</td>
<td>Child and adolescent sample from public schools (USA)</td>
<td>CDI (27)</td>
<td>α = 0.89</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Reynolds et al. 1985</td>
<td>16</td>
<td>8-12</td>
<td>Child sample from public schools (USA)</td>
<td>CDI (27)</td>
<td>α = 0.90</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Saylor et al. 1984b</td>
<td>72</td>
<td>10-13</td>
<td>Child sample from public schools (USA)</td>
<td>CDI (27)</td>
<td>Kuder-Richardson coefficient = 0.94</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Note:
N = Number of participants in the study sample
PPV = Positive predictive value
NPV = Negative predictive value
AUC = Area under the curve analysis
CDI = Children's Depression Inventory
α = Cronbach’s alpha reliability co-efficient
κ = Cohen's kappa reliability co-efficient
Kinder-DIPS: The Diagnostic Interview for Psychiatric Disorders in Children and Adolescents
KID-SCID: The child version of the Structured Clinical Interview for DSM Disorders
K-SADS-PL: The Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version
K-SADS-E: The Schedule for Affective Disorders and Schizophrenia for School-Age Children – Epidemiological Version
K-SADS-III-R: The Schedule for Affective Disorders and Schizophrenia for School-Age Children – the revision of the third edition
DICA: The Diagnostic Interview for Children and Adolescents
**Table 2.** Validation evidence for the Beck Depression Inventory (BDI) in child and adolescent samples

<table>
<thead>
<tr>
<th>Source</th>
<th>N</th>
<th>Age &amp; Gender (%m, %f)</th>
<th>Sample (location)</th>
<th>Scale name (no. of items)</th>
<th>Reliability</th>
<th>Criterion</th>
<th>Cutoff</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PP V</th>
<th>NP V</th>
<th>AU C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dolle et al. 2012</td>
<td>88</td>
<td>13-16</td>
<td>Adolescent psychiatric patients (Germany)</td>
<td>BDI-II (21)</td>
<td>α = 0.94</td>
<td>Kinder-DIPS</td>
<td>≥23</td>
<td>0.88</td>
<td>0.92</td>
<td>0.8</td>
<td>0.95</td>
<td>0.9</td>
</tr>
<tr>
<td>Blom et al. 2010</td>
<td>73</td>
<td>14-18 (0, 100)</td>
<td>Adolescent girls who were psychiatric patients (Sweden)</td>
<td>BDI-A1 (21)</td>
<td>-</td>
<td>DAWBA</td>
<td>&gt;14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.8</td>
</tr>
<tr>
<td>Krefetz et al. 2002</td>
<td>10</td>
<td>12-17 (44, 56)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>BDI-II (21)</td>
<td>α = 0.92</td>
<td>PRIME-MD</td>
<td>24</td>
<td>0.74</td>
<td>0.70</td>
<td>0.7</td>
<td>0.67</td>
<td>0.7</td>
</tr>
<tr>
<td>Kumar et al. 2002</td>
<td>10</td>
<td>12-17 (45, 55)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>BDI-II (21)</td>
<td>α = 0.94</td>
<td>PRIME-MD</td>
<td>21</td>
<td>0.85</td>
<td>0.83</td>
<td>0.8</td>
<td>0.83</td>
<td>0.9</td>
</tr>
<tr>
<td>Bennett et al. 1997</td>
<td>32</td>
<td>11-19 (41.5, 58.5)</td>
<td>Child and adolescent psychiatric inpatients and outpatients of a clinic for depression (USA)</td>
<td>BDI (21)</td>
<td>-</td>
<td>K-SADS</td>
<td>13</td>
<td>0.87</td>
<td>0.71</td>
<td>0.8</td>
<td>0.80</td>
<td>-</td>
</tr>
<tr>
<td>Jolly et al. 1994</td>
<td>75</td>
<td>12-17 (47, 53)</td>
<td>Child and adolescent inpatient</td>
<td>BDI (21)</td>
<td>α = 0.88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Study</td>
<td>Sample Description</td>
<td>Measure</td>
<td>α or κ</td>
<td>K-SADS-III-R</td>
<td>CAS</td>
<td>DICA</td>
<td>SADS-16</td>
<td>BDI (21)</td>
<td>18</td>
<td>0.63</td>
<td>0.70</td>
<td>0.1</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------</td>
<td>--------------</td>
<td>-----</td>
<td>------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Ambrosini et al. 1991</td>
<td>Child and adolescent outpatients referred to a clinic for depression (USA)</td>
<td>BDI</td>
<td>0.91</td>
<td>K-SADS-III-R</td>
<td>16</td>
<td>0.86</td>
<td>0.82</td>
<td>0.9</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marton et al., 1991</td>
<td>Adolescent inpatients and outpatients of a children's mental health centre with a current DSM-III Axis I psychiatric disorder (Depressed, n = 60 vs Non-Depressed, n = 59; USA)</td>
<td>BDI</td>
<td>-</td>
<td>DIC</td>
<td>16</td>
<td>0.68</td>
<td>0.81</td>
<td>0.7</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barrera and Garrison-Jones 1988a</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>BDI</td>
<td>0.86</td>
<td>CAS</td>
<td>11</td>
<td>0.82</td>
<td>0.53</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strober et al. 1981</td>
<td>Child and adolescent patients (USA)</td>
<td>BDI</td>
<td>0.79</td>
<td>SADS</td>
<td>16</td>
<td>0.81</td>
<td>0.81</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Clinical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell et al. 2012</td>
<td>Adolescents from three schools (India)</td>
<td>BDI</td>
<td>0.06</td>
<td>SADS-PL</td>
<td>18</td>
<td>0.63</td>
<td>0.70</td>
<td>0.1</td>
<td>4</td>
<td>0.97</td>
<td>0.07</td>
<td>0.6</td>
</tr>
<tr>
<td>Adewuya et al. 2007</td>
<td>Adolescents attending secondary school (Nigeria)</td>
<td>BDI</td>
<td>0.82</td>
<td>K-SADS-E</td>
<td>18</td>
<td>0.91</td>
<td>0.97</td>
<td>0.8</td>
<td>0.8</td>
<td>0.98</td>
<td>0.09</td>
<td>0.9</td>
</tr>
<tr>
<td>Study</td>
<td>Sample Size</td>
<td>Age Range</td>
<td>Gender</td>
<td>Country</td>
<td>Diagnosis</td>
<td>Measure 1</td>
<td>Measure 2</td>
<td>Measure 3</td>
<td>Measure 4</td>
<td>Measure 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------</td>
<td>---------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canals et al. 2001</td>
<td>30</td>
<td>17.5-18.5</td>
<td>Male</td>
<td>Spain</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>SC-16</td>
<td>K-SADS</td>
<td>0.90</td>
<td>0.96</td>
<td>0.4</td>
<td>0.99</td>
</tr>
<tr>
<td>Roberts et al. 1991a</td>
<td>17</td>
<td>14-18</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>K-SADS</td>
<td>11</td>
<td>0.84</td>
<td>0.81</td>
<td>0.1</td>
<td>0.99</td>
</tr>
<tr>
<td>Roberts et al. 1991b</td>
<td>80</td>
<td>14-18</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>K-SADS</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>Roberts et al. 1991c</td>
<td>90</td>
<td>14-18</td>
<td>Female</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>K-SADS</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>Kashan i et al. 1990</td>
<td>10</td>
<td>13-18</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>DICA</td>
<td>16</td>
<td>0.48</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whitaker et al. 1990</td>
<td>35</td>
<td>14-17</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>Semi-structured clinical interview based on DSM-III criteria administered by child psychiatrist or health professional</td>
<td>16</td>
<td>0.80</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barrera and Garrison-Jones 1988b</td>
<td>49</td>
<td>12-18</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td>CAS</td>
<td>16</td>
<td>1.00</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teri 1982</td>
<td>56</td>
<td>14-17</td>
<td>Male</td>
<td>USA</td>
<td>Adolescents</td>
<td>BDI (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Note:
N = Number of participants in the study sample
PPV = Positive predictive value
NPV = Negative predictive value
AUC = Area under the curve analysis
BDI = Beck Depression Inventory
α = Cronbach’s alpha reliability coefficient
κ = Cohen’s kappa reliability coefficient
Kinder-DIPS: The Diagnostic Interview for Psychiatric Disorders in Children and Adolescents
DAWBA: The Development and Wellbeing Assessment
PRIME-MD: The Primary Care Evaluation of Mental Disorders
K-SADS: The Schedule for Affective Disorders and Schizophrenia for School-Age Children
K-SADS-III-R: The Schedule for Affective Disorders and Schizophrenia for School-Age Children – the revision of the third edition
CAS: The Child Assessment Schedule
K-SADS: The Schedule for Affective Disorders and Schizophrenia
K-SADS-PL: The Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version
K-SADS-E: The Schedule for Affective Disorders and Schizophrenia for School-Age Children – Epidemiological Version
DICA: The Diagnostic Interview for Children and Adolescents
SCAN: Schedules for Clinical Assessment in Neuropsychiatry
### Table 3. Validation evidence for the Center for Epidemiologic Studies Depression Scale (CES-D) in child and adolescent samples

<table>
<thead>
<tr>
<th>Source</th>
<th>N Age &amp; Gender (%m, %f)</th>
<th>Sample (location)</th>
<th>Scale name (no. of items)</th>
<th>Reliability</th>
<th>Criteron</th>
<th>Cutoff Sensitivity Specificity PP V NP V AU C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logsdon and Myers 2010</td>
<td>5 9 13-18 (0, 100)</td>
<td>Adolescents mothers at 4-6 weeks postpartum (USA)</td>
<td>CES-D (20)</td>
<td>α = 0.84</td>
<td>K-SADS-PL 16</td>
<td>0.7</td>
</tr>
<tr>
<td>Aebi et al. 2009</td>
<td>1 4 15.5 (33, 67)</td>
<td>Adolescents diagnosed with major depressive disorders (Switzerland)</td>
<td>CES-D (20)</td>
<td>α = 0.83</td>
<td>Clinical interview 21</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Non-Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betancourt et al. 2012</td>
<td>3 6 10-17 (33, 67)</td>
<td>Children and adolescents (Rwanda)</td>
<td>CES-DC (20)</td>
<td>α = 0.86</td>
<td>MINI-KID ≥30</td>
<td>0.82</td>
</tr>
<tr>
<td>Cuijpers et al. 2008</td>
<td>1 3 14-16 (52, 48)</td>
<td>Adolescents (Netherlands)</td>
<td>CES-D (20)</td>
<td>α = 0.93</td>
<td>MINI 22</td>
<td>0.9</td>
</tr>
<tr>
<td>Thrane et al. 2004</td>
<td>1 2 9-16 (54, 46)</td>
<td>Adolescents from three American Indian reservations (USA)</td>
<td>CES-D (20)</td>
<td>α = 0.80</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yang et al. 2004</td>
<td>2 4 12-16 (52, 48)</td>
<td>Adolescents (Taiwan)</td>
<td>CES-D (20)</td>
<td>α = 0.9</td>
<td>K-SADS-E 90% tile 41</td>
<td>0.41</td>
</tr>
<tr>
<td>Prescott et al. 1998</td>
<td>5 6 16.8</td>
<td>Adolescents students from grades 9-12 (USA)</td>
<td>CES-D (20)</td>
<td>-</td>
<td>DISC 16</td>
<td>0.79</td>
</tr>
<tr>
<td>Garrison et al. 1991a</td>
<td>1 2 12-14 (100, 0)</td>
<td>Child and adolescent boys from school sample (USA)</td>
<td>CES-D (20)</td>
<td>α = 0.81</td>
<td>K-SADS-P 12</td>
<td>0.85</td>
</tr>
<tr>
<td>Garrison et al. 1991b</td>
<td>1 2 12-14 (0, 100)</td>
<td>Child and adolescent girls from a school</td>
<td>CES-D (20)</td>
<td>α = 0.86</td>
<td>K-SADS-P 22</td>
<td>0.83</td>
</tr>
<tr>
<td>Study</td>
<td>Subjects</td>
<td>Age Range</td>
<td>Sample Size</td>
<td>CES / K-SADS</td>
<td>α</td>
<td>AUC</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>Roberts et al. 1991a</td>
<td>Adolescents of nine senior high schools (USA)</td>
<td>14-18</td>
<td>7 (47, 53)</td>
<td>CES-D</td>
<td>α = 0.89</td>
<td>K-SADS</td>
</tr>
<tr>
<td></td>
<td>Adolescents male sample of nine senior high schools (USA)</td>
<td>14-18</td>
<td>8 (100, 0)</td>
<td>CES-D</td>
<td>-</td>
<td>K-SADS</td>
</tr>
<tr>
<td></td>
<td>Adolescents female enrolment of nine senior high schools (USA)</td>
<td>14-18</td>
<td>9 (0, 100)</td>
<td>CES-D</td>
<td>-</td>
<td>K-SADS</td>
</tr>
<tr>
<td>Fendrich et al. 1990</td>
<td>Children and adolescents at risk for depression according to their parents’ diagnosis (USA)</td>
<td>12-18</td>
<td>2 (0)</td>
<td>CES-DC</td>
<td>α = 0.89</td>
<td>K-SADS-E</td>
</tr>
</tbody>
</table>

Note:
- N = Number of participants in the study sample
- PPV = Positive predictive value
- NPV = Negative predictive value
- AUC = Area under the curve analysis
- CES-D = Center for Epidemiologic Studies Depression Scale
- CES-DC = Center for Epidemiologic Studies Depression Scale Child Version
- α = Cronbach’s alpha reliability coefficient
- K-SADS-PL: The Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version
- MINI-KID: The Mini-International Neuropsychiatric Interview
- MINI: The Mini-International Neuropsychiatric Interview for children
- K-SADS: The Schedule for Affective Disorders and Schizophrenia for School-Age Children
- K-SADS-E: The Schedule for Affective Disorders and Schizophrenia for School-Age Children – Epidemiological version
- DISC: The National Institute of Mental Health Diagnostic Interview Schedule for Children
- K-SADS-P: The Schedule for Affective Disorders and Schizophrenia for School-Age Children Present Episode version
Table 4. Validation evidence for the Reynolds Adolescent Depression Scale (RADS) in child and adolescent samples

<table>
<thead>
<tr>
<th>Source</th>
<th>N</th>
<th>Age &amp; Gender (%)</th>
<th>Sample (location)</th>
<th>Scale name (no. of items)</th>
<th>Reliability</th>
<th>Criteron</th>
<th>Cutoff</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AU C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osman, A. et al. 2010</td>
<td>196</td>
<td>14-17 (43, 57)</td>
<td>Adolescent inpatients (USA)</td>
<td>RAD S-2 (30)</td>
<td>α = 0.95</td>
<td>BHS</td>
<td>67</td>
<td>0.64</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.80</td>
</tr>
<tr>
<td>Krefetz et al. 2002</td>
<td>100</td>
<td>12-17 (44, 56)</td>
<td>Child and adolescent psychiatric inpatients (USA)</td>
<td>RAD S (30)</td>
<td>α = 0.91</td>
<td>PRIME-MD</td>
<td>70</td>
<td>0.86</td>
<td>0.49</td>
<td>0.6</td>
<td>0.7</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>Non-Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weber and Terhors t 2010</td>
<td>265</td>
<td>13-18 (73, 27)</td>
<td>Adolescent LGBTIQ youth (USA)</td>
<td>RAD S-2 (30)</td>
<td>α = 0.92</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Walker et al. 2005</td>
<td>989</td>
<td>12-18 (46, 54)</td>
<td>Child and adolescent school students (New Zealand)</td>
<td>RAD S (30)</td>
<td>α = 0.94</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boyd et al. 1997</td>
<td>785</td>
<td>11-18 (49, 52)</td>
<td>Child and adolescent school students (Australia)</td>
<td>RAD S (30)</td>
<td>α = 0.85</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Reynolds &amp; Miller 1985a</td>
<td>26</td>
<td>Mean: 17.3</td>
<td>Intellectually delayed adolescent school students (USA)</td>
<td>RAD S (30)</td>
<td>α = 0.87</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Reynolds &amp; Miller 1985b</td>
<td>26</td>
<td>Mean: 16.7</td>
<td>Non-intellectually delayed adolescent school students (USA)</td>
<td>RAD S (30)</td>
<td>α = 0.97</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Note:
N = Number of participants in the study sample
PPV = Positive predictive value
NPV = Negative predictive value
AUC = Area under the curve analysis
RADS = Reynolds Adolescent Depression Scale
α = Cronbach’s alpha reliability coefficient
BHS: Beck Hopelessness Scale
PRIME-MD: The Primary Care Evaluation of Mental Disorders
LGBTIQ: Lesbian, Gay, Bisexual, Transgender, Intersexual, Questioning.
Highlights:

- Symptom screening scales are often used to determine a clinical diagnosis of depression.
- We examined the diagnostic utility of these scales using systematic review and meta-analysis.
- Commonly used screening scales for depression are reliable measures among adolescents.
- Using cutpoints to determine a clinical diagnosis may produce high misclassification rates.