#JAACAP-D-19-00794R2

Revised Version, 30 September 2020

Systematic review: The measurement properties of the Children's Depression Rating Scale-Revised in adolescents with major depressive disorder

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors for the conduct of the research and/or preparation of the article.

This article has been published in Stallwood et al. (2021). Systematic review: The measurement properties of the Children's Depression Rating Scale-Revised in adolescents with major depressive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry,* 60(1):119-133. doi: 10.1016/j.jaac.2020.10.009. Following peer review and can also be viewed on the journal's website at https://doi.org/10.1016/j.jaac.2020.10.009

LAY SUMMARY

The Children's Depression Rating Scale-Revised (CDRS-R) is the most commonly used method to measure depression in treatment studies of teens with depression, but it is unknown whether the CDRS-R is appropriate for the purpose of measuring depression in adolescents. This study aimed to identify all existing evidence of the key measurement properties of the CDRS-R (for example, how well the scale measures what it is supposed to measure) in teens with depression, and to evaluate these properties using a well-established method developed by the COSMIN Initiative (https://www.cosmin.nl/). The study concludes that it is unclear whether the CDRS-R can appropriately measure depression symptom severity in treatment studies of teens with depression based on current available evidence. It is important that the best methods are used to measure outcomes to ensure that results from clinical research studies in teens with depression are meaningful and useful to relevant stakeholders, including patients, caregivers, health care providers, researchers, and policymakers.

ABSTRACT

Objective: To systematically appraise existing evidence of the measurement properties of the Children's Depression Rating Scale-Revised (CDRS-R) in adolescents with MDD. The CDRS-R is the most commonly used scale in adolescent depression research yet was originally designed for use in children six to twelve years old.

Method: Seven databases were searched for studies that evaluated the measurement properties of the CDRS-R in adolescents (ages 12 to 18 years). Of 65 studies screened by full-text, six were included. Measurement properties were appraised using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines. The COSMIN minimum requirements for recommending the use of an outcome measurement instrument are: (1) evidence for sufficient content validity (any level of evidence) and (2) at least low quality evidence for sufficient internal consistency.

Results: Four studies assessed an English language version of the CDRS-R; the other two assessed German and Korean versions, respectively. No study assessed content validity, cross-cultural validity/measurement invariance, or measurement error of the CDRS-R in adolescents with MDD. Low quality evidence was found for sufficient construct validity (n=4 studies) and responsiveness (n=2 studies) assessed via comparator instruments. Very low quality evidence was found for sufficient inter-rater reliability (n=2 studies). The results for structural validity (n=3 studies) and internal consistency (n=5 studies) were inconclusive.

Conclusion: It remains unclear whether the CDRS-R appropriately measures depressive symptom severity in adolescent MDD. Before use of the CDRS-R in adolescent MDD research

can be recommended, evidence of sufficient psychometric properties in adolescents with MDD is needed.

INTRODUCTION

Major depressive disorder (MDD) is a common psychiatric condition worldwide with an estimated 11% lifetime prevalence rate in adolescents.[1,2] The onset of MDD in adolescence is associated with impaired functioning and an increased risk of suicide that can extend into adulthood.[3,4] Despite various treatments being available for adolescent MDD including different modes of psychotherapy and/or medications, many adolescents with MDD remain unresponsive to treatments.[5]

To advance clinical decision-making, studies of adolescent MDD treatments require well-measured and well-reported health outcomes that enable the interpretation, comparison, and synthesis of study results. Two recent studies reporting outcomes of depression treatments for adolescents with MDD demonstrated that vast heterogeneity exists in the outcome measurement instruments (OMIs, e.g., rating scales, questionnaires) used across studies.[6,7] For example, the majority of studies measured "depression symptoms" as the primary outcome, yet this outcome was measured using 27 different OMIs.[6,7] Such heterogeneity is problematic because it introduces variability in the measurement and analysis of outcomes that may undermine the interpretation and synthesis of results and their translation into advancements in care.[8,9] Importantly, for researchers and evidence end-users (e.g., patients, clinicians) to be able to evaluate which OMIs are the most useful, the measurement properties of these OMIs need to be systematically assessed to determine whether they are actually appropriate for measuring "depression symptoms" (e.g., depressive symptom severity) in adolescents with MDD.

The Children's Depression Rating Scale-Revised (CDRS-R)[10] is the most commonly used OMI in adolescent MDD treatment studies.[6,7] First developed over 40 years ago,[11,12] the CDRS-R is a 17-item rating scale that was modelled after the adult-specific Hamilton Rating Scale for Depression.[13] The CDRS-R scale is completed based on semi-structured interviews that a clinician or trained professional conducts with a child and/or adult informant(s) who knows the child well (e.g., caregiver, teacher) in order to assess the presence and severity of 17 symptoms of depression.[10] Three items are rated by the clinician on the basis of the child's non-verbal behavior. The CDRS-R was originally developed for children aged six to twelve years.[11,12] While the CDRS-R was developed to be a valid and reliable OMI for "children with depressive disorders,"[10,12,14] it has been widely used in adolescents in both observational studies and intervention trials.[6,7] It remains unknown whether the CDRS-R has sufficient measurement properties, however, to be appropriate for use in adolescents with MDD.

The purpose of this review was to systematically identify and appraise the existing empirical literature on the sufficiency of the measurement properties of the CDRS-R for use in adolescents with MDD.

METHODS

This study is registered as part of a multi-stage project to develop a core outcome set for trials assessing treatment interventions for adolescents with MDD,[15] as detailed elsewhere.[16] See Table S1, available online, for the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)[17] Checklist completed for this review.

Eligibility Criteria

We followed the Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines in setting the eligibility criteria for this review.[18-20] In brief,

empirical research studies from any country that evaluated the CDRS-R in adolescents, when available in full-text and published in English, were eligible. Studies were eligible if the study population was adolescents aged 12 to 18 years, defined by the National Institute of Child Health and Human Development,[21] or the mean/median age of participants was between 12 and 18 years. Studies with a subgroup analysis of participants inclusive of ages 12 to 18 years were also included. Studies were eligible if at least half of the participants were diagnosed with MDD using a diagnostic tool tied to International Classification of Diseases (ICD) or Diagnostic and Statistical Manual (DSM) criteria (e.g., Kiddie Schedule for Affective Disorders and Schizophrenia [K-SADS], the National Institute of Mental Health Diagnostic Interview Schedule for Children) or if participants had a primary clinical diagnosis of MDD. Studies that were eligible for inclusion had the following primary objective[18,19,22]: to assess at least one or more of the CDRS-R's measurement properties or the development of the CDRS-R. Case reports, protocols, editorials, interviews, conference abstracts, and "grey" literature (i.e., unpublished commercial information or reports that are inaccessible via bibliographic databases) were not included.

Information Sources

The following databases were searched from inception to July 1, 2019: Medical Literature Analysis and Retrieval System Online (MEDLINE), Embase, Psychological Information Database (PsycINFO), Cochrane Central Register of Controlled Trials, Cumulative Index of Nursing and Allied Health Literature (CINAHL), Web of Science, and Scopus. The CDRS-R manual was also reviewed for any relevant studies and citations.[10] Google was searched to identify any missing relevant studies. Reference lists of full-text eligible studies were also checked for any relevant studies to be included.

Search Strategy

A sensitive search strategy was developed in consultation with a medical librarian and included synonyms for the "Children's Depression Rating Scale-Revised" (e.g., CDRS-R; see Table S2, available online, for search strategy and yield details) for electronic database and Google searching. No search filters were applied.

Study Selection

Title and abstract screening for eligible studies was completed by one reviewer (ES) and a randomly selected 15% sample of included studies were independently screened by a second reviewer with training in clinical epidemiology (AM). The kappa score between the two reviewers was 0.76; discrepancies were resolved by discussion between reviewers. Full-text screening for eligible studies was completed by one reviewer (ES) and a second reviewer trained in clinical research methodology (CR) independently screened a 10% sample yielding a kappa score of 0.70; discrepancies were resolved by discussion between reviewers. Full-text screening was performed with screening forms (available online)[23] developed using Research Electronic Data Capture (REDCap) data management software.[24]

Data Collection

The following characteristics were extracted from each eligible study and recorded in REDCap[24]: study first author, number of study participants, participants' mean and median age, age range, age subgroups (if applicable), gender, percentage of sample with MDD, country where the study was conducted, training and job title of individual who administered the CDRS-R, measurement properties assessed, comparator instruments used, and diagnostic criteria used for MDD. See Table 1 for details.

Measurement Properties

To evaluate the measurement properties of the CDRS-R, we followed the COSMIN guidelines for appraising outcome measurement instruments.[18,19,25] The following eight measurement properties, listed in COSMIN's order of importance, were eligible for appraisal using COSMIN guidelines: (i) content validity, (ii) structural validity and internal consistency (i.e., internal structure), and (iii) the remaining measurement properties: cross-cultural validity/measurement invariance, reliability, measurement error, hypotheses testing for construct validity, and responsiveness.[18,26] See Table 2 for definitions. Criterion validity (i.e., the degree to which the scores of an instrument are adequate reflection of a "gold standard") was not assessed in this review, as this is only relevant to OMIs tested against an established gold-standard."[18] As no acceptable "gold-standard" for quantifying depression symptom severity in adolescents was identified by the review team, criterion validity was not considered evaluable for the CDRS-R, and as recommended,[18] the single study that reported evaluating "criterion validity" was included as an evaluation of construct validity.[27]

Three reviewers (ES, SM, MO) independently recorded which of these nine measurement properties each study assessed according to COSMIN's measurement properties definitions. Across studies, there were 16 instances where a measurement property was assessed. The reviewers agreed on 15 out of these 16 measurement properties and the one discrepancy was resolved by discussion between reviewers. NJB verified the measurement properties identified. We included three reviewers and a verifier in this phase of the project because many of the included studies did not clearly identify or define which measurement properties they were assessing. Having three reviewers and a verifier with collective expertise in youth psychiatry, clinical epidemiology, and clinimetrics ensured a consensus-based decision on the tested measurement properties in each study.

Methodological Quality of Included Studies: Assessing Risk of Bias

Two trained reviewers (ES and SM) independently evaluated the methodological quality of each included study using the COSMIN Risk of Bias Checklist.[19] The Risk of Bias Checklist appraises the methodological quality of a study's evaluation of a measurement property and serves to identify risk of bias in assessment of each measurement property. Thus, the methods for each measurement property assessed in each included study were evaluated independently and rated as "very good", "adequate", "doubtful", "inadequate", or "not applicable." The COSMIN Risk of Bias Checklist includes 3 to 31 standards per measurement property for a total of 81 items categorized into the nine measurement properties.[19] The COSMIN Risk of Bias checklist for studies on reliability, relevant to clinician-reported outcome measurement instruments, was used for assessments of inter-rater reliability.[25] The kappa score between reviewers was 0.70 and the remaining discrepancies were resolved through discussion. NJB verified all Risk of Bias results.

As per COSMIN guidelines, to determine the overall quality of the method evaluating a measurement property, the "worst score counts" principle was used (i.e., the lowest rating given for the method of assessing a measurement property was the overall rating).[18,28] For example, if all of the COSMIN Risk of Bias Checklist items that assess the methodology of measuring structural validity were rated as "very good" methodology for a study except for one item that was rated as "doubtful" methodology, the overall rating of the methodology of how the study established structural validity would have been classified as "doubtful" for that study.

Sufficiency of the Measurement Properties

The results of each study were rated against the COSMIN criteria for a "sufficient measurement property" (Table 3) by two reviewers independently (ES, MO). The kappa score

between reviewers was 0.74 and remaining discrepancies were resolved through discussion between the reviewers. Table 3 summarizes the criteria used for designating "sufficient," "insufficient," and "indeterminate" measurement properties from the COSMIN guidelines' criteria for good measurement properties.[18]

The studies that used hypothesis testing for construct validity[27,29-31] and responsiveness[27,31] via comparison with other instruments did not define any hypotheses (i.e., in what direction and magnitude a comparator instrument is hypothesized to relate to the CDRS-R). Our review team hypothesized the direction and magnitude of how the comparator instrument should relate to the CDRS-R, as per the COSMIN guideline for situations in which the study did not define a hypothesis regarding the relationship to the comparator instrument.[18] The comparator instruments used in the included studies that measured hypothesis testing for construct validity[27,29-31] of the CDRS-R were: the Reynolds Adolescent Depression Scale,[32] Hamilton Depression Scale (HDRS),[13] Children's Depression Inventory (CDI),[33] the Beck's Depression Inventory (BDI),[34] Clinical Global Impression-Severity (CGI-S),[35] Clinical Global Impression-Improvement (CGI-I),[35] the Children's Global Assessment Scale (CGAS),[36] a depression subscale score generated from the Kiddie-Schedule for Affective Disorders and Schizophrenia-Present and Lifetime (KSADS-PL) version by the authors of the comparator study,[27,37,38] and subscales on the Child Behavior Checklist (CBCL).[39-41] The CGI-I was used to assess responsiveness on the English CDRS-R by correlating change scores on the CDRS-R with CGI-I total scores.[31] The CDI, BDI, CGI-S, CGAS, and subscales on the CBCL[27] were used to assess responsiveness on the Korean CDRS-R by correlating mean change scores with mean CDRS-R change scores.

The Reynolds Adolescent Depression Scale,[32] HDRS,[13] CDI,[33] BDI,[34] and KSADS-PL depression "subscale" measure depressive symptoms. The COSMIN guidelines[18] suggest that correlations with instruments measuring similar constructs should be greater than or equal to 0.50 for sufficient hypothesis testing for construct validity and responsiveness. We thus hypothesized that the Pearson r correlation between each of these respective OMIs' total scores and CDRS-R total scores would have a positive correlation that was greater than or equal to 0.5[18] because these OMIs are thought to measure a similar construct as the CDRS-R and like the CDRS-R, higher scores on these instruments are interpreted as more severe depression symptoms.[18] Relatedly, we hypothesized correlations to be greater than 0.5 for the following assessed CBCL subscales: withdrawn, anxious/depressed, and internalizing problems, for construct validity and responsiveness, respectively.[28] The same hypotheses were used for assessing responsiveness via mean change scores for the CDI, BDI, and CBCL subscales.

The CGI-S[35] assesses global illness severity and the CGI-I[35] measures illness improvement. The COSMIN guidelines suggest that correlations with instruments measuring related but dissimilar constructs should have lower correlations, e.g., 0.30 to 0.50, when determining the sufficiency of hypothesis testing for construct validity and responsiveness.[18] Thus, we hypothesized the r correlation of scores between each of these respective OMIs' scores and the CDRS-R's scores would have a positive correlation greater than 0.3; a higher score on these instruments indicate greater illness severity, or worsened symptoms/functioning, respectively. The same hypothesis was used for correlations between mean change scores on the CGI-S and mean change scores on the CDRS-R with an r value greater than 0.3. We hypothesized a negative correlation of -0.3 or less (e.g., r = -0.4, -0.5) between mean change scores on the CDRS-R and CGI-I scores. We hypothesized the CDRS-R correlation to be greater

than 0.3 for the following related but dissimilar constructs of the CBCL: delinquent behaviour, aggressive behaviour, externalizing problems.[28] For the CGAS[36] comparator instrument, which measures the child's overall general functioning, we hypothesized a negative correlation less than -0.30[18] with CDRS-R scores, and mean change scores of the CDRS-R and CGAS; a higher score on the CGAS is defined as a better overall global functioning ability and the construct "child's functioning" is dissimilar but related to the construct of "depressive symptom severity."

Synthesis of Results

Two reviewers (ES and MO) independently summarized the evidence for sufficient measurement properties across studies.[18,19,22] There were no discrepancies between reviewers. As per the COSMIN guidelines, we prospectively defined that if 75% of the results or greater in all studies were rated as having a "sufficient" measurement property (Table 3), the overall rating of that measurement property would be defined as "sufficient."[18] For example, if at least 75% of studies assessing structural validity found that structural validity of the CDRS-R fulfilled the COSMIN criteria for a "sufficient" measurement property (Table 3), then structural validity would be reported as "sufficient."[18] If the results across studies were inconsistent for unexplained reasons (e.g., we found a measurement property to be "sufficient" in similar studies, while in other similar studies we found it to be "insufficient"), and less than 75% of the findings across studies were consistent, the findings would be graded as inconsistent.

Adapted GRADE Approach

Two reviewers (ES and SM) independently graded the quality of the evidence across studies using the modified version of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach adapted by COSMIN.[18] The kappa score between reviewers was 0.85; discrepancies were resolved through discussion. NJB verified all

GRADE results. The quality of evidence across all included studies was graded as one of four predefined levels of quality of evidence: high, moderate, low, or very low evidence (see Table S3, available online, for definitions). This adapted GRADE approach is explained in detail in the COSMIN guidelines.[18] In brief, quality of evidence was downgraded when there was risk of bias (i.e., related to quality of the study methodology), unexplained inconsistency (i.e., if the results of included studies were inconsistent), imprecision (i.e., related to total sample size of the available studies), and/or indirectness of evidence (i.e., evidence derived from populations other than adolescents with MDD).[18] For example, if two studies that assessed internal consistency were rated as having "adequate methodological quality" (i.e., low risk of bias), their findings were consistent, total sample size across both studies was more than 100, and both studies included only adolescents with MDD (i.e., direct evidence), then the overall level of evidence for internal consistency would be graded as "high."

For indirectness, the COSMIN guidelines state that if the studies of a measurement property are performed in a population other than the specific population of interest, the reviewer can downgrade the level of evidence by one or two levels for "serious" or "very serious indirectness."[18] In this review, for studies that included participants who were outside the adolescent age range (i.e., 12 to 18 years old) and/or were not diagnosed with MDD (i.e., less than 100% of study population was diagnosed with MDD), the evidence was downgraded one level due to "serious indirectness."

Recommendation Category

To formulate recommendations based on our findings for the use of the CDRS-R in adolescents with MDD, we applied COSMIN guidelines to categorize the CDRS-R into one of the following categories[18]:

(A) "Evidence for sufficient content validity (any level) AND at least low quality evidence for sufficient internal consistency." *Outcome measurement instruments in category A can be recommended for use.*

(B) "Categorized not in A or C." *The measurement properties of outcome measurement instruments in category B need to be further appraised before recommending for use.*

(C) "High quality evidence for an insufficient measurement property." *Outcome measurement instruments in category C should not be recommended for use.*

RESULTS

The complete dataset can be accessed on our Open Science Framework webpage.[23]

Study Search and Selection

The search yielded 2272 studies; after duplicates were removed there were 1381 studies for title and abstract screening. This resulted in 65 eligible studies for full-text screening, of which six met eligibility criteria. See Figure 1 for reasons for exclusion after full-text screening. Two of the included studies were published in English but assessed CDRS-R use in another language (Korean and German, respectively).[27,30] Study characteristics are described in Table 1.

Reported Measurement Properties

None of the six included studies assessed all nine COSMIN measurement properties (Tables 1 and 5). No studies assessed content validity, cross-cultural validity, or measurement error of the CDRS-R in adolescents. Three studies assessed structural validity (i.e., dimensionality),[27,42,43] five studies assessed internal consistency,[27,30,31,42,43] two studies assessed inter-rater reliability,[29,27] four studies assessed hypothesis testing for construct validity,[27,29-31] and two studies assessed responsiveness.[27,31]

Methodological Quality of Included Studies: Assessing Risk of Bias

Table 5 summarizes the Risk of Bias checklist results from assessing the methodological quality of each measurement property from the included studies.[19] There was at least one study appraised as having "very good" or "adequate" methodological quality for each of the following measurement properties: structural validity,[36,43] internal consistency,[27,43] and hypothesis testing for construct validity (Reynolds Adolescent Depression Scale only[29]). No study was found with "very good" or "adequate" methodology for reliability (inter-rater),[29] hypothesis testing for construct validity (all other comparator instruments)[27,29-31] and responsiveness[27,31] (all comparator instruments). Content validity, cross-cultural validity, and measurement error of the CDRS-R were not assessed in any study.

Sufficiency of the Measurement Properties and Synthesis of Results

Table 5 shows the sufficiency rating of the measurement properties in each included study, as well as the overall sufficiency rating of the measurement properties (i.e., whether the measurement property was found to be sufficient across all included studies, as according to the COSMIN criteria; Table 3).[18] Overall, there were three measurement properties that were found to be sufficient based on the overall evidence: reliability (inter-rater; ICC=0.98 on English version; ICC=0.96 on the Korean version),[29,27] hypothesis testing for construct validity,[27,29-31] and responsiveness.[27,31] Structural validity[27,42,43] and internal consistency[27,30,31,42,43] were rated as indeterminate.

As predicted for hypothesis testing for construct validity the Reynolds Adolescent Depression Scale (n=1 study; r=0.77),[29] the HRDS (n=1 study; r=0.92),[29] the CDI (n=2 studies; r=0.78, r=0.86),[27,29] and the BDI (n=1 study, r=0.85),[27] each had a positive correlation with the CDRS-R greater than 0.5 (Table 5).[27,29] As predicted, the CGI-S (n=3 studies)[27,30,31] and the CGI-I (n=1 study)[31] scores each had a positive correlation with the CDRS-R scores that was

greater than 0.3 (r=0.55 to 0.93, r=0.92, respectively).[27,30,31] As predicted, the CGAS scores had a negative correlation less than -0.3 with the CDRS-R scores (r=-0.52 to -0.86).[27,31] As predicted, the CBCL scores of the withdrawn, anxious/depressed, and internalizing problems subscales correlations were greater than 0.5 (n=1 study; r=0.65, 0.67, 0.66, respectively).[27] As predicted, the CBCL scores of the delinquent behaviour, aggressive behaviour, and externalizing problem subscales were greater than 0.3 (n=1 study; r=0.48, 0.5, 0.45, respectively). Thus, after pooling the results of hypothesis testing for construct validity across studies, we rated hypothesis testing for construct validity as a sufficient measurement property (Table 5).

Responsiveness also met criteria for a sufficient measurement property (Table 5) though findings were mixed.[29] On the English language version, CDRS-R change in scores correlated with CGI-I scores, as predicted (r=-0.82).[31] The rest of the results were derived from the Korean version. The correlations between mean change scores on the CDRS and its most direct comparators, the CDI and BDI, nearly met our hypothesis of a positive correlation greater than 0.5 (r=0.46 each).[27] As predicted, the CGI-S had a positive correlation with CDRS-R scores that was greater than 0.3 (r=0.82).[27] As predicted, CGAS mean change scores had a negative correlation less than -0.3 with the CDRS-R mean change scores (r=-0.80).[27] In contrast to our predictions of a correlation greater than 0.5 between the CDRS-R and the CBCL subscale mean change scores for the withdrawn, anxious/depressed, and internalizing problems constructs, correlations were r=0.38, 0.44, and 0.37, respectively. In contrast to our predictions of a correlation greater than 0.3 between the CDRS-R and the CBCL subscale mean change scores for the delinquent behaviour, aggressive behaviour, and externalizing problem constructs, correlations were 0.11, 0.24, 0.25, respectively. Given the sufficient correlations demonstrated between the CDRS-R and the CGI-I on the English language version of the CDRS-R, and the

CGI-S and CGAS (and nearly sufficient correlations with the BDI and CDI) in the Korean study, CDRS-R responsiveness was overall rated as sufficient.

Structural validity and internal consistency were rated as indeterminate because it is unknown if these measurement properties of the CDRS-R are sufficient (Table 5). The studies that assessed structural validity only performed exploratory factor analyses, found factors that differed in number and content, and did not include a confirmatory factor analysis.[27,42,43] Internal consistency was rated as indeterminate because there was no evidence of sufficient structural validity,[27,30,31,42,43] which is required for a sufficient rating of internal consistency (Table 3).

Adapted GRADE Approach

No measurement property was graded as having "high" quality of evidence using the adapted GRADE approach (Tables 4 and 5). Each measurement property that had been evaluated in the included studies was downgraded, due to risk of bias in the methodology, and/or the indirectness (i.e., heterogeneous samples where the samples were not solely comprised of adolescents with MDD). No studies were downgraded due to inconsistencies or imprecision in results across studies.

Evaluations of measurement properties were downgraded to either moderate, low or very quality of evidence (Tables 4 and 5), as follows:

Structural validity was downgraded to "moderate" level of evidence due to the indirectness of evidence (i.e., evidence derived from a different population), as one of the three studies that assessed structural validity included patients without MDD in their structural validity analysis[27] (Table 1). Since there were two studies of "adequate methodological quality," structural validity was not downgraded for risk of bias.

Internal consistency was downgraded to "moderate" level of evidence due to indirectness of evidence for similar reasons, as studies included patients without MDD[27,31] and children under the age of 12 in the final analysis (Table 1).[42,43] Since there was at least one study of "very good methodological quality," internal consistency was not downgraded for risk of bias.

Reliability (inter-rater) was downgraded to "very low" level of evidence due to risk of bias and indirectness of evidence. Risk of bias was found because in one study, the assessors completed the CDRS-R at the same time as other measures (i.e., the Schedule for Affective Disorders and Schizophrenia (SADS), Present Episode version and the HDRS), and CDRS-R scores were primarily based on the information obtained from the SADS interview questions.[29] The other study did not report any details of reliability assessment methodology including timing of assessments.[27]

Hypothesis testing for construct validity was downgraded to "low" quality of evidence as all four studies that evaluated this were rated as having high risk of bias (i.e., multiple studies of "doubtful" methodological quality and one study of "adequate" methodological quality available)[27,29-31] and indirectness (i.e., the overall sample size had participants who were not diagnosed with MDD).[27,31] Issues of methodological quality across studies that assessed hypothesis testing for construct validity were largely related to the overall lack of evidence provided and found by our review team for the measurement properties of the comparator instrument in adolescents with MDD. This information is critical to report to demonstrate that the comparator instrument is actually fit for purpose as the comparator for assessing the use of the CDRS-R in adolescents with MDD.

Responsiveness, assessed by only two studies,[27,31] was downgraded to " low" quality of evidence for the same reasons as hypothesis testing for construct validity, due to risk of bias

(related to the overall lack of reporting and evidence for the measurement properties of the comparator instruments in adolescents with MDD), and indirectness (due to overall sample including participants without MDD).[27]

Recommendation Category

From the collective results of this review, the CDRS-R falls into Category B of COSMIN's guidelines for recommending outcome measurement instruments, such that "the measurement properties of the CDRS-R need to be further appraised before recommending for use."[18]

DISCUSSION

The evidence for the CDRS-R, the most commonly used OMI in studies assessing treatments of adolescents with MDD,[6,7] is limited. Only six relevant studies that spanned three language versions were found for evaluation, and no studies that assessed content validity, cross-cultural validity/measurement variance, or measurement error in adolescents with MDD were identified. The risk of bias in the methodology used for evaluating each measurement property was frequently found to be of "doubtful" or "inadequate" methodological quality.

Notably, the CDRS-R met COSMIN criteria for sufficient inter-rater reliability, hypothesis testing for construct validity, and overall, for responsiveness. However, the level of evidence for these measurement properties using the adapted GRADE approach ranged from low to very low, due to methodological risk of bias and/or the inclusion of children and participants without MDD. The findings for the structural validity and internal consistency of the CDRS-R in adolescents with MDD were inconclusive, due to a lack of the appropriate analysis for sufficient structural validity (i.e., a confirmatory factor analysis was not conducted); evidence of sufficient structural validity is required for sufficient internal consistency. Notably, exploratory factor

analyses results indicated that the structure of the CDRS-R is not unidimensional, with between two and five factors found across studies. As for other commonly used depression scales that rely on sum scores, this creates challenges in the interpretability of the CDRS-R total sum score and its change scores for change in depressive symptom severity.[44]

According to the COSMIN guidelines, the minimal requirements for recommending the use of an OMI in a specific population is that the OMI must have demonstrated sufficient content validity (with any quality of evidence) and at least low quality of evidence for internal consistency.[18] We found a striking absence of studies assessing content validity of the CDRS-R in adolescents, a tool that was originally developed for children ages six to twelve years old.[10] To assess the content validity of any clinician-reported OMI, clinicians' input on the relevance (i.e., all items are appropriate), comprehensiveness (i.e., all concepts are captured by the OMI), and comprehensibility (i.e., all items are well-understood by the assessor) of the OMI are essential to evaluate.[22,45] In the case of the CDRS-R, patient input would also be important to obtain to determine if the patient understands what they are being asked during the semi-structured interview that informs completion of items 1 through 15 (i.e., comprehensibility of the patient).[45] In the absence of a complete assessment of content validity in adolescents, it remains unclear whether the CDRS-R scores reported in past evaluative studies[6,7] are valid reflections of change in depressive symptom severity. For example, some item scores on the CDRS-R could potentially be related to age-specific issues in adolescents rather than a symptom of depression itself (e.g., fatigue or self-esteem).

The second COSMIN minimal requirement for recommending an OMI for use, internal consistency, was found to be indeterminate in the context of CDRS-R use in adolescents because there was no evidence from the included studies of sufficient structural validity[18,19,22]; evidence

of sufficient structural validity is required for sufficient internal consistency. It is therefore uncertain if the scores of the CDRS-R are an adequate reflection of the dimensionality of the depressive symptom severity.

The current quality of evidence of the measurement properties of the CDRS-R in adolescents with MDD ranged from moderate to very low. Any future studies assessing measurement properties of the CDRS-R should consider using the newly developed COSMIN design checklist,[46] and follow recommended methodology,[46] an appropriate study sample, and completely report the comparator instrument's characteristics (e.g., the construct the comparator instrument measures and measurement properties), and run the appropriate statistical methods (e.g., confirmatory factor analysis for structural validity).[46] There is currently no consensus on how to best measure depression in adolescents in the context of clinical trials.[7] As part of an international effort to develop a core outcome set for adolescent depression clinical trials,[16] other OMIs used to measure depression symptom severity in adolescents[6,7] will be evaluated in order to help identify the best tools for use.

The strengths of the current study include the comprehensive sensitive search and selection methods which ensure that no major evaluative studies were missed, and the application of a comprehensive, rigorous, and well validated OMI study appraisal system developed by COSMIN.[18,19,22] Additionally, the complex appraisal process was undertaken using consensus methods by reviewers with a breadth of knowledge in mental health, clinical research methodology, and pediatric mental health clinical work. All phases of this review involved two reviewers, with initial inter-rater agreement rates that ranged from 80 to 100%. Any identified discrepancies were due to overlooking a detail in an included study, which was readily resolved

through reviewer discussion. There were no domains that were consistently rated differently between raters (e.g., a consistent discrepancy for ratings of structural validity).

There are several limitations to this review. First, this review focused on studies with a majority sample of adolescents diagnosed with MDD and excluded studies outside of this scope. As such, our study findings do not pertain to the measurement properties of the CDRS-R used for children under the age of twelve. The CDRS-R was originally designed for use in children in the six to twelve year age group,[11,12] and studies have looked at measurement properties such as (but not limited to) content validity,[11] internal consistency,[47] reliability,[48] and hypothesis testing for construct validity.[47,49] To our knowledge, however, such studies have not been appraised using the systematic methods here (i.e., through application of the COSMIN guidelines and adapted GRADE approach). We defined adolescents as youth ages 12 to 18 years, in accordance with published guidance,[21] and as the included studies used varying definitions of adolescents (e.g., youth ages 12-17, 12-18, 13-18 years), the definition used helped ensure that we captured the few relevant studies that evaluated the measurement properties of the CDRS-R in an adolescent population.

Second, two of the six included studies assessed the CDRS-R in Korean and German, respectively.[27,30] The findings of these studies may not be generalizable to the English language version of the CDRS-R items.[50] Given the limited evidence on the measurement properties of the CDRS-R for adolescent MDD, there were not enough studies to assess the measurement properties of the CDRS-R per language version separately. Third, the method used in this review, the COSMIN guidelines, are used to appraise patient-reported OMIs, while the CDRS-R is a clinician-reported OMI. However, the COSMIN guideline states that their appraisal methodology can be used for clinician-reported OMIs with modifications, as appropriate,[18] and

has been used for clinician-reported OMIs previously.[51,52] The recently released modified risk of bias checklist relevant to clinician-reported OMIs for studies of reliability was implemented.[25] In this review, no adaptations were needed to be made to the COSMIN methodology.

It is unclear whether the CDRS-R appropriately measures depressive symptom severity in adolescents with MDD. There is an imperative to ensure that the results from clinical research studies in adolescents with MDD are meaningful and useful to relevant stakeholders, including patients, caregivers, clinicians, researchers, and policy-makers. Before the use of the CDRS-R in adolescent MDD can be recommended, it needs to meet the minimum COSMIN requirements for measurement properties: sufficient content validity and internal consistency in adolescents.

References

1. World Health Organization. *Depression and other common mental disorders: global health estimates.* Geneva: World Health Organization;2017.
2. Avenevoli S, Swendsen J, He J, Burstein M, Merikangas K. Major depression in the national comorbidity survey-adolescent supplement: Prevalence, correlates, and treatment. *J Am Acad Child Adolesc Psychiatry.* 2015;54(1):37-44.
3. Weissman MM, Wolk S, Goldstein RB, et al. Depressed adolescents grown up. *JAMA.* 1999;281(18):1707-1713.
4. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. *Lancet.* 2012;379(9820):1056-1067.
5. Arroll B, Macgillivray S, Ogston S, et al. Efficacy and tolerability of tricyclic antidepressants and SSRIs compared with placebo for treatment of depression in primary care: A meta-analysis. *Ann Fam Med.* 2005;3(5):449-456.
6. Krause K, Bear H, Edbrooke-Childs J, Wolpert M. What outcomes count? Outcomes measured for adolescent depression between 2007 and 2017. *J Am Acad Child Adolesc Psychiatry.* 2019;58(1):61-71.
7. Mew EJ, Monsour A, Saeed L, et al. Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. *J Clin Epidemiol.* 2020;126:71-79.
8. Szatmari P, Offringa M, Butcher N, Monga S. Counting what counts: The case for harmonized outcomes in child youth and mental health research. *J Am Acad Child Adolesc Psychiatry.* 2019;58(7):656-658.
9. Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *J Affect Disord.* 2017;208:191-197.
10. Poznanski E, Mokros H. *Children's Depression Rating Scale-Revised (CDRS-R).* Los Angeles: WPS; 1996.
11. Poznanski E, Cook SC, Carroll BJ. A depression rating scale for children. *Pediatrics.* 1979;64(4):442-2450.
12. Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating Scale - Revised. *Psychopharmacol Bull.* 1985;21(4):979-989.
13. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23(1):56.
14. Poznanski E, Grossman J, Buchsbaum Y, Banegas M, Freeman L, Gibbons R. Preliminary studies of the reliability and validity of the Children's Depression Rating Scale. *J Am Acad Child Adolesc Psychiatry.* 1984;21(4):979-989.
15. Monga S, Butcher N, Monsour A, Mew E, Szatmari P, Offringa M. Core set of outcomes for adolescents with major depressive disorder: A tool of standardized outcomes for clinical research. http://www.comet-initiative.org/studies/details/1122?result=true%E2%80%8B. Published 2018. Accessed November 11, 2019.
16. Monga S, Monsour A, Stallwood E, et al. Core outcome set development for adolescent major depressive disorder clinical trials: A registered report. *J Am Acad Child Adolesc Psychiatry, in press.* 2020.
17. Moher D, Liberati A, Tetzlaff J, Altman D, PRISMA GROUP. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* 2009;6(7):e1000097.

18. Prinsen C, Mokkink L, Bouter L, et al. COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res.* 2018;27(5):1147-1157.
19. Mokkink L, de Vet H, Prinsen C, et al. COSMIN Risk of Bias Checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1171-1179.
20. Terwee C, Prinsen C, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Qual Life Res.* 2018;27(5):1159-1170.
21. Williams K, Thomson D, Seto I, et al. Standard 6: Age groups for pediatric trials. *Pediatrics.* 2012;129(3):1159-1170.
22. Terwee C, Prinsen C, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Qual Life Res.* 2018;27(5):1159-1170.
23. Stallwood E, Monsour A, Rodrigues C, et al. Systematic review: The measurement properties of the Children's Depression Rating Scale-Revised for adolescents with major depressive disorder. https://osf.io/crfm8/. Published 2019. Accessed December 11, 2019.
24. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde J. Research electronic data capture (REDCap) - a meta data-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377-381.
25. Mokkink LB, Boers M, Vleuten Cvd, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol.* 2020;preprint.
26. Prinsen C, Vohra S, MR. R, et al. Guideline for selecting outcome measurement instruments for outcomes included in a Core Outcome Set. https://cosmin.nl/wp-content/uploads/COSMIN-guideline-selecting-outcome-measurement-COS.pdf. Published 2016. Accessed August 23, 2019.
27. Kim K, Nam S, Choi J, et al. Psychometric properties and factor structures of the korean version of Children's Depression Rating Scale-Revised. *J Child Adolesc Psychopharmacol.* 2018;28(4):285-292.
28. Terwee C, Mokkink L, Knol D, Ostelo R, Bouter L, de Vet H. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Qual Life Res.* 2012;21(4):651-657.
29. Shain BN. Comparison of self-rated and clinican-rated measures of depression in adolescents. *Am J Psychiatry.* 1990;147(6):793-795.
30. Plener PL, Grieb J, Sprober N, et al. Convergence of Children's Depression Rating Scale-revised scores and clinical diagnosis in rating adolescent depressive symptomatology. *Mental Illness.* 2012;4(1):29-31.
31. Mayes TL, Bernstein IH, Haley CL, Kennard BD, Emslie GJ. Psychometric properties of the Children's Depression Rating Scale-Revised in adolescents. *J Child Adolesc Psychopharmacol.* 2010;20(6):513-516.
32. Reynolds WM. Reynolds Adolescent Depression Scale. *The Corsini Encyclopedia of Psychology.* 2010:1.
33. Kovacs M. Rating scales to assess depression in school-aged children. *Acta Paedopsychiatrica: Int J Child & Adolesc Psychiatry.* 1981;46(5-6):305-315.
34. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry.* 1961;4(6):561-571.

35. Guy W. *ECDEU assessment manual for psychopharmacology.* US Department of Health, Education, and Welfare, Public Health Service; 1976.
36. Shaffer D, Gould MS, Brasic J, et al. A Children's Global Assessment Scale (CGAS). *Arch Gen Psychiatry.* 1983;40(11):1228-1231.
37. Kaufman J, Birmaher B, Brent D, et al. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): Initial reliability and validity data. *J Am Acad Child Adolesc Psychiatry.* 1997;36(7):980-988.
38. Kim Y, Cheon K, Kim B, et al. The reliability and validity of Kiddie Schedule for Affective Disorders and Schizophrenia Present and Lifetime version Korean version (K-SADS-PL-K). *Yonsei Med J.* 2004;45:81-89.
39. Achenbach T, Edelbrock C. *Manual for the Child Behavior Checklist and Revised Behavior Profile.* Burlington, Vermont: University of Vermont Department of Psychiatry; 1983.
40. Achenbach T, Rescorla L. *Manual for the ASEBA school-age forms & profiles: An integrated system of mult-informant assessment.* Burlington, Vermont: University of Vermont, Research Center for Children, Youth and Families; 2001.
41. Achenbach T, Department of Psychiatry UoV. *Manual for the Child Behavior Checklist/4-18 and 1991 profile.* Burlington, Vermont: Department of Psychiatry, University of Vermont; 1991.
42. Isa A, Bernstein I, Trivedi M, Mayes T, Kennard B, Emslie G. Childhood depression subscales using repeated sessions on Children's Depression Rating Scale - Revised (CDRS-R) scores. *J Child Adolesc Psychopharmacol.* 2014;24(6):318-324.
43. Guo Y, Nilsson ME, Heiligenstein J, Wilson MG, Emslie G. An exploratory factor analysis of the Children's Depression Rating Scale-Revised. *J Child Adolesc Psychopharmacol.* 2006;16(4):482-491.
44. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychol Assess.* 2016;28(11):1354-1367.
45. Powers JH, Patrick DL, Walton MK, et al. Clinician-reported outcome assessments of treatment benefit: Report of the ISPOR clinical outcome assessment emerging good practices task force. *Value Health.* 2017;20(1):2-14.
46. Mokkink L, Prinsen C, Patrick D, et al. COSMIN study design checklist for patient-reported outcome measurement instruments. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf#. Published 2019. Accessed October 7, 2019.
47. Jain S, Carmody TJ, Trivedi MH, et al. A psychometric evaluation of the CDRS and MADRS in assessing depressive symptoms in children. *J Am Acad Child Adolesc Psychiatry.* 2007;46(9):1204-1212.
48. Myers K, Winters NC. Ten-year review of rating scales. II: Scales for internalizing disorders. *J Am Acad Child Adolesc Psychiatry.* 2002;41(6):634-659.
49. Yee AM, Algorta GP, Youngstrom EA, Findling RL, Birmaher B, Fristad MA. Unfiltered administration of the YMRS and CDRS-R in a clinical sample of children. *J Clin Child Adolesc Psychol.* 2015;44(6):992-1007.
50. Breugelmans R. Dangers in using translated medical questionnaires. *Chest.* 2009;136(4):1175-1177.

51. Jaspers M, van Haasterecht L, van Zuijlen P, LB. M. A systematic review on the quality of measurement techniques for the assessment of burn wound depth and healing potential. *Burns.* 2018;45(2):261-281.
52. Abedi A, Mokkink L, Zadegan S, et al. Reliability and validity of the AOSpine thoracolumber injury classification system: A systematic review. *Global Spine J.* 2019;9(2):231-242.
53. Mokkink L, Terwee C, Patrick D, 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):737-745.

Table 1. Characteristics of the Included Studies Assessing the Measurement Properties of the CDRS-R

First Author, Publication Year	N of Participants at Enrolment	Age Range (years) [subgroup analysis age range in years, if applicable]	Mean Age (years)	% Female	% with MDD	Country & CDRS Language Version	CDRS-R Rapporteur	Measurement Properties Measured[a]	Diagnostic Assessment Criteria
Kim 2018[27]	66	12-17	Not Reported	61	55	South Korea / Korean	Interviewers (occupation not reported; n=2[b])	☒Structural validity ☒ Internal consistency ☒Reliability ☒ Hypothesis testing for construct validity ☒Responsiveness	K-SADS-PL
Isa 2014[42]	234 (includes participants from Mayes 2010[31])	8-17	12.6	44	100	United States / English	Treating clinician (n=not reported)	☒Structural validity ☒ Internal consistency	Clinical diagnosis of MDD
Plener 2012[30]	1	N/A (one subject, age 15)	N/A	100	100	Germany / German	Trainees, professionals, others (n=32)	☒ Internal consistency ☒ Hypothesis testing for construct validity	ICD-10
Mayes 2010[31]	152[c]	12-18	Not reported	47	100[c]	United States/ English	Child psychiatrist (n=not reported)	☒ Internal consistency ☒ Hypothesis testing for construct validity ☒Responsiveness	K-SADS-PL
Guo 2006[43]	315 [144 adolescents]	7-17 [13-18]	12.7	48	100	United States/ English	Not reported	☒Structural validity ☒ Internal consistency	Clinical diagnosis of MDD
Shain	45	12-18	14.9	64	100	United	Interviewer[d]	☒Reliability	DSM-III-R

1990[29]						States/ English		☒ Hypothesis testing for construct validity	

Note: CDRS-R = Children's Depression Rating Scale-Revised; DSM-III-R = Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised; ICD-10 = International Statistical Classification of Diseases and Related Health Problems, tenth version; K-SADS-PL = Kiddie-Schedule for Affective Disorders and Schizophrenia-Present and Lifetime version; MDD = Major Depressive Disorder; N/A = not applicable.

[a]According to Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) definitions of measurement properties;[53] Measurement properties are ordered in order of importance specified by COSMIN;[18] see Table 2 for definitions.
[b]Master of Education and Master of Art degree indicated.
[c]Of the 152 enrolled, the n=94 and n=88 with a diagnosis of MDD were assessed on the included measurement properties.
[d]CDRS-R likely performed and scored by a psychiatrist; reporting unclear.

Table 2. COSMIN Definitions of Measurement Properties

Term			Definition
Domain	**Measurement Property**	**Measurement Property Aspect**	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the CDRS-R (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal Consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to "true"[a] differences between patients
	Measurement Error		The systematic and random error of a patient's score that is not attributed to "true"[a] changes in the construct to be measured
Validity			The degree to which the CDRS-R measures the construct(s) it purports to measure (depressive symptom severity)
	Content Validity		The degree to which the content of the CDRS-R is an adequate reflection of the construct to be measured
		Face Validity	The degree to which (the items of) the CDRS-R indeed looks as though they are an adequate reflection of the construct to be measured
	Construct Validity		The degree to which the scores of the CDRS-R are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the CDRS-R validly measures the construct to be measured
		Structural Validity	The degree to which the scores of the CDRS-R are an adequate reflection of the dimensionality of the construct to be measured
		Hypothesis Testing	Item construct validity
		Cross-cultural validity/measurement invariance[b]	The degree to which the performance of the items on a translated or culturally adapted CDRS-R are an adequate reflection of the performance of the items of the original version of the CDRS-R.
Responsiveness	Responsiveness		The ability of the CDRS-R to detect change over time in the construct to be measured

Note: Adapted from Mokkink et al[53] (https://doi.org/10.1016/j.jclinepi.2010.02.006; published under Creative Commons license CC BY: https://creativecommons.org/licenses/by/4.0/). CDRS-R = Children's Depression Rating Scale-Revised; COSMIN = COnsensus-based Standards for the selection of health Measurement Instruments.[18]

[a]The word "true" means any observation is composed of two components: a true score and error associated with the observation. "True" is the average score that would be obtained if the scale were given an infinite number of times. It refers to the consistency of the score and not to its accuracy.
[b]Relevant to studies that have evaluated cross-cultural validity for instruments across "culturally different populations," which COSMIN defines broadly and inclusive of e.g., different ethnicities, language groups, age groups, or genders[19]

Table 3. COSMIN Criteria for Assessing Measurement Properties as Sufficient, Insufficient, or Indeterminate

Measurement Property	Rating[a]	Criteria for Rating
Structural Validity	+	Classical Test Theory: Confirmatory Factor Analysis: Comparative Fit Index or Tucker-Lewis Index or comparable measure >0.95 OR Root Mean Square Error of Approximation <0.06 or Standardized Root Mean Residuals <0.08[2]
	?	Classical Test Theory: Not all information for + reported
	-	Criteria for + not met
Internal Consistency	+	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale
	?	Criteria for "At least low evidence for sufficient structural validity" not met
	-	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale
Cross-Cultural Validity/Measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important differential item functioning for group factors (McFadden's R2 < 0.02)
	?	No multiple group factor analysis OR differential item functioning analysis performed
	-	Important differences between group factors OR differential item functioning was found
Reliability	+	Intraclass correlation coefficient or weighted Kappa ≥ 0.70
	?	Intraclass correlation coefficient or weighted Kappa not reported
	-	Intraclass correlation coefficient or weighted Kappa < 0.70
Measurement Error	+	Smallest detectable change or limits of agreement < minimal important change
	?	Minimal important change not defined
	-	Smallest Detectable Change or Limits of Agreement > Minimal Important Change
Criterion Validity	+	Correlation with gold standard ≥ 0.70 OR Area Under the Curve ≥ 0.70
	?	Not all information for + reported
	-	Correlation with gold standard < 0.70 OR Area Under the Curve < 0.70
Hypothesis Testing for Construct Validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis
Responsiveness	+	The result is in accordance with the hypothesis OR area under the curve ≥ 0.70
	?	No hypothesis defined (by the review team)

	-	The result is not in accordance with the hypothesis OR area under the curve < 0.70

Note: Measurement properties are ordered in order of importance specified by COSMIN. Adapted from Prinsen et al[18] (https://doi.org/10.1007/s11136-018-1798-3; published under Creative Commons license CC BY: https://creativecommons.org/licenses/by/4.0/). COSMIN = COnsensus-based Standards for the selection of health Measurement Instruments.

[a]"+" = sufficient; "-" = insufficient; "?" = indeterminate

Table 4. Levels of Evidence for the Measurement Properties of the Children's Depression Rating Scale-Revised in Adolescents with MDD using the Adapted Grades of Recommendation, Assessment, Development and Evaluation Approach

Adapted GRADE[a] Approach	Methodological Risk of Bias[b]	Inconsistency[c]	Imprecision[d]	Indirectness of Evidence[e]	GRADE Level (total number of evidence downgrades)
Measurement Properties					
Content Validity (no studies assessed)	N/A	N/A	N/A	N/A	**N/A**
Structural Validity (three studies)[27,42,43]	NO Two studies of adequate methodological quality[42,43]	NO	Assessed in risk of bias checklist[f]	YES (-1) One study with 55% study sample with MDD[27]	**Moderate (-1)**
Internal Consistency (five studies)[27,30,31,42,43]	NO Two studies of very good methodological quality[27,43]	NO	NO N>100	YES (-1) One study with 55% study sample with MDD[27,31] Two studies included children in analysis[42,43]	**Moderate (-1)**
Cross-Cultural Validity (no studies assessed)	N/A	N/A	N/A	N/A	**N/A**
Inter-rater Reliability (two studies)[29,27]	YES (-2) One study of doubtful methodological quality and one study of inadequate methodological quality[29,27]	NO	NO N>100	YES (-1) One study with 55% study sample with MDD[27]	**Very Low (-3)**
Measurement	N/A	N/A	N/A	N/A	**N/A**

Error (no studies assessed)					
Hypothesis Testing for Construct Validity (four studies)[27,29-31]	YES (-1) Multiple studies of doubtful quality and one study of adequate quality[27,29-31]	NO	NO N>100	YES (-1) One study with 55% study sample with MDD[27]	**Low (-2)**
Responsiveness (two studies) [27,31]	YES (-1) Multiple studies of doubtful[27,31]	NO[g]	NO N>100	YES (-1) One study with 55% study sample with MDD[27]	**Low (–2)**

Note: MDD = Major depressive disorder; N/A = Not applicable, i.e., no included studies assessed that particular measurement property.

[a]"-1","-2","-3" refer to the level of the downgraded evidence according to COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) adapted Grades of Recommendation, Assessment, Development and Evaluation (GRADE) approach.[18]
[b]The adapted GRADE approach[18] guidelines state that if "There are multiple studies of at least adequate quality, or there is one study of very good quality available," do not downgrade for risk of bias; See Table 5 for details.
[c]There were no inconsistencies of measurement property results across studies.
[d]Downgraded one level if N=50-100, downgraded two levels if N<50, as per the adapted GRADE approach guidelines.[18]
[e]Downgraded one level if all participants did not have MDD and/or are were between the ages of 12-18.
[f]COSMIN's adapted GRADE approach[18] does not downgrade imprecision for structural validity because sample size for structural validity is already considered and graded in Risk of Bias checklist for this measurement property. See COSMIN guidelines for details.[18]
[g]Not downgraded because inconsistent results may be due to different language versions/populations

Table 5. Methodological Quality, Sufficiency of the Measurement Properties, and Overall Level of Evidence Across the Included Studies Appraising the Children's Depression Rating Scale-Revised

	Structural Validity		**Internal Consistency**		**Reliability** (Inter-rater)		**Hypothesis Testing for Construct Validity** (Comparison with other instruments)		**Responsiveness** (Comparison with other instruments)	
Study	**Quality**	**Rating**	**Quality**	**Rating**	**Quality**	**Rating**	**Quality**	**Rating**	**Quality**	**Rating**
Kim et al. 2018[27]	inadequate	?	very good	?	doubtful	+	doubtful[a]	+[a]	doubtful[b] doubtful[c]	+[b] -[c]
Isa et al. 2014[42]	adequate	?	inadequate	?	NA	NA	NA	NA	NA	NA
Plener et al. 2012[30]	NA	NA	doubtful	?	NA	NA	doubtful[d]	+[d]	NA	NA
Mayes et al. 2010[31]	NA	NA	inadequate	?	NA	NA	doubtful[e]	+[e]	doubtful[f]	+[f]
Guo et al. 2006[43]	adequate	?	very good	?	NA	NA	NA	NA	NA	
Shain et al. 1990[29]	NA	NA	NA	NA	inadequate	+	+[g] +[h]	doubtful[g] adequate[h]	NA	
Overall rating of sufficiency of measurement properties	**Indeterminate (?)**		**Indeterminate (?)**		**Sufficient (+)**		**Sufficient (+)**		**Sufficient (+)**	
Quality of evidence (adapted GRADE)[e]	**Moderate**		**Moderate**		**Very Low**		**Low**		**Low**	

Note: No studies assessed content validity, cross-cultural validity, measurement error; NA, indicates measurement property was not assessed by study. Rating of Sufficiency of Measurements Properties Legend: "+" = sufficient measurement property; "?" = indeterminate measurement property. Very good, adequate, doubtful, and inadequate refers to methodological quality of each study for each measurement property using the COSMIN Risk of Bias Checklist; GRADE = Grading of Recommendation, Assessment, Development, and Evaluation Approach. [e]Levels of evidence are defined in Table S3.

[a]Results of the CDRS-R correlations with the CDI, BDI, CGI-S, CGAS at all assessments, and the K-SADS-PL depression subscale score and CBCL at trial baseline.[27]

[b]Results of the CDRS-R mean change correlations with the CGI-S, CGAS mean change correlations.[27]

[c]Results of the CDRS-R mean change correlations with CDI, BDI, CBCL depressive subscale.[27]

[d]Results of the CDRS-R correlations with the CGI-S at trial baseline.[30]

[e]Results of the CDRS-R correlations with the CGI-S, CGI-I, and CGAS at trial baseline and exit.[31]

[f]Results of the CDRS-R mean change correlations with the CGI-I.[31]

[g]Results of the CDRS-R correlations with the CDI, and Hamilton scale.[29]

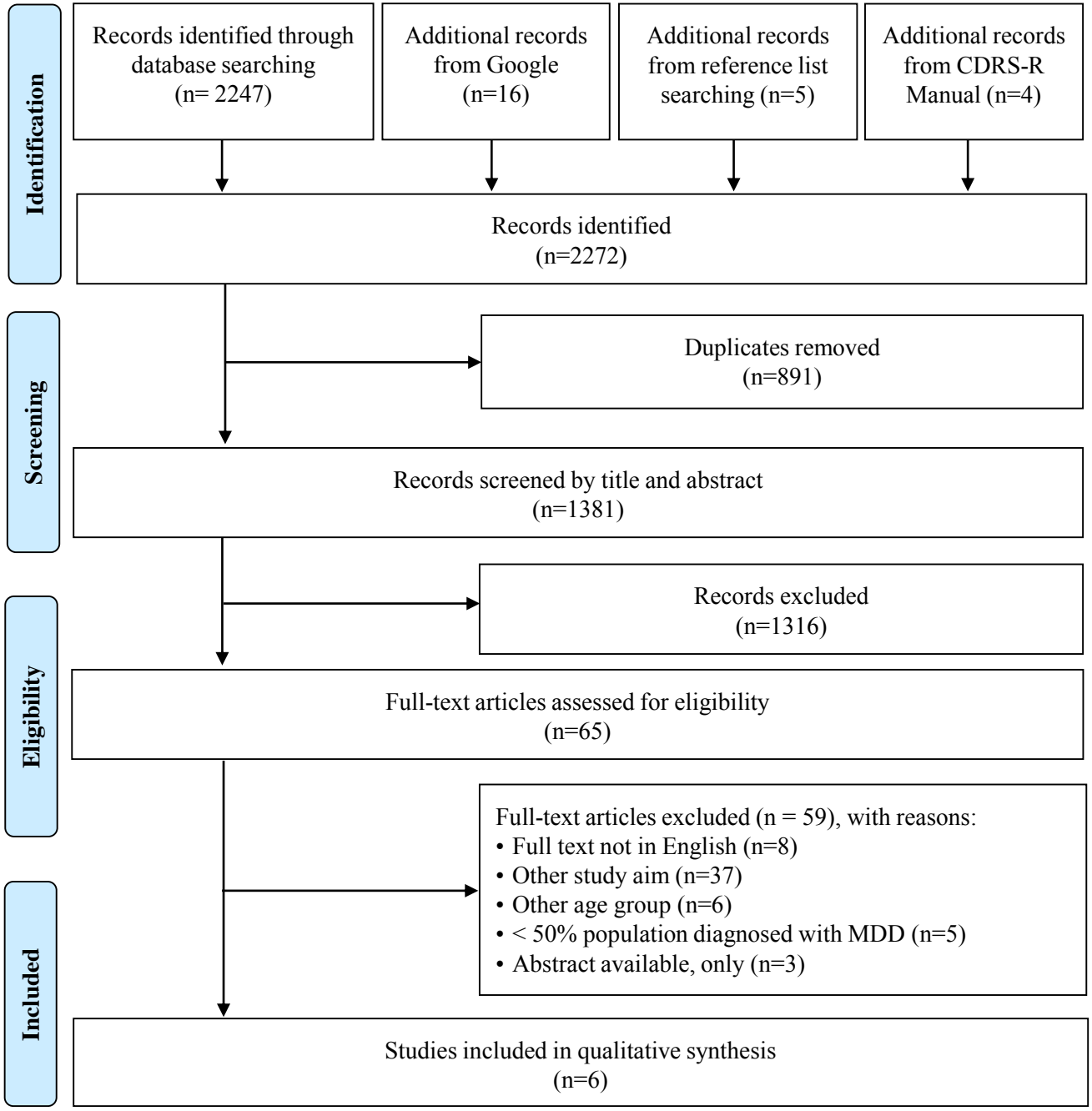
[h]Results of the CDRS-R correlations with the Reynolds scale.[29]

Figure Titles and Legends

Figure 1. Flow Diagram of Information Through the Different Phases of the Systematic Review

Note: Adapted from Moher et al[17] (https://doi.org/10.1371/journal.pmed.1000097; published under Creative Commons license CC BY: https://creativecommons.org/licenses/by/4.0/)**.** MDD = Major Depressive Disorder; CDRS-R = Children's Depression Rating Scale-Revised.

Figure 1


Identification
Records identified through database searching (n= 2247)
Additional records from Google (n=16)
Additional records from reference list searching (n=5)
Additional records from CDRS-R Manual (n=4)
Records identified (n=2272)
Screening
Duplicates removed (n=891)
Records screened by title and abstract (n=1381)
Eligibility
Records excluded (n=1316)
Full-text articles assessed for eligibility (n=65)
Full-text articles excluded (n = 59), with reasons:
• Full text not in English (n=8)
• Other study aim (n=37)
• Other age group (n=6)
• < 50% population diagnosed with MDD (n=5)
• Abstract available, only (n=3)
Included
Studies included in qualitative synthesis (n=6)

PRISMA 2009 Checklist

Table S1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Checklist.

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5-6
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	5
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	6
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	6-7/Table S2
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	7/Fig1
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	7-8
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	7-8
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	8-10
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	12-13/Table3
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	12-13

PRISMA 2009 Checklist

Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	8-10
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	12-13
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	14
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	14/Table 5
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Table 5
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	15-17
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	14-15/Table 4/Table 5
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	17-19
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	19
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	22
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	23
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Title page

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Table S2. Search Strategy used to Identify Empirical Studies that Evaluated the Measurement Properties of the Children's Depression Rating Scale-Revised in Adolescents with Major Depressive Disorder (MDD)

A. Strategy designed for the Ovid Medical Literature Analysis and Retrieval System Online (MEDLINE)® Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE® Daily 1946 to July 1, 2019.

CDRS-R	
Term Used	**Yield**
1. "CDRS-R".tw	129
2. "Children's Depression Rating Scale Revised".tw	170
3. "Children's Depressive Rating Scale Revised".tw	4
4. "Child Depression Rating Scale Revised".tw	14
5. "Child Depressive Rating Scale Revised".tw	0
6. "Child's Depression Rating Scale Revised".tw	1
7. "Child's Depressive Rating Scale Revised".tw	0
8. 1 or 2 or 3 or 4 or 5 or 6 or 7	213

B. Strategy designed for the Ovid PsycINFO® 1806 to July 1, 2019 database.

CDRS-R	
Term Used	**Yield**
1. "CDRS-R".tw	132
2. "Children's Depression Rating Scale Revised".tw	178
3. "Children's Depressive Rating Scale Revised".tw	5
"Child Depression Rating Scale Revised".tw	15
4. "Child Depressive Rating Scale Revised".tw	1
5. "Child's Depression Rating Scale Revised".tw	0
6. "Child's Depressive Rating Scale Revised".tw	0
7. 1 or 2 or 3 or 4 or 5 or 6 or 7	226

C. Strategy designed for the Ovid *Excerpta Medica database (Embase®) 1947 to July 1, 2019.*

CDRS-R	
Term Used	**Yield**
1. "CDRS-R".tw	262
2. "Children's Depression Rating Scale Revised".tw	249
3. "Children's Depressive Rating Scale Revised".tw	4
4. "Child Depression Rating Scale Revised".tw	31
5. "Child Depressive Rating Scale Revised".tw	0
6. "Child's Depression Rating Scale Revised".tw	1
7. "Child's Depressive Rating Scale Revised".tw	0
8. 1 or 2 or 3 or 4 or 5 or 6 or 7	360

D. Strategy designed for the Cochrane Central Register of Controlled Trials®, 2005 to July 1, 2019.

CDRS-R	
Term Used	**Yield**
1. "CDRS-R".tw	146
2. "Children's Depression Rating Scale Revised".tw	133
3. "Children's Depressive Rating Scale Revised".tw	1
4. "Child Depression Rating Scale Revised".tw	0
5. "Child Depressive Rating Scale Revised".tw	0
6. "Child's Depression Rating Scale Revised".tw	1
7. "Child's Depressive Rating Scale Revised".tw	0
8. 1 or 2 or 3 or 4 or 5 or 6 or 7	192

E. Strategy designed Cumulative Index to Nursing and Allied Health Literature®, 1961 to July 1, 2019.

CDRS-R	
Term Used	**Yield**
1. CDRS-R	148
2. Children's Depression Rating Scale Revised	127
3. Children's Depressive Rating Scale Revised	2
4. Child Depression Rating Scale Revised	127
5. Child Depressive Rating Scale Revised	2
6. Child's Depression Rating Scale Revised	127
7. Child's Depressive Rating Scale Revised	2
8. 1 or 2 or 3 or 4 or 5 or 6 or 7	183

Note: Search Strategy was for Title OR Abstract (i.e., TI or AB).

F. Strategy designed for Web of Science®, 1990 to July 1, 2019.

Term Used	**Yield**
1. CDRS R	205
2. Children's Depression Rating Scale Revised	208
3. Children's Depressive Rating Scale Revised	153
4. Child Depression Rating Scale Revised	438
5. Child Depressive Rating Scale Revised	246
6. Child's Depression Rating Scale Revised	307
7. Child's Depressive Rating Scale Revised	191
8. 1 or 2 or 3 or 4 or 5 or 6 or 7	547

Note: Search Strategy was for Title OR Abstract.

G. Strategy designed for Scopus®, 1960 to July 1, 2019.

Term Used	**Yield**
TITLE-ABS(cdrs AND r) OR TITLE-ABS(children's AND depression AND rating AND scale AND revised) OR TITLE-ABS(children's AND depressive AND rating AND scale AND revised) OR TITLE-ABS(child AND depression AND rating AND scale AND revised) OR TITLE-ABS(child AND depressive AND rating AND scale AND revised) OR TITLE-ABS(child's AND depression AND rating AND scale AND revised) OR TITLE-ABS(child's AND depressive AND rating AND scale AND revised)	526

H. Strategy designed for Google without date restrictions.

Term Used	**Yield**
Children's Depression Rating Scale Revised	16

Note: Only the first 4 pages of Google were searched.

Table S3. The Definitions of the Quality of Evidence Levels According to the Adapted GRADE Approach

Quality of Evidence	Definition
High	We are very confident that the true measurement property lies close to that of the estimate of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very Low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

Note: Adapted from Prinsen et al[18] (https://doi.org/10.1007/s11136-018-1798-3; published under Creative Commons license CC BY: https://creativecommons.org/licenses/by/4.0/). GRADE: Grades of Recommendation, Assessment, Development and Evaluation.

References

1. Prinsen C, Mokkink L, Bouter L, et al. COSMIN guideline for systematic reviews of Patient‐Reported Outcome Measures. *Qual Life Res*. 2018;27(5):1147-1157.