**Assessing the heterogeneity of Chinese scales for screening depression among children and adolescents**

Haoyuan Wang 1, Mengzhen Hu 1, Liuqing Tian 1, Weibiao Liu 1, Yuanyuan An1, Ying Li 2, Hu Chuan-Peng 1#

(1School of Psychology, Nanjing Normal University, Nanjing, China; 2

Department of Psychiatry, Beijing Children’s Hospital, Capital Medical University, Beijing, China)

**CrediT Author Statement**

**Haoyuan Wang：**Conceptualization, Data curation, Investigation, Writing - Original Draft. **Mengzhen Hu：**Data curation, Visualization, Investigation. **Liuqing Tian:** Investigation. **Weibiao Liu:** Investigation. **Yuanyuan An:** Writing- Reviewing and Editing. **Ying Li:** Investigation, Writing- Reviewing, and Editing. **Chuan-Peng Hu：**Conceptualization, Supervision**,** Investigation, Project administration, Writing-Reviewing and Editing.

Corresponding author: Hu Chuan-Peng, email: [hcp4715@hotmail.com](mailto:hcp4715@hotmail.com)

# 1. Introduction

[儿童青少年抑郁障碍的严重性（自杀率）]

Major depression, a highly prevalent mental disorder, imposes substantial personal, social, and economic burdens both in China (Fu & Zhang, 2023) and globally (Herrman et al., 2022; McGrath et al., 2023). Depression can occur as early as the age of 10 (Dattani, 2022), with a peak onset at 19.5 years (Solmi et al., 2022), which means adolescence a critical period for the onset of depression. Depression is accompanied by self-harm and suicide among children and adolescents (Zeynep Başgöze et al., 2021), resulting in over 800,000 deaths annually among individuals aged 15-29 (Amaltinga & Mbinta, 2020). Therefore, measuring, preventing, or alleviating depression is an urgent issue in China (Ma et al., 2023) and globally (UNICEF China, 2021).

[不同测量方法抑郁患病率不同]

While significant attention is directed to the effectiveness of interventions (Cuijpers et al., 2020), an equally fundamental issue is often overlooked: how to measure depression (Fried et al., 2022). The reported prevalence rates of depression differ significantly when the measurements used vary (Chen et al., 2022; Huang et al., 2022; Yu et al., 2022; Zhang et al., 2022). For example, employing a strict, clinical interview-based standard, Deng et al. (2023) estimated that the prevalence of major depression among Chinese children aged 6-16 years is approximately 2% to 3% (Deng et al., 2023; Li et al., 2022). In contrast, the estimated prevalence based on data from self-reported scales was much higher: 14.6% among elementary school students, 23.6% to 24.2% among middle school students, and 28.0% among high school students (Chen et al., 2022; Huang et al., 2022; Yu et al., 2022; Zhang et al., 2022).

[梳理先前研究发现的抑郁测量问题]

The overlooked issues in measuring depression hinder the diagnosis and treatment of this mental disorder (Fried et al., 2022). A prominent problem is the heterogeneity of symptoms that define depression. Fried (2017) analyzed the content of seven widely used depression scales (e.g., BDI-II, HRSD, CES-D, IDS, QIDS, MADRS, SDS), revealed a low overlap between scales in the symptoms, questioned the notion that different measurements of depression are measuring the same "depression". In a separate study, Fried et al. (2015) investigated the number of unique symptom profiles reported by 3,703 depressed outpatients and found that nearly half of the symptoms were endorsed by only one individual, highlighting the variability in diagnostic criteria in clinical settings. Similarly, Veal et al. (2024) identified 3,888 different measures of outcomes among 450 randomized controlled trials (RCTs) of unipolar and bipolar depression. All these findings suggest it is urgent to address the issue of measuring depression.

[抑郁测量问题的严重性，临床，可重复性和普遍性的挑战，影响政策等]

The importance of how depression is measured is self-evident. First, the scales or methods we use for measuring depression should reflect how we conceptualize and operationalize depression. However, most of the scales used in practice were developed decades ago and have not been updated with our recent understandings of depression, especially the non-medical factors (Eaton et al., 2023). Secondly, research can influence mental health policy development, adoption, and implementation by providing scientifically validated choices and data (Whiteford, 2001). Nonetheless, varying depression measurement methods can produce different outcomes, such as changes in detection rates, which may subsequently affect public policies and intervention strategies. Moreover, the heterogeneity of symptoms that define depression can challenge our ability to compare data across different studies or populations, ultimately affecting the reliability of depression-related research and the efficacy of interventions designed to address this pervasive mental health issue.

[尚未解决的三个问题]

Although recent studies have revealed the heterogeneity issue in measuring depression (Fried, 2017), three problems remain unaddressed. First, whether the measurement of depression among children and adolescents shares the aforementioned problem is unknown. Studies have shown that depression in children and adolescents is different from that of adults in multiple ways (Kułak-Bejda et al., 2022). Juvenile depression is characterized by a distinct set of symptoms, including mood disorders, behavioral disorders, anxiety, and self-destructive behavior. The duration of depression among juveniles is usually shorter. Also, the treatment strategies are different between children/adolescents and adults: medication is commonly used to treat depression in adults but is rarely used for treating children and adolescents (American Psychological Association, 2023). So far, whether the heterogeneity issue in measuring depression among children and adolescents exists is unknown.

Second, the situation outside Europe and the US is largely unknown. Previous psychological research has primarily focused on Western, educated, industrialized, rich, and democratic (WEIRD) countries (Chen et al., 2023; Muthukrishna et al., 2020). The same issue applies to psychometrics, with few studies focusing on measurements in non-English-speaking countries. For example, in Fried (2017) study, none of the seven depression scales included were developed in non-Western countries. In psychological research, WEIRD samples are among the least representative populations (Henrich et al., 2010). In the field of mental health and psychiatric research, this issue similarly affects the diagnosis and treatment of psychological disorders, including depression (Wilson, 2024). Including non-WEIRD samples would alleviate concerns about the validity, reliability, generalizability, and robustness of psychological research (Tindle, 2021). Having a systematic review of the scales used for measuring depression in China would provide one piece of such information.

Third, previous studies focused on widely used scales, but it is unknown how many scales were used in measuring depression and how frequently each was used. A recent meta-research revealed that researchers often avoid using existing models, concepts, and measurements (Anvari et al., 2024), which leads to an issue: most measurements were used only once or twice, and new measurements were created instead of revising the existing ones. This slows down cumulative science (Elson et al., 2023). However, it remains unclear whether this issue exists in the measurement of depression in China.

[总结本研究的工作]

To address the three issues above, we conducted a comprehensive content analysis of available Chinese depression scales. Our findings revealed that among the scales used for measuring depression in Chinese children and adolescents, the overlap between scales is low. We also found that a few scales were mostly used, while most other scales were seldom used. Meanwhile, our results provided an extensive list of symptoms used for screening depression among Chinese children and adolescents, laying a solid foundation for the further development of measurements. Additionally, these results offered insights into the potential psychopathology and social pathways of depression in Chinese society (and, by extension, East Asian societies) and will inspire future development of scales for measuring depression in clinical settings and public health policies.

# 2. Methods

We took three steps to extract symptoms from all scales that measure depressions among the Chinese student’s population (see Fig 1 for the flowchart). Firstly, we identified all scales that has been used for screening depression. Secondly, we identified unique symptoms of each scale. Thirdly, we compared the symptoms across all scales. The latter two steps followed Fried (2017) with minor modifications (see details below).

## 2.1 Identify and screen scales

We identified scales that measure depression from four recent meta-analyses which synthesized the prevalence of different mental health problems among four Chinese students populations: elementary school, middle school, high school, and college (Chen Y. et al., 2022; Huang et al., 2022; X. Yu et al., 2022; Zhang Y. et al., 2022). We extracted all papers included in these meta-analyses and selected scales that were used for screening depression. In total, we identified 34 scales from 441 articles from all articles included in these four meta-analyses.

We then screened versions of scales and identified the most valid version for later analysis (See supplementary material for details). If a scale has multiple versions, we choose one of these version based on the following criteria: (1) If the reference(s) of the scale cited in articles was in English, we search for the Chinese versions because all participants in the above mentioned articles are Chinese students; (2) If the reference(s) of the scale included both English and Chinese versions of the scale, we included only the Chinese version; (3) If there were multiple Chinese versions and if the one(s) published later in time mythologically improved the previous ones, we choose the later version; (4) If there were multiple Chinese versions and no obvious methodological advances were reported, we choose the one explicitly stated symptom names; (5) If all other condition were equal, we selected the most frequently cited one. For instance, CES-D was first translated by Wang et al. (1999), which was used by 14 of all 441 papers included in the four menta-analyses. However, Zhang et al. (2010) improved the translation of the 20th item, thus, we chose the revised version by (Zhang J. et al., 2010).

## 2.2 Identify symptoms within scales

In this phase, four trained coders focused on identifying items that assessed identical or similar symptoms within each scale. The procedure was as follows: First, the four coders independently identified and consolidated items within each scale. Then, they formed two pairs, with each pair reviewing their results and resolving any discrepancies within the pair. Subsequently, the results from two pairs of coders were cross-checked and any remaining inconsistencies were discussed and resolve with the corresponding author when necessary. The final consolidated version of the results underwent independent verification by a clinically certified psychiatrist (Y. L).

## 2.3 Compare symptoms across different scales

In this step, the same four coders compared symptoms across all scales. The procedure was the same as the identification of symptoms within scales: independent individual coding, discussion by pairs, cross-checking between pairs, discussion with the corresponding author, and verification by a clinically certificated doctor.

The coding processes were as follow. We first pooled symptoms from all scales together and identified unique symptoms. In this step, we retained both compound symptoms and specific symptoms, as in Fried (2017). Compound symptoms are symptoms that include a range of related symptoms, whereas specific symptoms are more concrete and describe specific patterns. For example, “appetite changes” is a compound symptom, it includes two specific symptoms: “appetite increased” and “appetite decreased”, and all three of them were treated as unique symptoms. We employed an approach that maximize the number of different symptoms. More specifically, if the items describe similar symptom using different words and that the words have significantly different meanings under the Chinese context, we treat them as belonging to the same compound symptom but are different specific symptoms. For instance, there are many different words for describe depressed mood in different scales, we used 'depressed moods' as the compound symptom but distinguish different specific symptoms such as: 'blue', 'low mood', 'sad', and 'anhedonia'. This approach is slightly different from Fried (2017), where he coded all these item as a specific symptom 'Sad moods'.

Then, we assigned score of all scales on all unique symptoms. More specifically, a scale was scored as “0” on a symptom if it does not have items that measures this symptom. For instance, the Children's Depression Inventory (CDI) has no item for 'Depressed mood', we assigned “0” for CDI on this symptom. If a scale has an item that directly measures a symptom, compound or specific, it was coded as 2 on that symptom. Note if a scale has an item measures a compound symptom, then, this scale not only had a score of 2 on that compound symptom but also has a score of 1 on all specific symptoms of this compound symptom. For example, CDI has an item directly measures the compound symptom “appetite change” and scored 2 on this compound symptom. Importantly, even CDI does not have items for 'appetite increased' and 'appetite decreased', it scored 1 on these two specific symptoms (see Fig 2 and Supplementary Materials for details). However, if the item measures a specific symptom under a compound symptom, this scale was coded “2” on that specific symptom but still “0” on the corresponding compound symptom.

[Insert Fig 2 later]

**Figure 2***. Number of articles that used each scale for screening depression.* Total number of articles were based on those included in four meta-analyses (XXX). Scales with \* were excluded from our analyses.

## 2.4 Data analyses

We reported the descriptive summary of scales as well as the symptoms within scale. We highlighted symptoms that are used in DSM-5 for diagnosis of depression. More specifically, there are 28 symptoms are overlapped with the DSM-5 symptoms for depression, where were derived from the nine symptoms in DSM-5 and all their specific symptoms, see Fried et al. (2015) for details.

We used Jaccard index to quantify the degree of overlap between different scales (Fried, 2017). The formal of Jaccard index is s/(u1 + u2 + s), where "s" represents the number of items shared by two scales, and "u1" and "u2" denote the number of items that are exclusively present in each of the two scales. Jaccard index ranges from 0 (no overlap among scales) to 1 (complete overlap). We interpreted Jaccard index as in Fried (2017): very weak 0.00–0.19, weak 0.20–0.39, moderate 0.40–0.59, strong 0.60–0.79, and very strong 0.80–1.0. Moreover, we explored the relationship between the mean Jaccard coefficient, adjusted scale length and the number of captured symptoms (i.e., number of symptoms included in the scale) by employing Spearman correlation (see Supplementary Material 5 for more details).

[Insert Fig later]

**Figure 1**

*Research flowchart*

# 3 Result

# 3.1 A summary of scales

Among all papers included in these four meta-analyses (citations), 441 reported depressions. We identified 34 unique scales in these articles. Among the 34 scales, six of which were developed by Chinese researchers (MSSMHS, CSSDS, ADI, CSSMHS, CEPS, Ji\_2007). The seven most frequently used scales in this study are SDS (XXX articles out of XXX), SCL-90 (XXX articles out of XXX), CES-D (XXX articles out of XXX), CDI (XXX articles out of XXX), DSRSC (XXX articles out of XXX), BDI-I (XXX articles out of XXX), and MSSMHS (XXX articles out of XXX). Among which, the CES-D and SDS are consistent with the scales selected by Fried (2017) and XXX were specifally for children and adolescent. Nineteen scales were each used in only one article; for details, see our Supplementary Material 1. See Fig 2 for the frequency of citations of these scales among all 441 empirical papers in the meta-analyses (Chen Y. et al., 2022; Huang et al., 2022; X. Yu et al., 2022; Zhang Y. et al., 2022).

To conduct content analysis, we searched for items for all 34 scales and the items of four scales, the Mini International Neuropsychiatric Interview for children and adolescents (Mini-KID), WHO-CIDI 3.0, Psychological Health Inventory (PHI), and the Symptom Checklist 45, were not findable. Items of another scale, Short Depression Scale, was not findable because of unidentifiable citations. Moreover, the Beck Depression Inventory (Zhang Yuxin Revised Edition) was merged with BDI because its symptoms measured are the same as BDI. The items used in 'Gu & Chen (2020) 'and 'Ji (2007)' were identical but in different languages, thus we also merged these two as one scale, referred them as 'Ji (2007)'. Finally, the boys’ and girls’ version of the Child Behavior Checklist (CBCL) were treated as one scale. In total, 27 scales were included in the current study.

# 3.2 Items and symptoms within scales

For the 27 scales included, there are 425 items in total. Among them, 73 items were merged into 31 symptoms. Also, there were two cases where one item measured two symptoms. The item from Ji (2007), "During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing your usual activities?" measures both 'sad' and 'Sense of hopelessness'. And the 8th item from PHQ-9, "Actions or speech slowed down to a noticeable extent, or conversely—feeling restless or agitated, being unable to sit still, more than usual", measures both 'Agitation' and 'Retardation'. Thus, a total of 385 symptoms were identified across all 27 scales (See supplementary materials for number of items and symptoms of each included scale).

The comparison of 385 symptoms across 27 scales resulted in unique 84 symptoms (see Figure 3). Among these, eight are compound symptoms, including *Depressive mood, Irritability*, *Self-abasement*, *Interest/pleasure loss*, *Somatization*, *Appetite changes*, *Somnipathy*, and *Reduced socialization*. Among all 27 scales, 19 of them did not include any idiosyncratic symptoms. For the other eight scales the rate of idiosyncratic symptom varied from 3.9% to 22.2%. Interestingly, all scales include symptoms that were not covered in DSM-5. The DSI has the highest proportion of DSM-5 symptoms for depression, 71.42% of the total nine DSM-5 depression symptoms. Please see the supplementary matierals for detailed information.

Among 27 scales, XXX scales (%XX) include items for children and adolescent. After merging items, there were 22 symptom that were specific for children and adolesnce. XXX scales (%XX) includ items that are cultural adopted, resulted in 6 distinct symptoms (See Supplementary Material 1 and the figure x for details).

[Insert Fig later]

**Figure 3***. Content Overlap Across* *Twenty-seven Depression Scales.* Each row represents a symptom, each column represents a scale.If a scale measures a symptom, then there is a dot or a circle on that row. The former represent compound symptoms and the latter represent specific symptoms. Symptoms in bold font are from DSM-5. ADI: Adolescent Depression Inventory, CDI: Children's Depression Inventory, HAMD: Hamilton Depression Rating Scale for Depression, DSI: Depression Status Inventory, SDS: Self- Rating Depression Scale, MFQ-C: Mood and Feelings Questionnaire, CBCL: Child Behavior Checklist, BDI-II：Beck Depression Inventory-II, DSRSC: Depression Self-rating Scale for Children, BDI-I: Beck Depression Inventory, KADS-11: Kutcher Adolescent Depression Scale, CES-D: The Center for Epidemiological Studies Depression Scale, PHQ-9: Patient Health Questionnaire-9 items, CSSDS: China Middle school students' depression scale, CES-D-C: Center for Epidemiologic Studies Depression Scale for Children, UPI: University Personality Inventory, SMFQ: Short Mood and Feelings Questionnaire, SCL-90: Symptom Checklist 90, CES-D-13: Short version of Center for Epidemiologic Studies Depression Scale, CCSMHS: Chinese College Student Mental Health Scale, DASS-21: The Depression Anxiety Stress Scale, BSRS-5: Brief Symptom Rating Scale, Sakuma\_2010: Sakuma et al.(2010) self-designed questionnaire, MSSMHS: Middle-school students Mental Health Scale, CEPS: China Education Panel Survey, HADS: Hospital Anxiety and Depression Scale, Ji\_2007: Comprehensive Survey Report on Health-Related/Risk Behaviors among Chinese Adolescents.The right side of the figure is *Number of symptoms that appear across combinations of scales.* There are 18 symptoms that appear only in one scale; these symptoms are referred to as idiosyncratic symptoms.

# 3.3 Symptoms across scales

Among the 84 symptoms, 18 (21.43%) were idiosyncratic symptoms and only appeared in one scale.

None of the 84 symptoms were present in all scales. The most frequently presented symptom, appeared in 22 out of 27 scales, was *Sense of hopelessness*. The second most measured symptom, 18 out of 27, was *Interest loss*. Note that *markedly diminished interest or pleasure*, a key symptom of diagnosis of major depression in DSM-5, is splitted into two specific symptoms: *interest loss* and *pleasure loss* in this study*.* We found *Pleasure loss* was observed less frequently than *Interest loss*, being measured in 9 out of 27 scales. Another frequently measured symptom is the compound symptom *Depressed mood*, which was directly measured in 5 scales. However, this compound symptom includes several specific symptoms: *blue* appeared in 10 scales, *low mood* in 15 scales, *sad* in 13 scales, and *anhedonia* in 16 scales*.* Combined as a cluster, depressed mood and related specific symptom presented in 25 out of 27 scale and are the most frequently measured symptom.

The degree of overlap between scales was calculated using the Jaccard coefficient. The mean overlap across all scales was 0.19, range from 0.09 to 0.25, indicating a very low level of similarity between these scales (see Figure 5). CES-D has the highest mean degree of overlap with other scale. The highest overlap, 0.75, appeared between two versions of CES-D: CES-D for adults and CES-D-C for children. The second highest overlap, 0.72, was between DSI and SDS. Many scales that have no overlap with each other. For example, there was no overlap between China Education Panel Survey with Middle-school students Mental Health Scale, PHQ-9, CSSDS, SMFQ, CSSMHS. Note that because Ji\_2005 has only one item that measures two symptoms, it has no overlap with PHQ-9, KADS-11.

We found a correlation between the mean Jaccard coefficient of each scale and the scale length, *r* = 0.54, 95% CI [0.14, 0.78]. Similarly, the correlation between mean overlap of scales and the number of captured symptoms is *r* = 0.70, 95% CI [0.39, 0.87]. These findings suggests that longer scales exhibit increased overlap with other scales, thus demonstrating enhanced representativeness.

[Insert Fig later]

**Figure 5.** *Overlap of item content of 27 depression scales.* ADI: Adolescent Depression Inventory, CDI: Children's Depression Inventory, HAMD: Hamilton Depression Rating Scale for Depression, DSI: Depression Status Inventory, SDS: Self- Rating Depression Scale, MFQ-C: Mood and Feelings Questionnaire, CBCL: Child Behavior Checklist, BDI-II：Beck Depression Inventory-II, DSRSC: Depression Self-rating Scale for Children, BDI-I: Beck Depression Inventory, KADS-11: Kutcher Adolescent Depression Scale, CES-D: The Center for Epidemiological Studies Depression Scale, PHQ-9: Patient Health Questionnaire-9 items, CSSDS: China Middle school students' depression scale, CES-D-C: Center for Epidemiologic Studies Depression Scale for Children, UPI: University Personality Inventory, SMFQ: Short Mood and Feelings Questionnaire, SCL-90: Symptom Checklist 90, CES-D-13: Short version of Center for Epidemiologic Studies Depression Scale, CCSMHS: Chinese College Student Mental Health Scale, DASS-21: The Depression Anxiety Stress Scale, BSRS-5: Brief Symptom Rating Scale, Sakuma\_2010: Sakuma et al.(2010) self-designed questionnaire, MSSMHS: Middle-school students Mental Health Scale, CEPS: China Education Panel Survey, HADS: Hospital Anxiety and Depression Scale, Ji\_2007: Comprehensive Survey Report on Health-Related/Risk Behaviors among Chinese Adolescents. Mean overlap is detailed in the supplementary materials.

# 4. Discussion

[第一段: 总结结果，并强调本研究的特色]

This study aimed at assessing the heterogeneity of Chinese scales that used for screening depression among children and adolescents. We identified 84 unique symptoms from 27 scales used in for screening depression from recent four meta-analyses (citations). We found a low overlap among scales: there was no single symptom appeared in all the scales and more than one fifth of the symptoms appreared in only one scale. Also, we found only a few scales included age-specific or cultural specific items. Our results confirmed the heterogeneity of depression scales among children and adolescents, suggesting that the issue of measuring depression exists among a different population other than WEIRD adults.

4.1. Key findings

[进一步介绍具体量表的结果，并与已有研究进行比较]

Similar to some previous content analysis studies, we examined the item content and overlap of depression rating scales. A key finding is the strong heterogeneity among depression scales. Unlike Fried (2017), our study found that the CES-D had the highest overlap with other scales. This discrepancy may be attributed to our inclusion of a broader range of scales, increased the possiblity that some symptoms of CES-D overlap with other scales' symptom. For example, in Fried (2017), the items "people were unfriendly" or "I felt that people disliked me" of CES-D were idiosyncratic symptoms, but we found these two items were also measured by XXX and XXX. However, one cannot conclude that the CES-D is the best scale. Previous research has found that the CES-D is one of the least representative scales (Shafer, 2006). The quality of assessment scales should primarily be discussed in the context of evaluating the actual issues experienced by individuals with depression (Fried, 2017).

[量表长度与重叠度的关系]

We also found a positive correlation between the length of the scale and the overlap. This is not surprising because longer scales usually included more items. It is important to note that a longer scale does not necessarily equate to a better one. Previous research has demonstrated that excessively long scales may negatively impact response quality, reliability, and response rates (Sharma, 2022). On the other hand, overly brief scales, such as single-item scale, are insufficient for capturing complex psychological constructs and fail to produce reliable scores (Allen et al., 2022). Indeed, we found the single-item measurement of depression (e.g., Ji\_2005) have the lowest average overlap with other scales (0.09).

[对捕捉到DSM-5症状的讨论]

Our results revealed that no scale includes all DSM-5 symptoms. This result is also consistent with previous finding (Fried et al., 2016). Because many symptoms in the DSM-5 are not empirically based, the extent to which a scale covers DSM-5 should not be used to judge its quality (see Fried, 2017).

[文化特异性项目的讨论]

Interestingly, the symptoms that specific to children and adolescents in Chinese society are also not strongly overlapped. The four scales that measure the depression of children and adolescents, MSSMHS, CSSDS, ADI, and CSSMHS, have XXX, XXX, XXX, and XXX age-specific items. Among these items, XXX appeared in all scales, suggesting that XXX is a common indictor of depression. There are also idiosyncratic symptoms. For example, only XXX measured XXX, sugestting XXX .

4.2 Heterogeneity of the depressive syndrome

Why is there heterogeneity in depression scales? Fried (2017) proposes two possible reasons: First, different scales reflect varying clinical perspectives on depression; Second, the original purposes when creating these scales were not the same. These two reasons also apply to our findings for the heterogeneity among scales that measures adults' depression. A third reason for heterogeneity may come from the time frames of different scales. For instance, the PHQ-9 asks respondents to recall feelings over the past two weeks, while the CES-D focuses on the past week. Different time frame may cause different symptoms. For example, XXX. Currently, there is no universally accepted time frame for evaluating depressive symptoms, but the heterogeneity caused by the time frame should be noted.

The inclusion of scales specifically designed to measure children and adolescents' depression under Chinese context may increase the heterogeneity in the current study. While some studies used these scale for measuring children and adolescents' depression, most other studies did not. Instead, they used scales that adopted from adults' scales. Among the seven most used scales (covered XX% of studies), only one of them were developed for children and adolescents. Thus, scales in the current study mixed those for adults (or general population) and those specifically for children and adolescent, increase the heterogeneity of the scales. Similarly, the current study includes both scales that were translated and adopted from English scales and those developed by Chinese researchers, which may also increase the heterogeneity. For example, symptoms such as Lack of patience and Mood swings are captured only by scales developed in China, reflecting unique insights of Chinese researchers into depression.

4.3 Implication

[大规模调查是问卷筛查，引导政策，需要格外注意]

In our empirical research based on children and adolescents, we confirmed the heterogeneity in measuring depression, i.e., Jingle-fallacy (REF). These results called for careful selection of a specific scale, especially when the stake is high. For instance, many large-scale surveys included item for depression among children and adolescents. The Report on national mental health development in China (REF) only used XXX; The Prevalence and influencing factors of anxiety and depression (REF) only used XXX; and the China Family Panel Study (REF) only used XXX (Fu & Zhang, 2023; Kong et al., 2023; Xie et al., 2014). The results of these large scale survey were usually intepreted as a representative data of the national depression level. For example, in study the relationship between XXX and depression, XXX used the the China Family Panel Study's data to represent the depression level of China. The General Office of the National Health Commission of China, in exploring a service plan for the prevention and treatment of depression, used only the PHQ-9 scale for screening (National Center for Mental Health and Mental Health, 2024).The findings from the document review and key informant interviews suggest that vital global initiatives and local epidemiological and political factors have collectively influenced the policy on depression prevention and treatment (Ma et al., 2023).

2、[关注儿童青少年的抑郁测量]

Our findings call for attention to the measurement of children and adolescents' depression. Although many methods for measuring depression in children and adolescents have been developed (Pegg et al., 2023), in practical research within China, the most widely used scales are the SDS, SCL-90, and CES-D. None of these three questionnaires were specifically developed to measure adolescent depression. The overlap between the SDS and the ADI, a scales specifically developed in China to measure adolescent depression, is only 0.22, lower than the average overlapping between scales. Without considering the culture- and age- specific factors in the most widely used scales for depression, the large-scale survey may produce biased results of the reality.

3 [中国抑郁测量的历史和问题]

Our findings also call for attention to the development of measurement that suit Chinese context. Currently, the measurement of depression primarily relies on translated scales, with very few scales developed locally. For instance, among the 27 scales included in this study, only 6 were developed by local researchers. Out of 441 papers, only 22 used scales that were developed by Chinese researchers (see Supplementary Material 1 for details). Since most psychological scales are in English (Gronier, 2023), non-English-speaking countries may face several challenges when measuring depression. First, translation errors: translation errors can occur during the process. For example, in Wang et al. (1999) version, the 20th item of the CES-D was translated as "I walk very slowly," which were corrected after ten years by Zhang et al. (2010). This leads to inconsistencies in the measured content, even when using the same CES-D scale in China. Second, lack of cross-cultural adaptation frameworks: Although recent studies have proposed frameworks for the necessary steps in cross-cultural adaptation and validation of scales (Gronier, 2023), many scales were translated before these frameworks were established. For instance, the SDS was translated as early as 1984！！！！(Wang & Chi, 1984) but without revision or update to check its cultural adaptation under the newer standard. Third, incomplete statistical analyses: Gronier (2023) summarized the statistical analyses commonly used in scale translation, including measuring internal consistency, factor analysis (exploratory and confirmatory), convergent validity, time constancy, and socio-demographic analyses. However, many Chinese translations were completed earlier without applying these statistical methods. For example, Wang et al. (1984) only reported the correlation between SDS and HAMD, and Zhang et al. (2010) did not conduct exploratory factor analysis in their study of CES-D. The challenges in China's depression measurement affect the replicability and universality of depression research significantly. These issues may also applied to other non-English-speaking countries.

4[抑郁测量方法的一致性]

From the meta-analysis data, among the 34 identified scales, 19 were used in only one article. The SDS, SCL-90, and CES-D scales were used far more frequently than other scales. SDS and CES-D are widely used globally for measuring depression (Fried, 2017), while the high usage of SCL-90 in China may be due to its familiarity among Chinese researchers. SCL-90 is the most commonly used mental health scale in Chinese universities (Yu et al., 2020), typically employed for screening first-year college studetns. Most questionnaires were used infrequently, and researchers often do not report the reasons for choosing a particular depression questionnaire from among many available ones, which is not recommended by recent best practices (Flake & Fried, 2018). Depression measurement should standardize its measurement criteria, but currently, there isn't such a standard. Scientific progress in a field largely accumulates gradually, requiring researchers to use the same concepts to describe their subjects, reach a consensus on the definition of these concepts, and employ comparable standards or measurement methods in their investigations (Anvari et al., 2024). Building a cumulative evidence base in psychology necessitates the standardization of our measures and protocols (Elson et al., 2023). This ensures that researchers can reliably compare results across different studies, thereby advancing scientific knowledge in a cohesive and systematic manner. More critically, although these scales are labeled as depression scales, they actually measure different content. This Jingle Fallacy is a common threat to the reproducibility and validity of psychological research (Weidman et al., 2017).

5 [抑郁测量的方法基础]

Finally, our results underscored the importance of methodological foundations of depression measurement. The existing scales were created primarily based on developers' clinical experience and personal viewpoints, rather than rigorously exploring the construct through item development, expert reviews, or focus groups (Fried et al., 2022). Also, many depression tools were developed without a clear and explicit theory about the nature of depression. Some underlying theories of depression, such as the notion that depression symptoms arise from a common cause, are inconsistent with actual research findings (Fried et al., 2022). While contemporary theories continue to evolve, the depression measurement scales has not kept up with these theoretical updates.

4.4 Strengths and limitations

[strength 1我们关注了儿童和青少年的抑郁测量。2，我们关注了非西方国家的抑郁测量。]

This study has several key strengthes. First, it focuses on the measurement of depression in children and adolescents, which differs considerably from the measurement of depression in adults, both in terms of scale selection and depression symptoms (Fried, 2017). We provide a comprehensive symptom list for depression in Chinese children and adolescents, facilitating future research in this demographic. Second, our study focuses on the measurement of depression in a non-Western country and revealed findings that may also apply to other non-Western countries. Unlike other content analysis studies (Kook et al., 2022; Wall & Lee, 2022), ours is the only one to incorporate scales developed in non-Western context. This enhances researchers' understanding of non-WEIRD regions and provides insights into the challenges of measuring depression in China and other non-English speaking countries.

[Limitations 1 内容分析很难客观，2我们纳入了更多的量表，可能会带来更高的异质性。]

Our study also has limitations. Despite being independently examined by different groups and ultimately validated by a psychiatrist to ensure the reliability of the results, content analysis is essentially subjective. We surmise that slight differences in results may occur if the study were to be redone by another research group. In the spirit of open science, we will make our coding public, allowing others to reanalyze this data. Another limitation of our study is that our meta-analytical dataset includes 27 scales. The inclusion of more scales would contribute to increased heterogeneity. Similarly, the fact that some scales have fewer items may also increase heterogeneity. If the analysis is only limited to those scales that are used more frequently, the final average overlap might be slightly higher.

# References

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single Item Measures in Psychological Science: A Call to Action. *European Journal of Psychological Assessment*, *38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

Amaltinga, A. P. M., & Mbinta, J. F. (2020). Factors associated with depression among young people globally: A narrative review. *International Journal Of Community Medicine And Public Health*, *7*(9), 3711. https://doi.org/10.18203/2394-6040.ijcmph20203949

American Psychological Association. (2023, January). *Clinical Practice Guideline for the Treatment of Depression in Children and Adolescents*. https://www.apa.org/depression-guideline/children-and-adolescents

Anvari, F., Alsalti, T., Oehler, L., Hussey, I., Elson, M., & Arslan, R. C. (2024). *A fragmented field: Construct and measure proliferation in psychology*. https://doi.org/10.31234/osf.io/b4muj

Chen Y., Zhang Y., & Yu G. (2022). Prevalence of mental health problems among college students in mainland China from 2010 to 2020: A meta-analysis. *Advances in Psychological Science*, *30*(5), 991–1004. https://doi.org/10.3724/SP.J.1042.2022.00991

Chen, Z., Hu, B., Liu, X., Becker, B., Eickhoff, S. B., Miao, K., Gu, X., Tang, Y., Dai, X., Li, C., Leonov, A., Xiao, Z., Feng, Z., Chen, J., & Chuan-Peng, H. (2023). Sampling inequalities affect generalization of neuroimaging-based diagnostic classifiers in psychiatry. *BMC Medicine*, *21*(1), 241. https://doi.org/10.1186/s12916-023-02941-4

Cuijpers, P., Stringaris, A., & Wolpert, M. (2020). Treatment outcomes for depression: Challenges and opportunities. *The Lancet Psychiatry*, *7*(11), 925–927. https://doi.org/10.1016/S2215-0366(20)30036-5

Dattani, S. (2022). At what age do people experience depression for the first time? *Our World in Data*. https://ourworldindata.org/depression-age-of-onset

Deng, H., Wen, F., Xu, H., Yang, H., Yan, J., Zheng, Y., Cui, Y., & Li, Y. (2023). Prevalence of affective disorders in Chinese school-attending children and adolescents aged 6–16 based on a national survey by MINI-Kid. *Journal of Affective Disorders*, *331*, 192–199. https://doi.org/10.1016/j.jad.2023.03.060

Eaton, N. R., Bringmann, L. F., Elmer, T., Fried, E. I., Forbes, M. K., Greene, A. L., Krueger, R. F., Kotov, R., McGorry, P. D., Mei, C., & Waszczuk, M. A. (2023). A review of approaches and models in psychopathology conceptualization research. *Nature Reviews Psychology*, *2*(10), 622–636. https://doi.org/10.1038/s44159-023-00218-4

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren’t toothbrushes. *Communications Psychology*, *1*(1), 25. https://doi.org/10.1038/s44271-023-00026-9

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are ‘good’ depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Fried, E. I., & Nesse, R. M. (2015a). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR\*D study. *Journal of Affective Disorders*, *172*, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

Fried, E. I., & Nesse, R. M. (2015b). Depression sum-scores don’t add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Fu, X., & Zhang, kan. (2023). *BLUE BOOK OF MENTAL HEALTH REPORT ON NATIONAL MENTAL HEALTH DEVELOPMENT IN CHINA (2021-2O22)* (1st ed.). Social Sciences Academic Press.

Gronier, G. (2023). Psychometric Analyses in the Transcultural Adaptation of Psychological Scales. In S. Misciagna (Ed.), *Psychometrics—New Insights in the Diagnosis of Mental Disorders*. IntechOpen. https://doi.org/10.5772/intechopen.105841

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Herrman, H., Patel, V., Kieling, C., Berk, M., Buchweitz, C., Cuijpers, P., Furukawa, T. A., Kessler, R. C., Kohrt, B. A., Maj, M., McGorry, P., Reynolds, C. F., Weissman, M. M., Chibanda, D., Dowrick, C., Howard, L. M., Hoven, C. W., Knapp, M., Mayberg, H. S., … Wolpert, M. (2022). Time for united action on depression: A Lancet–World Psychiatric Association Commission. *The Lancet*, *399*(10328), 957–1022. https://doi.org/10.1016/S0140-6736(21)02141-3

Huang X., Zhang Y., & Yu G. (2022). Prevalence of mental health problems among primary school students in Chinese mainland from 2010 to 2010:A meta-analysis. *Advances in Psychological Science*, *30*(5), 953–964. https://doi.org/10.3724/SP.J.1042.2022.00953

Kong, X., Wu, Y., Wang, X., Sun, Y., Chen, K., Li, Q., & Li, J. (2023). Analysis of the prevalence and influencing factors of anxiety and depression in the Chinese population: A cross-sectional survey. *Heliyon*, *9*(5), e15889. https://doi.org/10.1016/j.heliyon.2023.e15889

Kook, M., Clinger, J. W., Lee, E., Schneider, S. C., Storch, E. A., & Guzick, A. G. (2022). A Content Analysis of Self-report Child Anxiety Measures. *Child Psychiatry & Human Development*. https://doi.org/10.1007/s10578-022-01455-z

Kułak-Bejda, A., Bejda, G., & Waszkiewicz, N. (2022). Depression of Children and Adolescents. *Progress in Health Sciences*, *12*(2), 109–117. https://doi.org/10.5604/01.3001.0016.1754

Li, F., Cui, Y., Li, Y., Guo, L., Ke, X., Liu, J., Luo, X., Zheng, Y., & Leckman, J. F. (2022). Prevalence of mental disorders in school children and adolescents in China: Diagnostic data from detailed clinical assessments of 17,524 individuals. *Journal of Child Psychology and Psychiatry*, *63*(1), 34–46. https://doi.org/10.1111/jcpp.13445

Ma, J., Zhou, H., Fu, Q., & Lu, G. (2023). Facilitators and barriers in the development and implementation of depression prevention and treatment policies in China: A qualitative study. *BMC Public Health*, *23*(1), 276. https://doi.org/10.1186/s12889-023-15201-0

McGrath, J. J., Al-Hamzawi, A., Alonso, J., Altwaijri, Y., Andrade, L. H., Bromet, E. J., Bruffaerts, R., De Almeida, J. M. C., Chardoul, S., Chiu, W. T., Degenhardt, L., Demler, O. V., Ferry, F., Gureje, O., Haro, J. M., Karam, E. G., Karam, G., Khaled, S. M., Kovess-Masfety, V., … Zaslavsky, A. M. (2023). Age of onset and cumulative risk of mental disorders: A cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, *10*(9), 668–681. https://doi.org/10.1016/S2215-0366(23)00193-1

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

National Center for Mental Health and Mental Health. (2024). *Notice of the General Office of the National Health Commission on Exploring the Development of Special Services for the Prevention and Treatment of Depression and Senile Dementia (Health Office Disease Control Letter (2020) No. 726)*. https://ncmhc.org.cn/channel/newsinfo/6266

Pegg, S., Green, H. E., Kesselring, I., Burkhouse, K. L., & Kujawa, A. (2023). Depression Assessment. In J. L. Matson (Ed.), *Handbook of Clinical Child Psychology* (pp. 1027–1061). Springer International Publishing. https://doi.org/10.1007/978-3-031-24926-6\_48

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. https://doi.org/10.1002/jclp.20213

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar De Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: Large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*(1), 281–295. https://doi.org/10.1038/s41380-021-01161-7

Tindle, R. (2021). Improving the global reach of psychological research. *Discover Psychology*, *1*(1), 5. https://doi.org/10.1007/s44202-021-00004-4

UNICEF China. (2021). *ADOLESCENT MENTAL HEALTH A UNICEF PRIORITY FOR 2021-2025*. https://china.un.org/en/176907-adolescent-mental-health

Veal, C., Tomlinson, A., Cipriani, A., Bulteau, S., Henry, C., Müh, C., Touboul, S., De Waal, N., Levy-Soussan, H., Furukawa, T. A., Fried, E. I., Tran, V.-T., & Chevance, A. (2024). Heterogeneity of outcome measures in depression trials and the relevance of the content of outcome measures to patients: A systematic review. *The Lancet Psychiatry*, *11*(4), 285–294. https://doi.org/10.1016/S2215-0366(23)00438-8

Wall, A. D., & Lee, E. B. (2022). What do Anxiety Scales Really Measure? *Journal of Psychopathology and Behavioral Assessment*, *44*, 1148–1157.

Wang, X., Wang, X., & Ma, H. (1999). *Manual of the Mental Health Rating Scale*. Chinese Journal of Mental Health.

Wang, Z., & Chi, Y. (1984). Self-Rating Depression Scale(SDS). *Shanghai Psychiatry*, *2*, 71–72.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/emo0000226

Whiteford, H. (2001). Can Research Influence Mental Health Policy? *Australian & New Zealand Journal of Psychiatry*, *35*(4), 428–434. https://doi.org/10.1046/j.1440-1614.2001.00919.x

Wilson, S. (2024). Sociodemographic reporting and sample composition over 3 decades of psychopathology research: A systematic review and quantitative synthesis. *Journal of Psychopathology and Clinical Science*, *133*(1), 20–36. https://doi.org/10.1037/abn0000871

Xie, Y., Hu, J., & Zhang, C. (2014). The China Family Panel Studies: Design and Practice. *Chin J Sociol*, *34*, 1–32.

Yu, X., Zhang, Y., & Yu, G. (2022). Prevalence of mental health problems among senior high school students in mainland of China from 2010 to 2020: A meta-analysis. *Advances in Psychological Science*, *30*(5), 978. https://doi.org/10.3724/SP.J.1042.2022.00978

Yu, Y., Wan, C., Zhao, X., Huebner, E. S., Tan, J., Xu, C., & Zhang, J. (2020). Undergraduate students’ norms for the Chinese version of the symptom check-List-90-R (SCL-90-R). *BMC Public Health*, *20*(1), 1588. https://doi.org/10.1186/s12889-020-09689-z

Zeynep Başgöze, Andrea Wiglesworth, Katherine A. Carosella, Bonnie Klimes-Dougan, & Kathryn R. Cullen. (2021). Depression, Non-Suicidal Self-Injury, and Suicidality in Adolescents: Common and Distinct Precursors, Correlates, and Outcomes. *Journal of Psychiatry and Brain Science*. https://doi.org/10.20900/jpbs.20210018

Zhang J., Wu Z., Fang G., Li J., Han buxin, & Chen Z. (2010). Development of the Chinese age norms of CES-D in urban area. *Chinese Mental Health Journal*, *24*(2), 139–143.

Zhang Y., Jin J., & Yu G. (2022). Prevalence of mental health problems among junior high school students in Chinese mainland from 2010 to 2020: A meta-analysis. *Advances in Psychological Science*, *30*(5), 965–977. https://doi.org/10.3724/SP.J.1042.2022.00965