**Assessing the heterogeneity of 27 Chinese scales for screening depression among children, adolescents, and young adults**

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（患病率、社会成本等）]

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). Adolescence is a critical period for the onset of depression: depression can occur as early as the age of 10 (Dattani, 2022), with a peak onset at 19.5 years (Solmi et al., 2022). Moreover, depression during the adolescence also accompanied with self-harm and suicide (Zeynep Başgöze et al., 2021), resulting in over 800,000 deaths annually among individuals aged 15-29 (Amaltinga & Mbinta, 2020). The prevention and alleviation of depression are urgent issues in China (Ma et al., 2023) and globally (UNICEF China, 2021).

While the effectiveness of interventions has garnered significant attention (Cuijpers et al., 2020), a more fundamental issue is often overlooked: how to *measure* the depression disorder (Fried et al., 2022). The reported prevalence rates of depression differed signficantly when the measurements they used varied (Chen et al., 2022; Huang et al., 2022; Yu et al., 2022; Zhang Y. et al., 2022). For example, employing a strict, clinical interview-based standard, Deng et al. (2023) found 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, studies that used self-reported scales reported much higher prevalence rates: 14.6% among elementary school students, 23.6% to 24.2% among middle school students, and 28.0% among high school students (Chen, Zhang, and Yu., 2022; Huang, Zhang, and Yu 2022; Yu, Zhang, and Yu 2022; Zhang, Jin, and Yu 2022).

These findings underscore the importance of how depression is conceptualized, operationalized, and measured. The lack of clarity in measuring depression hinders the diagnosis and treatment of this mental disorder (Fried et al., 2022). Indeed, Fried (2017) analyzed the content of seven widely used depression scales, revealed a low overlap between scale in the symptoms, suggesting that different measurements of depression are measuring different things. In a separate study, Fried et al. (2015) investigated the number of unique symptom profiles reported by 3703 depressed outpatients and found that nearly half of the symptom were endorsed by only one individual, further highlighting the variability in diagnostic criteria in clinicals settings. Similarly, Veal et al. (2024) found that among 450 randomized controlled trials (RCTs) of unipolar and bipolar depression and identified 3888 different measures of outcomes. All these findings suggest it’s urgent to address the issue of measuring depression.

However, all previous studies focused on measuring depression among adults, few attentions have been paid to the measurement of depression among children and adolescent. Previous studies have shown that depression in children and adolescents are different from that of adults (Kułak-Bejda et al., 2022). To understand how depression is measured among children and adolescent, we conducted a comprehensive content analysis of 27 Chinese depression scales to assess their heterogeneity. Our findings revealed a low overlapping among these scales and call for attention to the measurement of depression of children and adolescent. Our results also provided a comprehensive list of symptoms used for screening depression among children and adolescent, laid a solid ground for further development of measurements. 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 et al., 2022; Huang et al., 2022; 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 438 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 39 of all 438 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), 438 reported depressions. We identified 34 unique scales in these articles. Among all these scales, 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. The items of the other two scales, the Beck Depression Inventory (Zhang Yuxin Revised Edition) and Short Depression Scale, were not available either because of unidentifiable citations. These six scales were excluded from further analyses. The items used in 'Gu & Chen (2020) 'and 'Ji (2007)' were identical but in different languages, thus we regarded these two studies used the same scale referred them as 'Ji (2007)'. Also, 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. See Fig 2 for the frequency of citations of these scales among all 438 empirical papers in the meta-analyses (Chen et al., 2022; Huang et al., 2022; Yu et al., 2022; Zhang Y. et al., 2022). The seven most frequently used scales in this study are SDS, SCL-90, CES-D, CDI, DSRSC, BDI-I, and MSSMHS, among which, the CES-D and SDS are consistent with the scales selected by Fried (2017).

# 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.

[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 appreared in all the scales and more than one fifth of the symptoms appreared in only one scale. Our results revealed greater heterogeneity than previous findings on the depression scales for adults (Fried (2017), suggesting that depression among children and adolescents may have a varied pathological pathways and different depression scales are not interchangeable.

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]

1、与先前的一些内容分析研究类似，我们的研究仍然发现了抑郁量表的异质性很强。与Fried (2017)不同的地方是，本研究中CES-D跟其他量表重叠度是最高的。这是因为我们纳入了更多的量表，其中包括了CES-D的简版和儿童版。尽管先前研究发现CES-D是不具备代表性的量表(citations)，但是元分析数据集中只有CES-D量表会同时纳入了简版、儿童版等版本。

2、我们还发现，较长的量表往往与其他量表有更大的重叠，从而呈现出更有代表性的内容(Fried, 2017)。例如，单项测量(即Ji\_2005)与其他量表的平均重叠度最低(0.09)。先前的研究表明，较长的量表可能会对反应的质量、可靠性和反应率产生负面影响(Sharma, 2022)。然而，过于简短的量表可能导致大量信息的丢失。

3、相比于之前的研究，我们在量表选择上范围更大，并且是基于实际中国青少年研究量表的使用情况。我们发现有两个量表是单项题目，尽管单项题目可以节省时间，和降低数据处理成本，但是单项题目在测量情绪时无法计算rest reliable, 并且复杂的心理结构单项题目也不能够进行充分的捕捉(Allen et al., 2022)。本研究在抑郁的角度支持了上述的这种说法。

4、不能认为CES-D就是好的量表，因为什么。

4.2 [为什么会有异质性]

1、不同问卷反映了临床对抑郁症的不同看法。

2、不同问卷的开发目的不一样。

3、由于本研究是基于儿童青少年的真实数据纳入的量表，因此我们会纳入一些专门研究儿童青少年的量表，这会造成异质性的增加。例如：CDI、CSSDS、ADI、CCSMHS这些专门测量儿童、青少年的问卷都会包含了与学习和学校相关的症状，这些症状是其他问卷不会包括在内的。

4、在评估和概念化青少年抑郁症时，文化背景、种族和民族也很重要(Pegg et al., 2023)。ADI and CCSMHS这两个中国研究者编制的量表有一些独特症状，可能反应了中国研究者对抑郁的一些独到理解。

Implication

1、[大规模调查是问卷筛查，引导政策，需要格外注意]

我们基于儿童青少年的实际研究，发现了不同的量表测量到了抑郁的不同方面，因此为研究选择特定量表可能存在严重偏差结果的风险。而目前诸多儿童青少年的大规模调查仍在使用单独量表测量(citations)，考虑到大规模调查能够影响到政策(Ma et al., 2023), 这一课题值得更多的关注。

2、[关注儿童青少年的抑郁测量]

评估青少年抑郁症需要选择评估方法，确定信息提供者，决定如何评分和概念化抑郁症，以及考虑文化和背景因素(Pegg et al., 2023)。尽管目前已经开发出了很多儿童青少年抑郁测量方法(Pegg et al., 2023)，然而在中国实际研究中使用最多的问卷是前三名则是SDS、SCL-90、CES-D，这三个问卷都并不是专门开发用来测量青少年的量表。SDS和中国编制的专门测量青少年抑郁的ADI的重叠率只有0.22。我们必须考虑到这样一种可能性，关于中国青少年抑郁的研究结果一直存在偏差，因为最常用的量表并不是专门针对青少年而编制的，对于青少年抑郁的症状测量可能并不够全面。

3 [抑郁测量的方法基础]

回顾抑郁测量的方法基础，会发现抑郁测量有以下两个问题。

1、Today, the development and validation of psychometric instruments is a thorough process that occurs in three phases. In phase one, the substance of the construct is explored (for example, clarifying its nature, breadth and depth). In phase two, the structure of the instrument is investigated (for example, using item analyses and factor analysis). Finally, in phase three, the relation between the instrument and external constructs is researched (for example, by testing its ability to discriminate between groups known to be distinct). 抑郁量表的开发通常忽视了第一个阶段，开发者通常根据他们的临床经验和个人观点开发抑郁工具，或者是通过统计程序完全避开理论考虑, 而不是通过项目开发、专家评论或焦点小组对要测量的结构进行严格的探索(Fried et al., 2022)。

2、大多数抑郁工具是在没有关于抑郁本质的清晰明确的理论的情况下开发的，抑郁症的一些潜在理论，例如depression symptoms arise from a common cause，与实际研究的结果不一致(Fried et al., 2022)。

4 [中国抑郁测量的历史和问题]

中国抑郁测量主要还是依靠翻译国外量表，本土开发特别少，例如：本研究纳入的共27个量表，只有6个量表是中国人自己编制的。438篇文献中，只有23篇使用的是中国人自己编制的量表（详见补充材料1）。

这会导致中国抑郁测量上出现三个问题，会抑郁症研究的可复制性和普遍性构成了重大挑战。

1. 直接翻译量表同样只有第二阶段和第三阶段，忽视了第一阶段。
2. 直接翻译量表反应了中国量表开发同样没有一个明确的抑郁症理论。

2，翻译过来的量表翻译过程中可能出现一些翻译错误，例如，在汪向东（1999）年版本就将CES-D第20题翻译成了“我走路很慢”，直到2010年，章婕（2010）才指出其是不准确的翻译。这就造成了在中国即使都用的是CES-D量表，所测到的内容都不完全相同。

5[问卷开发及使用的建议]

1. 开发

1 开发量表不能与理论分开。 2量表应该迭代开发(Fried et al., 2022)。

2、使用

Considering that scales may measure different intrinsic aspects of depression, it implies that users need to identify the structure they wish to examine before selecting a scale, and should not interchangeably use these measures or interpret them (Kook et al., 2022). When determining which scale to choose, consideration needs to be given to the evaluation objectives, psychometric properties, treatment settings and cultural considerations, administration and format, as well as performance characteristics such as specificity and sensitivity (Wall & Lee, 2022). For Chinese researchers, even after deciding to use a specific scale, careful attention should be paid to the selection of translated versions. We recommend researching whether there is a more recent version that improves upon earlier editions. At present, there isn't any specific research evaluating which scales are suitable for depression. Given this situation, using multiple scales may enhance the robustness of the study (Fried, 2017).

# 4 讨论

[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 appreared in all the scales and more than one fifth of the symptoms appreared in only one scale. Our results revealed greater heterogeneity than previous findings on the depression scales for adults (Fried (2017), suggesting that depression among children and adolescents may have a varied pathological pathways and different depression scales are not interchangeable.

[第二段: 进一步介绍具体量表的结果，并与已有研究进行比较]

Our results revealed that CES-D has the highest mean overlap with other scales (0.25). This result differed from Fried (2017)’s finding that CES-D had the lowest mean overlap with other six scales (0.27). Also, we found that CES-D had no idiosyncratic symptoms, while in Fried (2017), CES-D displayed 33% of idiosyncratic symptoms. The primary reason for these differences lies in the fact that we included more scales.

尽管如此，依旧不能认为CES-D可以和其他量表进行互换使用，CES-D仍然与13个量表的重叠度属于very weak等级。而CES-D和其简版作为中国境内大规模调查使用的量表（CEPS、蓝皮书），我们的发现可以帮助大家更好的理解那些大规模调查的结果。

We also found that longer scales tend to have greater overlap with other scales, thus presenting more representative content (Fried, 2017). For example, the single item measurement (i.e., Ji\_2005) had the lowest mean overlap (0.09) with other scales. Despite previous research suggesting that longer scales may negatively affect the quality of responses, their reliability, and response rate (Sharma, 2022). However, excessively brief scales could result in a loss of substantial information. 这个结果提示了我们在选则量表的时候，如果对所测症状数量有一定要求的话，可以在其他条件同样的情况下去选择那些题目数更多的量表。

[对另一个关键结果的讨论，应该是关注儿童青少年的抑郁测量，DSM-5是否有覆盖其实没有那么重要，因为它本身就是有问题的]

The scales exhibit significant variations based on the categorization of symptoms according to the DSM-5. None of the scales captured all individual diagnostic criteria of the depression disorders in the DSM-5. Interestingly, the PHQ-9 was crafted based on the symptoms of DSM-5 (Kroenke et al., 2001). Yet, in our findings, it only captured 64.29% of all the DSM-5 symptoms, ranking it second highest. This is due to the Chinese version of the PHQ-9 scale containing significantly less information compared to the symptoms of DSM-5. For instance, the seventh symptom in DSM-5 is "feelings of worthlessness or inappropriate guilt," yet the corresponding scale item in the Chinese version of PHQ-9 is "feeling terrible about oneself, or feeling like a failure, or a disappointment to oneself or one's family," which does not carry the concept of guilt in the Chinese context. 使用中文PHQ-9想反应全部DSM-5症状的学者需注意这一点，我们的补充材料提供了详细的量表对DSM-5症状的反应。我们研究的一个主要贡献就是对量表所测的症状提供了一个参考，有需要的研究者可以参考进行选择。

[第三段: 分析异质性的原因，以及重新阐述问题的严重性]

The heterogeneity of depression scales may arise due to varied clinical perspectives on depression and the fact that scales are developed for different purposes (Fried, 2017). In addition to this, it may also be influenced by certain cultural factors. In China, school education has a significant negative impact on the mental health of teenagers. In our findings, the depression scales MSSMHS, CSSDS, ADI and CCSMHS designed by Chinese researchers all include symptoms related to learning and school. The ADI and CCSMHS also include some idiosyncratic symptoms, which may reflect the unique insights of Chinese researchers into depression.

This heterogeneity in content can pose problems in certain situations, as these scales are often perceived as interchangeable (e.g., in meta-analyses). However, the reality is that different scales assess various aspects of depression, hence choosing a specific scale for a study could carry a significant risk of producing biased results.

我们的研究的一个现实意义就是根据实际研究的数据，发现了中文量表存在更严重的异质性问题。

Therefore, it may not be prudent to combine research findings from different scales (for instance in meta-analyses), particularly those that have little overlap. In light of the fact that the results of large-scale surveys can even influence the formulation of policies (Ma et al., 2023), this problems becomes particularly crucial.

[第四段：如何解读结果，call for better standards]

We cannot simply assume that a low overlap equates to a poor-quality scale, or conversely, that a high overlap equates to a good quality scale. Low overlap merely implies that the results obtained from this type of scale are less likely to be extrapolated to other scales. Considering that scales may measure different intrinsic aspects of depression, it implies that users need to identify the structure they wish to examine before selecting a scale, and should not interchangeably use these measures or interpret them (Kook et al., 2022). When determining which scale to choose, consideration needs to be given to the evaluation objectives, psychometric properties, treatment settings and cultural considerations, administration and format, as well as performance characteristics such as specificity and sensitivity (Wall & Lee, 2022).

我们的研究的一个意义就是梳理了中国实际研究中使用各个量表的频率，在量表选择时，从使用频率的这个角度能够给中国研究者一些参考。我们发现了即使中国自编的量表会有中国特色的症状，但实际研究中中文本土自编量表使用并不算多。中国大多数研究使用的还是翻译过来的量表，使用翻译过来的量表就会面临一个不同翻译版本的问题，因此对中国研究者而言，应该格外注意这一点。

For Chinese researchers, even after deciding to use a specific scale, careful attention should be paid to the selection of translated versions. We recommend researching whether there is a more recent version that improves upon earlier editions. At present, there isn't any specific research evaluating which scales are suitable for depression. Given this situation, using multiple scales may enhance the robustness of the study (Fried, 2017).

[第五段: 展望未来]

Limitations and Future Directions

Our study has some 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 (Fried, 2017). 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.

Based on our findings, we urge researchers and clinicians to be thoughtful in using and promoting the results generated from different depression scales, because what they measure might not be consistent. This study can serve as a reference for researchers who are interested in a particular symptom of depression. In future research, if we could evaluate the quality of these scales, it would provide great convenience for the practical application of Chinese researchers. Lastly, it is imperative for Chinese researchers to focus on the development of depression scales that are localized. A majority of the scales referred to in this study are translations from foreign sources with only a minimal contribution from Chinese researchers. It is only when Chinese researchers create these resources that there will be an increased likelihood of incorporating and measuring those issues that reflect specific Chinese characteristics.

# 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

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

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

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., 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.

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

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

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

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

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

Sharma, H. (2022). How short or long should be a questionnaire for any research? Researchers dilemma in deciding the appropriate questionnaire length. *Saudi Journal of Anaesthesia*, *16*(1), 65. https://doi.org/10.4103/sja.sja\_163\_21

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

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.

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

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