AN ADAPTATION AND VALIDATION STUDY OF THE HUBLEY DEPRESSION SCALE FOR OLDER ADULTS (HDS-OA) IN THE GENERAL ADULT POPULATION IN CHINA

by

XUYAN TANG

B.Sc., Sun Yat-Sen University, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

An adaptation and validation study of the Hubley Depression Scale for Older Adults (HDS-OA) in the general adult population in China

submitted by Xuyan Tang in partial fulfillment of the requirements for

the degree of Master of Arts

in Measurement, Evaluation and Research Methodology

Examining Committee:

Dr. Anita Hubley, Measurement, Evaluation and Research Methodology
Supervisor

Dr. Yan Liu, Measurement, Evaluation and Research Methodology
Supervisory Committee Member

Dr. Bruno Zumbo, Measurement, Evaluation and Research Methodology
Supervisory Committee Member

Abstract

Depression tends to be under-identified in China, as indicated by a much lower prevalence rate, despite similar diagnostic criteria used in China and Western cultures. Commonly used depression screens (e.g., Chinese versions of the Beck Depression Inventory-II or Zung Self-rating Depression Scale) tend to be overly long, costly, dated, or lack consistently strong psychometric evidence in Chinese samples. Accordingly, use of these measures leave many depression cases undetected and a more effective screen is thus needed for the Chinese population. The Hubley Depression Scale for Older Adults (HDS-OA) is ideally suited. It was developed based on the latest diagnostic criteria, is relatively short (i.e., 16 items), and has exhibited satisfactory psychometric properties in both depressed and non-depressed groups. In this thesis, two studies were conducted. In Study 1, I adapted the English version of the HDS-OA into Chinese, through backward and forward translation as well as use of pilot testing. In Study 2, I evaluated the psychometric properties of the Chinese version of the HDS-OA with a nonclinical sample from the general population in China. The data demonstrated a unidimensional factor structure, good internal consistency reliability, and strong convergent and discriminant evidence for validity. The purpose of this research is to obtain a better screen of depression that might increase the accuracy of diagnosis and allow for more timely intervention for Chinese people with depression. As a consequence, depression relapse and premature death caused by suicide may be prevented and a higher level of well-being in the general population may be obtained.

Lay Summary

Existing depression measures in China tend to be overly long, costly or dated. In addition, evaluations of these measures raise questions about their quality and most do not seem to screen for depression effectively in Chinese samples. This suggests that a new, effective instrument is needed for the Chinese population. In Study 1, I translated the Hubley Depression Scale for Older Adults (HDS-OA) into Chinese following the latest guidelines in the field and had four bilingual experts and 26 Chinese-speaking adults evaluate the quality of the translation, which appeared to be satisfactory. In Study 2, I examined the quality of the Chinese HDS-OA based on results from 364 adults in China. Results indicated satisfactory quality and responses to each item could be summed to produce a total score interpreted as a measure of depressive symptomatology in Chinese-speaking men and women.

Preface

This thesis is an original work by Xuyan Tang. All work contained within was approved by the University of British Columbia's Behavioural Research Ethics Board (Certificate No. H19-03135 for Study 1 and Certificate No. H20-01105 for Study 2), under the project title "A Translation and Validation Study of the Hubley Depression Scale (HDS-OA) in the General Adult Population in China". None of the text of the thesis is taken directly from previously published or collaborative articles.

The research idea and design were formulated by me and my supervisor, Dr. Anita Hubley. The primary measure used in this study, which was the Hubley Depression Scale for Older Adults (HDS-OA), was developed by Dr. Anita Hubley. Dr. Owen Lo, Dr. Michelle Chen, Sirui Wu, Man Niu, Michelle Zhang, Sophie Ma Zhu, Cynthia Hsu, Xin Gao, Candice Yu, Clarence Chan, and You Yi provided significant assistance in the forward or backward translation of the HDS-OA. The other measures used were selected by me and Dr. Anita Hubley. I was primarily responsible for survey set-up, data collection, statistical analysis, as well as the majority of manuscript composition. Dr. Anita Hubley contributed heavily to participant recruitment, data analysis and manuscript revisions and edits.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the following people, without whom I would not have been able to complete my thesis work. I would like to first express my deepest appreciation to my supervisor, Dr. Anita Hubley, for being a supportive, strong guiding force throughout my graduate studies. Her deep insights and valuable feedback helped me in every step of my research. I also thank her for appreciating my research strengths, which motivated me to produce a level of work that I never expected.

Besides my supervisor, I would like to thank the rest of my thesis committee, Dr. Bruno Zumbo and Dr. Yan Liu, who provided me with precious suggestions and comments that inspired this thesis.

I would also like to express my sincere gratitude to my family and friends. My biggest thanks to my parents, Shengmin and Jianhua, for their constant encouragement and unconditional love. A heartfelt thanks to my wonderful friend Michelle, for being a source of emotional support and strength, especially in my difficult times over the last three years. Lastly, I wish to thank all my friends who ever contributed in the preparation and completion of this thesis.

Thank you so much everyone!

Chapter 1: Introduction

Depression is derived from the Latin verb *deprimere*, which means "to press down" ("History of depression", 2019). It is a complex emotional experience, accompanied by a variety of emotions, including, but not limited to, sadness, anger, guilt and, shame; as a consequence, it lasts longer than any single negative emotion and causes more suffering (Meng, 2005). Depression not only leads to impairments of cognitive functioning, such as diminished ability for visuo-spatial processing, deficits in executive function as well as episodic memory impairments, but also results in impairments of social functioning, such as difficulties in understanding social emotions, dysfunctions in interpersonal interactions and decreased academic or work performance (Kupferberg, Bicks, & Hasler, 2016; Lam, Kennedy, McIntyre, & Khullar, 2014). Due to its detrimental effects on physical and mental health, depression has become the single largest cause of disability worldwide (7.5%) and a major contributor to the overall global burden of disease (World Health Organization, 2018).

According to recent World Health Organization (WHO) estimates, depression is a common mental disorder that affects around 322 million people of all ages across the world, among which more than 17% (around 54 million) are Chinese ("WHO China Office Fact Sheet-Depression", 2017). Despite the fact that a large amount of people in China are suffering from depression, China has a lower prevalence of depression in comparison with many countries in the world. A cross-cultural study conducted by Guerra et al. (2016) measured the prevalence rates of late life sub-syndromal depression in China and eight other countries with the same level of income (i.e., Cuba, Dominican Republic, Puerto Rico, Mexico, Venezuela, Peru, India and Nigeria). By dividing the catchment areas into urban sites and rural sites, researchers found that China had the lowest prevalence rate among these low- and middle-income countries, regardless

of the sampling locations (Urban sites: China: 2.5% vs. other countries: ranging from 15.5% to 38.6%, Rural sites: China: 1.0% vs. other countries: ranging from 7.8% to 25.3%) (Guerra et al., 2016).

Similar results are found when the prevalence rates of major depressive episodes (MDE) were compared between China and high-income countries. As shown in a cross-national study that administered the WHO Composite International Diagnostic Interview (CIDI) for assessment of MDE, the lifetime prevalence estimate of MDE in China was 6.5% (Kessler & Bromet, 2013). By contrast, the lifetime prevalence rates in New Zealand, Netherlands, United States and France were found to be about three times higher at 17.8%, 17.9%, 19.2% and 21.0% respectively.

Generally, China always has a lower depression prevalence compared with other countries, regardless of the income level. An explanation for this phenomenon is that people with depression in China are often mistakenly diagnosed with neurasthenia, which is determined by predominantly somatic symptoms such as physical fatigue, tension headache, and sleep disturbance (Lee, Kim, & Cho, 2017). Since Chinese patients are inclined to report bodily sensation, the diagnosis of neurasthenia is more popular and prevalent in epidemiological research and clinical practice than depression (Lee, 1999). The ratio between Chinese psychiatric patients who were diagnosed with depression and those identified as having neurasthenia was 1:30 (Kleinman, 1986). However, when reassessing 100 neurasthenia patients in China, it appeared that 87% of them should have been diagnosed with some kind of depressive disorder (Kleinman, 1982). This explains why depression seems to be much less common than neurasthenia in China, unlike many Western countries (He, 2013; Zhou, 2012).

Though the prevalence rates of depression in Western countries were about three times higher than that in China (Kessler & Bromet, 2013), statistics from the WHO indicates that the

suicide rate in China (8.0 per 100, 000 population) is quite close to those in Western countries (Netherlands: 9.6, New Zealand: 11.6, France: 12.1, United States: 13.7, per 100, 000 population) ("Suicide rate estimates", 2018). That is to say, China has a relatively low prevalence rate of depression but comparable suicide rate. Given that depression is reported to be a main risk factor for suicide attempts and committed suicide in China, it can be inferred that many suicide cases have failed to be diagnosed with depression (Liu, Contreras, Muñoz, & Leykin, 2014). As a consequence, many individuals have missed the opportunity to receive treatment in the early stages and thus end their lives as their conditions worsen. Taking into account this problem, it is critical to find an effective tool for the screening of depression in the Chinese population, so that mental health specialists can provide intervention and treatment as early as possible, and prevent depression relapse as well as premature death. I plan to address this in my thesis using two studies——Study 1 to translate a recently developed screen (i.e., the HDS-OA) and Study 2 to provide preliminary reliability and validity evidence with a general Chinese sample.

In the introduction chapter, I have raised the problem of under-recognition of depression in China and shown the need for a new screen, so the remaining sections will serve to address this issue. What is coming next is a literature review chapter that compares the diagnostic criteria and screening instruments of depression between the Chinese and Western cultures, introduces the scale to be adapted in my thesis, and provides a theoretical framework for translation and validation. Following this chapter is a free-standing two-study manuscript. In this manuscript, Study 1 describes the adaptation procedures, which included one round of forward translation, two rounds of backward translation, a pre-test with a four-person expert panel and a pre-test with a participant panel of 26 Chinese adults. Study 2 presents psychometric evidence for the newly

translated depression measure in a sample of 364 Chinese adults drawn from the general population. Specifically, internal consistency reliability, internal (factor) structure, and convergent and discriminant validity were examined. Strengths and limitations of each study are also discussed in this chapter. The thesis ends with a conclusion chapter that offers a brief overview of findings from these two studies, summarizes some directions for future research and provides insight into how high-quality translation can be obtained.

Chapter 2: Literature Review

Types of depression

In clinical research, the term *depression* can be understood from multiple levels. To be more specific, it can be a symptom, syndrome or nosologic disorder (Kendall, Hollon, Beck, Hammen, & Ingram, 1987). As a symptom, it can mean depressed mood, which is a core manifestation of various depressive disorders. When being described as a syndrome, depression refers to depressive episodes with the occurrence of a cluster of symptoms such as sadness, loss of interest, fatigue, insomnia and weight loss. It can be seen in not only individuals diagnosed with depressive disorders, but those with physical illness such as stroke, thyroid disorder or cardiovascular disease (Goodwin, 2006; Zhou, 2012). Finally, as a nosologic category, the third level of depression denotes different types of depressive disorder, including but not limited to major depressive disorder (MDD), dysthymia, and psychotic depression. Each disorder has its own clear diagnostic criteria for inclusion and exclusion.

One of the most commonly seen subtypes of depressive disorder is MDD, which is also the focus of this study. According to the widely accepted Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), the severity of MDD can be classified as mild, moderate, or severe, depending on the number and combinations of symptoms. In any case, at least five of the following symptoms should be present for nearly every day during the same two-week period for the diagnosis of MDD: (1) depressed mood, (2) loss of interest in most activities, (3) significant change in weight (5% or more) or appetite, (4) insomnia or hypersomnia, (5) psychomotor agitation or retardation, (6) fatigue or loss of energy, (7) feeling worthless or excessively or inappropriately guilty, (8) diminished ability to think or concentrate or make

decisions, and (9) thoughts of death or suicide or having a suicide plan (American Psychiatric Association [APA], 2013).

In fact, these nine symptoms can be categorized into two dimensions in terms of symptomatology, which are somatic complaints and affective or cognitive disturbance. More specifically, somatic complaints include symptoms (3) to (6), while affective or cognitive disturbance comprises symptoms (1), (2), (7), (8) and (9) (Tylee & Gandhi, 2005). Because somatic symptoms of MDD may overlap with the symptoms of a variety of other disorders, such as anxiety disorders, chronic pain, and dementia, either of the core symptoms of depression, namely, depressed mood and loss of interest, must be present when making a definitive diagnosis of MDD (APA, 2013; Ellis, Robinson, & Crawford, 2006; Kapfhammer, 2006).

In addition to the DSM-5, many clinicians in China also adopt a local diagnostic system called the Chinese Classification of Mental Disorders Version 3 (CCMD-3). The first version of the CCMD was published in 1985 and the latest version, CCMD-3, was released in 2001 (Chinese Society of Psychiatry, 2001). It matches the fourth edition of the DSM (DSM-IV) and another popular classification standard named the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) to a great extent because its descriptive definitions and diagnostic criteria refer to these two international systems (Dai et al., 2014). At the same time, locally salient features were also included to ensure the adherence to etiology and pathology in the Chinese context; hence, it is widely accepted by psychiatrists throughout the country (Chen, 2002).

The symptoms listed in the CCMD-3 for diagnosis of MDD are almost identical to those in the DSM-5, except that the symptom (8) in the DSM-5 is rephrased as “having difficulties in making associations and diminished ability to think” and “reduced libido” is added as a

depressive symptom. The greatest difference between the CCMD-3 and DSM-5 is that the Chinese standard requires depressed mood and four or more of the other nine symptoms to be present for at least 2 weeks; namely, loss of interest is not a core symptom of depression in the CCMD-3. Still, the way that it defines depression does not seem to deviate from the most popular international classifications, which implies that the meaning of depression is more or less consistent across cultures.

Apart from MDD, another main subtype of depressive disorder is referred to as persistent depressive disorder (PDD), more commonly known as dysthymia. The diagnosis of PDD is somewhat similar to MDD, with depressed mood and two or more of the following symptoms lasting for at least 2 years: (1) poor appetite or overeating, (2) insomnia or hypersomnia, (3) low energy or fatigue, (4) low self-esteem, (5) poor concentration, and (6) feelings of hopelessness. Additionally, the DSM-5 proposes that the symptoms of PDD must not be absent for over two consecutive months during the two-year period (APA, 2013). As the severity is less intense and the duration is much longer compared with MDD, PDD can be considered a chronic form of mild depression.

Other subtypes of depression include psychotic depression, where individuals experience not only the symptoms of MDD but also psychotic symptoms, such as delusions and/or hallucinations (Hales, 2008). Depressive disorder is also common in women during pregnancy and in women after childbirth, due to hormonal changes in their body; therefore, this kind of depression is described as antenatal and postnatal depression, respectively (“Postpartum Depression Facts”, n.d.).

Cultural differences in depression

Although depression seems to be universal across cultures in terms of descriptive definitions and diagnostic criteria, previous studies provide support for group differences between China and Western countries in reporting depressive symptoms. A common pattern is that patients in the Chinese culture somatize depression, by reporting somatic symptoms more frequently than patients in Western cultures. Several explanations have been proposed to illustrate this somatization phenomenon.

The first explanation is that language has a direct impact on subjective psychopathological experiences, which results in different manifestations of depressive symptoms across cultures (Zhou, 2012). That is, a lack of vocabulary for affective states in the Chinese language increases difficulties in verbalizing inner feelings; hence Chinese people tend to articulate depression physically when they are asked to express their feelings (Kleinman, 1982; Zhu & Wang, 2011). However, when a scale was used to measure the difficulties with clearly identifying and describing emotional states, Chinese and Euro-Canadian psychiatric outpatients did not show significant difference in scores, indicating that the variations in languages are barely responsible for the tendency to report somatic symptoms (Ryder et al., 2008).

Another explanation is “somatosensory amplification”, which refers to the tendency to experience bodily sensations as being particularly intense and disturbing due to the effect of psychological distress on the perception of psychiatric patients (Nakao & Barsky, 2007). Simon and his colleagues (1999) examined this explanation using a sample of 1146 patients with major depression from 14 countries in Western Europe, North America, South America, Africa and Asia. They computed the proportion of patients with at least three medically unexplained, currently bothersome, somatic symptoms. The results implied that the reporting of multiple

unexplained somatic symptoms was common across sites (50%), and there were no significant variations between China and other 13 countries. That is to say, the true physical conditions of people with depression are in fact consistent across cultures; therefore, somatosensory amplification does not seem to account for the tendency for somatization in the Chinese group.

The third explanation emphasizes the denial of psychological distress. As mental illnesses have been stigmatized in the Chinese culture for a long time, people attempt to camouflage their psychological problems as physical disorders to "inhabit the sick role in their societies without bearing the burden of stigma" (Ryder et al., 2008, p. 302). From this perspective, even among individuals who share similar psychological experiences across cultures, patients in China are still more likely to express physiological components of depressive symptoms and suppress the psychological ones. However, results from the same study conducted by Simon et al. (1999) showed that the proportions of patients rejecting the two most apparently psychological symptoms (i.e., depressed mood and feelings of guilt or worthlessness) did not vary significantly among the sites and the overall prevalence was only 11%. Therefore, stigmatization of mental illness may not serve as the main cause of the somatization tendency.

The last interpretation is that somatization is a situation-oriented coping pattern that brings adaptive or maladaptive consequences depending on the health care systems of a society (Kleinman, 1982). Due to the shortage of psychiatrists and the unpopularity of psychotherapy, most people with depression in China are more willing to meet a physician who does not specialize in psychiatry and emphasize physical complaints when describing the symptoms (Mao, 2013). This way, they can get medical treatment as soon as possible and save money. Consequently, somatization, as a socially efficacious way for obtaining health care resources in Chinese settings, is more common in China than in other Western countries.

Simon et al.'s (1999) research examined this explanation by comparing the proportion of patients who reported only somatic symptoms as the reason for seeking help from the physician. It turned out that this tendency was significantly more common in countries which offered walk-in care without an ongoing patient-physician relationship (e.g., Greece: 91%, China: 87%, Japan: 77%) than countries which offered a more personal form of primary care with detailed medical records and scheduled appointments (e.g., Chile: 68%, Italy: 53%, France: 45%) (Simon et al., 1999). Therefore, it is the health care system, instead of the cultural background or economic development, that appears to lead to the cross-national differences in somatization.

Parker, Cheah and Roy's (2001) study also provided support for this interpretation. In this research, participants were asked to nominate the most important clinical feature that had initially encouraged them to seek psychiatric assistance. It turned out that somatic symptoms were more frequently nominated as the most important feature by the Chinese sample (Chinese: 60% vs Australian: 13%), whereas symptoms related to depressed mood or cognition were more frequently reported by the Euro Australian sample (Australian: 47% vs Chinese: 25%).

Generally, such evidence suggests that the tendency to report somatic symptoms is not due to an inability to describe affective states, a different experience of physical conditions, nor an unwillingness to express psychological distress, but a culturally effective presentation mode to seek help from health services in the Chinese context. Hence, even though people with depression in China are inclined to express depression somatically compared with those in Western countries, we can still say that the construct of depression itself remains consistent across cultures.

Forms of depression assessment

Traditionally, there has been a wide range of assessment methods for depression: projective tests that can reflect underlying psychological processes, structured or semi-structured interviews that need to be administered by trained clinicians, and symptom inventories that have little requirement for users to have a psychological background or experience (Srivatsan et al., 2018). Among these assessments, self-report measures are shown to be an appropriate approach to evaluate the level of depression in the Chinese population, regardless of the somatization tendency (Ryder et al., 2008).

A study conducted by Ryder et al. (2008) examined the effect of evaluation methods on somatization in Chinese and North Americans with three assessment modalities: (1) a spontaneous problem report on an unstructured clinical interview, (2) a structured clinical interview administered by clinicians, and (3) a symptom rating questionnaire consisting of multiple depression scales. Chinese subjects were significantly more likely to report somatic symptoms on the spontaneous problem report and the structured clinical interview compared with North Americans. However, the effect disappeared entirely when a questionnaire was given to each participant privately. That is to say, even if people have different modes of presenting depression in various cultures, affective states would still be manifest when individuals evaluate their own experiences by self-report inventories. This finding indicates the effectiveness of self-report symptom inventories as a method for measuring depression in the Chinese group and the appropriateness of adapting Western measures into the Chinese language, despite the tendency of somatization in this specific cultural context.

Therefore, this research will focus on self-report scales, which are one of the most important and common assessment modalities for depression. Over the past several decades, researchers have developed a large number of depressive symptom measures with Western

cultures, such as the Beck Depression Inventory (BDI), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hamilton Depression Rating Scale (HAMD), Hospital Anxiety and Depression Scale (HADS), Patient Health Questionnaire-9 (PHQ-9), and the Zung Self-rating Depression Scale (SDS).

Most of these depressive symptom inventories were developed for screening purposes; namely, they are used to estimate the level of risk of depression and determine if any further assessment is warranted (Anderson, Michalak, & Lam, 2002). Some scales are also valid for quantitatively evaluating patients' responses to treatment or monitoring their severity over time, such as the BDI and HAMD (Anderson et al., 2002). However, results from these measuring or screening tests are merely an indication of the level of depression, not a definitive diagnosis. Unfortunately, most of the time this fact is neglected by physicians, who have misused self-report screening scales as a diagnosis tool and prescribed medication without further clinical examinations (Lee et al., 2010). In fact, to establish the presence or absence of depression, clinical interviews that are administered by a qualified professional should be offered to symptomatic individuals after a positive screening test result (Ruf, Morgan, & Mackenzie, 2017).

Western measures of depression

Important evidence to support the use of depression instruments is good reliability and validity in the target population. Research has provided support for many commonly used depression scales in Western populations and, subsequently, some researchers have translated them into Chinese versions and applied them to Chinese samples. Using the keywords "depression", "depressive symptoms", "scale", "questionnaire", "inventory", "China" and "Chinese" to search papers in five databases, Yan, Xiao, and Hu (2016) found that the top three translated scales that have been widely used in China were the SDS, HAMD and SCL-90 (see

Table 2.1). Another systematic review conducted by Jin and Zhang (2017) showed similar results that more than 16,000 articles adopted the Chinese version of SDS during the past 25 years (see Table 2.2). Thus, the SDS ranked first among widespread depression scales in China, followed by the HAMD, SCL-90, CES-D and HADS (Jin & Zhang, 2017; Yan et al., 2016). It is perhaps not surprising that the SDS is the most commonly used measurement tool because it is one of the earliest introduced depression inventories in the country (Wang & Chi, 1984). Despite this, the psychometric properties of the SDS have not been as well explored as some depression measures available for epidemiological research. Instead, among available papers exploring the

Table 2.1.

Number and Proportion of Articles using Western Depression Measures Published between 2009 and 2013 in Yan et al.'s (2016) Review

Measure Name	Measure Acronym	Number	Proportion
Zung Self-Rating Depression Scale	SDS	5812	34.35%
Hamilton Depression Scale	HAMD	5521	32.63%
Symptom Check List 90	SCL-90	2563	15.15%
Hospital Anxiety and Depression Scale	HADS	797	4.71%
Beck Depression Inventory	BDI	540	3.19%
Center for Epidemiological Studies Depression Scale	CES-D	480	2.84%
Geriatric Depression Scale	GDS	408	2.41%
Edinburgh Postnatal Depression Scale	EPDS	364	2.15%
Depression Self-Rating Scale for Children	DSRSC	93	0.55%
Children's Depression Inventory	CDI	80	0.47%
Patient Health Questionnaire-9	PHQ-9	74	0.44%

Table 2.2.

Number and Proportion of Articles using Western Depression Measures Published between 1992 and 2016 in Jin and Zhang's (2017) Review

Measure Name	Measure Acronym	Number	Proportion
Zung Self-Rating Depression Scale	SDS	16070	80.88%
Center for Epidemiological Studies Depression Scale	CES-D	1147	5.77%
Hospital Anxiety and Depression Scale	HADS	840	4.23%
Geriatric Depression Scale	GDS	604	3.04%
Edinburgh Postnatal Depression Scale	EPDS	514	2.59%
Beck Depression Inventory	BDI	297	1.49%
Patient Health Questionnaire-9	PHQ-9	262	1.32%
Children's Depression Inventory	CDI	135	0.68%

psychometric properties of depression scales in China, most articles focused on the PHQ-9 and CES-D, accounting for 36.4% and 31.2%, respectively (Sun, Li, Yu, & Li, 2017; see Table 2.3). Based on this review, only a handful of articles provided validity evidence related to the BDI, HAMD and GDS (Sun et al., 2017). It is worth noting that the main reason why the Chinese version of PHQ-9 has not been extensively used is that this scale was not translated into Chinese until 2007 (Jin & Zhang, 2017). Nevertheless, it is one of the depression measures with considerable validity evidence examined with the Chinese population and has become increasingly popular in recent years (Jin & Zhang, 2017).

Based on the frequency of use in clinical and research practice and the availability of psychometric evidence, I will selectively describe three depressive symptom instruments – the SDS, CES-D, and PHQ-9 – in more detail.

Table 2.3.

Number and Proportion of Articles Validating Translated Western Depression Measures Published before May, 2016 in Sun et al.'s (2017) Review

Measure Name	Measure Acronym	Number	Proportion
Patient Health Questionnaire-9	PHQ-9	16	36.36%
Center for Epidemiological Studies Depression Scale	CES-D	14	31.82%
Geriatric Depression Scale	GDS	5	11.36%
Beck Depression Inventory	BDI	5	11.36%
Hamilton Depression Scale	HAMD	4	9.09%

Zung Self-rating Depression Scale. The SDS consists of 20 items assessing depression for patients diagnosed with depressive disorder, in which 10 questions are regularly keyed and 10 questions are reverse keyed (Zung, 1965). According to the common characteristics of depression, these items can be classified into four subscales: pervasive affect, physiological equivalents, other disturbances, and psychomotor activities (Lipovac et al., 2010). Each item is scored using a Likert-type response format ranging from 1 ("*a little of the time*") to 4 ("*most of the time*"); total scores range from 20 to 80. Higher scores represent a higher level of depression.

The psychometric studies of the Chinese version of the SDS (C-SDS) showed inconsistent results. In a sample of 501 women in rural areas, whose average age was 50 years, Peng et al. (2013) reported Cronbach's alpha coefficients for four subscales as .35, .47, .60, and .74, and .78 overall, which would be considered unacceptable, poor, questionable, and acceptable, respectively (George & Mallery, 2003). Research using a sample of 193 elderly with an average age of 73 years old, however, yielded a two-factor structure and only reported an overall Cronbach's alpha of .91 (Lee et al., 1994). They found high correlations between the

SDS and Chinese versions of the HAMD (r = .86) and the GDS (r = .88) (Lee et al., 1994). However, these analyses were problematic because a multidimensional structure negates the use of a total score or an alpha based on a total score. In addition, some researchers have suggested that the SDS puts an emphasis on physical pain, which may inflate reliability when it is calculated in an elderly group (Robinson et al., 1997).

Finally, another study exploring test criterion evidence for the validity of the inferences made from the C-SDS in non-psychotic outpatients reported that the area under the curve (AUC) was less than 0.7 (Duan & Sheng, 2012), suggesting that the measure may not be useful for screening depression (Metz, 1978). The specificity (36%) in this sample with a recommended cutoff score of 40 was not very good either. Generally, good reliability evidence for an elderly group was obtained through problematic analyses, and the overall psychometrics properties of the C-SDS are not satisfactory for the general population.

Center for Epidemiological Studies - Depression Scale. The CES-D is a 20-item multiple-choice screening tool used to measure the level of depressive symptoms in the general population for epidemiological studies, mainly focusing on the affective component or depressed mood (Radloff, 1977). It is worth noting that the purpose of the CES-D differs from the SDS, which was designed "chiefly for diagnosis at clinical intake and/or evaluation of severity of illness over the course of treatment" (Radloff, 1977, p. 385). Participants are asked to rate how often they experienced depressive symptoms in the past week, including sleep disturbance, feelings of hopefulness, loss of appetite, and so on. The items are scored on a scale of 0 ("*Rarely or none of the time [less than 1 day]*") through 3 ("*Most or all of the time [5-7 days]*"), with a total score ranging from 0 to 60. Individuals scoring higher on the CES-D have a higher level of depression.

Although the Chinese version of the CES-D (C-CES-D) is not as popular as the C-SDS, its reliability and validity have been more thoroughly investigated in the general population, with more than 20 articles studying the psychometric properties in a wide array of groups, such as young adolescents, middle-aged adults and community elderly people. He et al. (2013) and Zhang and Li (2011) carried out validity studies with samples of 30,801 and 16,047 (respectively) nonclinical Chinese people of all ages. The results indicated good internal consistency, with Cronbach's alpha between .85 and .90, in both samples. In Zhang et al. (2015), similar results (Cronbach's alpha = .85) could be found using a sample of Chinese patients with type 2 diabetes. A subsample of 40 patients was chosen for retest within 2-4 weeks, from which poor test-retest reliability with coefficients of .64 was obtained. With respect to validity, the C-CES-D was moderately strong correlated with a measure of mental health (r = - .69) (Zhang & Li, 2011). But the correlation between the C-CES-D and the Beck Depression Inventory – II (BDI-II), was relatively smaller (r = 0.61) (Zhang & Li, 2011), which did not make sense, because the BDI-II was supposed to measure the same construct as the CES-D.

In a sample of 3686 Chinese adult primary care patients, Chin and colleagues (2015) found high correlations between the C-CES-D and the Chinese versions of the PHQ-9 (r = .78) and Short Form-12 Health Survey (r = -.75). But an AUC of .75 indicated that the CES-D did not differentiate groups with and without depression very well. Excellent two-week test-retest reliability (r = .91) was reported based on a small sample of 383 adults who agreed to take the second evaluation. These studies provided good internal consistency evidence for the C-CES-D, but whether this scale reliably and validly measures depression in the Chinese samples is still questionable in view of the inconsistent test-retest reliability coefficients and convergent validity, as well as barely acceptable criterion validity.

Patient Health Questionnaire-9. This self-administered instrument was designed to assist health care professionals in assessing and monitoring the severity of depression (Kroenke, Spitzer, & Williams, 2001). It consists of nine items that correspond to the nine symptoms described in the DSM-IV for MDD. Each item generates a score ranging from 0 ("*not at all*") to 3 ("*nearly every day*"), so the total scores range from 0 to 27. Higher scores reflect a higher level of depression.

There are also over 20 papers exploring psychometric properties of the Chinese version of the PHQ-9 (C-PHQ-9), using various populations (Sun et al., 2017). Compared with the reliability of C-SDS and C-CES-D scores, C-PHQ-9 scores appear to be more reliable in the general population. In two samples of 6,028 and 1,045 people from the general population (Yu, Tam, Wong, Lam, & Stewart, 2012; Wang et al., 2014), the Cronbach's alpha coefficients were acceptable at .82 and .86, respectively. The two-month test-retest reliability estimate was .76 (Yu et al., 2012) and the two-week test-retest reliability estimate was .86 (Wang et al., 2014).

The C-PHQ-9 also yielded good validity evidence, as it was found to be moderately correlated with a measurement relevant to mental health (r = - .60) and weakly or less highly correlated with a health status questionnaire containing eight subscales (r ranging from -.11 to -.47) (Wang et al., 2014). Despite its good reliability and construct validity evidence, some researchers argue that the PHQ-9 has a drawback of high specificity but low sensitivity (Sahni & Agius, 2017). For example, Zhang et al. (2013) found acceptable specificity (84.2%) but poor sensitivity (56.5%) with the recommended cut-off score of 10. Thus, many depressed patients may not be identified and receive treatment. In general, the C-PHQ-9 demonstrated satisfactory reliability and validity evidence, except that sensitivity was not high enough to warrant the use of this scale to screen depression.

Though thousands of studies have used these three translated depression inventories and some of the results have revealed adequate reliability and validity, there may be issues with the translation process used. Going back to the original sources, either nothing was reported about the translation process or only a brief description of the translation process was found, typically indicating that the measures were first translated into a Chinese version and then proofread by different translators (Robinson et al., 1997; Zhang, 1993). Importantly, all of these measures, except the PHQ-9, were translated before 1996, when no single complete standard for adapting tests was ever released and translators did not necessarily follow a rigorous and systematic procedure in doing their work (Yan, Xiao, & Hu, 2016). Therefore, Yan et al. (2016) pointed out that some items in these measures might not have been precisely interpreted, which could have led to a lack of equivalence of meaning between the English and Chinese versions. For example, in the SDS, the English item "I enjoy looking at, talking to, and being with attractive women/men" was translated as "I feel as happy as I used to be when I interact with the opposite sex" in the Chinese version (Zhang, 1993). Other examples, in the CES-D, included the English items "I had crying spells", which was translated to "I used to cry", and "I could not get going", which was translated as "I walked slowly" in the Chinese version (Zhang, 1993). In addition, in the CES-D, the item "I was bothered by things that usually don't bother me" was translated as "I was troubled by some small things", and the item "I thought my life had been a failure" was translated to "I thought my life was nothing" (Zhang, 1993).

Obviously, the meaning of these items was not equivalent to that in the original scales, and such inaccurate translations would impair the effectiveness of these measures (He et al., 2013; Yan et al., 2016). This may be one of the important reasons for the questionable or even poor reliability and validity evidence when using the SDS, CES-D, and PHQ-9 in some samples

(Yan et al., 2016). Considering the potential translation and adaptation problems, these depression screens may not be a good choice for the Chinese population, despite the fact that they have been widely accepted for a very long time and sometimes they have yielded acceptable psychometric properties.

Chinese measures for depression

Apart from making direct translations of Western surveys, a few Chinese researchers have made efforts to adapt some English scales by removing/adding items or simply develop new measures within the Chinese culture, in order to make them more culturally sensitive. But it has to be noted that none of these measures are as popular as the directly translated ones. Citation counts reveal that these depression scales have only been used in a few studies, so the psychometric properties have not been thoroughly examined yet. In order to give a clear picture of the existing Chinese-developed depression measures, Table 2.4 is provided with a brief description of the scales, including the names, original articles, psychometric properties, studies that have used them, and how many times they have been used.

It can be seen that, though most of these scales have been in the literature for over a decade, only a few (SSDA, CDSS) have been used by other researchers. Therefore, it is difficult to justify which scales should be chosen to evaluate the level of depression in the Chinese population, as evidence for the quality of these scales has not been sufficiently examined yet.

In addition to a lack of reliability and validity evidence, another problem is that most of the researchers did not give a detailed explanation of how they developed the instrument in their paper. Therefore, readers have no information about how they collected and chose items to form their scale, which makes it even harder to determine the measures' quality. As a result, neither the contexts nor characteristics of the instrument are able to be fully understood; more

Table 2.4.

Summary of Chinese Measures for Depression

Measure Name	Measure Acronym	Measure citation	Psychometric properties	Number of items and sample item	Studies that used the measure	Citation counts
Self-rating Scale for Depression in Adolescents	SSDA	Wang, Qiu, & He, 1997	Cronbach's alpha = .95, split-half reliability = .87, test-retest reliability = .72 (55 days), .76 (45 days), .80 (10 days), a correlation of .67 with the CES-D	20 (I felt depressed.)	Li, Qiu & Wang, 2001; Wang & Ding, 2003; Wang & Wang, 2005	17
Chinese Depressive Symptom Scale	CDSS	Lin, 1989	Cronbach's alpha = .90, a correlation of .92 with the CES-D	22 (I feel suspicious of others.)	Yan, Robins, & Lin, 2000	13
The 9-item version of the CES-D	CES-D-9	He, Chen, Guo, Zhang, Yang, & Wang, 2013	Cronbach's alpha = .85-.88, eight-week test-retest reliability = .49, a correlation of .94-.96 with the CES-D	9 (I felt lonely.)	Huang, Guo, Wang, & Chen, 2017; Liu, Zheng, & Ye, 2016	4

Measure Name	Measure Acronym	Measure citation	Psychometric properties	Number of items and sample item	Studies that used the measure	Citation counts
Depression Symptoms Rating Scale of Traditional Chinese Medicine	DSRS-TCM	Wang, Hu, Chen, & Chen, 2005	Cronbach's alpha = .92, split-half reliability = .89, twenty-day test-retest reliability = .92, a correlation of .87 with the HAMD	21 (I felt fatigued.)	Ma, Zhao, Li, Wang, & Rong, 2014; Shi, Liang, & Gao, 2016	2
Adolescent Student Self-Rating Depression Scale	ASSR	Wang, 2007	Cronbach's alpha = .89, split-half reliability = .84, fourteen-day test-retest reliability = .72, a correlation of .80 with the SDS, sensitivity = 78.2%, specificity = 84.3%	31 (I did not feel like eating; my appetite was poor.)	Zhang, 2011	1
Depression Rating Scale for the Elderly	DRSE	Zhang, Shen, & Zhang, 1992	Cronbach's alpha = .85, split-half reliability = .88, a correlation of .88 with the HAMD	47 (I lost weight.)	Zhang & Ma, 2008	1
Depression Screening Scale for Community Elderly	DSSCE	Lei, He, Cao, Zhao, Wang, & Wang, 2013	Cronbach's α = .98, one-week test-retest reliability = .68, a correlation of .71 with the SDS, sensitivity = 91.2%, specificity = 93.0%	30 (Do you feel you are worthless?)		0

Measure Name	Measure Acronym	Measure citation	Psychometric properties	Number of items and sample item	Studies that used the measure	Citation counts
College Student Self-Rating Depression Scale	CSRDS	Song & Liu, 2007	Cronbach's alpha = .85, split-half reliability = .91	20 (No items provided in the article.)		0
Depression Inventory	DI	Zheng, 1990	Cronbach's alpha = .91, split-half reliability = .90, a correlation of .78 and .62 with the BDI and HAMD	20 (I felt nervous.)		0
The 10-item version of the CES-D	CES-D-10	Xiong, 2015	Cronbach's alpha = .78-.81, eight-week test-retest reliability = .56-.63, a correlation of .95 with the CES-D and 0.56 with the BDI-II	10 (I felt fearful.)		0

specifically, any assumptions and theoretical positions that the researchers held towards the concept of depression and what domains of knowledge they privileged and omitted is unknown (Rowan & Wulff, 2007). Taking these problems into account, I will selectively describe a couple of Chinese-specific depression inventories that have been used more than five times and have a clear description of the development process in the paper, which can at least give us some hints about the quality. The measures are listed as follows:

Chinese Depressive Symptom Scale (CDS). The 22-item CDS consists of two segments. In the first segment, the 16 regular-keyed items in the CES-D were retained but all of the reverse-keyed ones were eliminated because the author argued that "unusually high average ratings" indicate that those items might have been biased (Lin, 1989, p. 126). The second segment included six new items that were designed to reflect "psychiatric complaints that originated in past unpleasant experiences and social relations" (Lin, 1989, p. 123). To be more specific, these items (in English) are: (1) I feel I have a lot to talk about, but can't find the opportunity to say it; (2) I feel suffocated; (3) I feel suspicious of others; (4) I don't think others trust me; (5) I don't think I can trust others; and (6) I remember unpleasant things from the past. Each item generates a score ranging from 0 ("*never*") to 3 ("*from time to time*"), such that the total scores range from 0 to 66, with higher scores reflecting greater symptom severity.

Lin (1989) showed that the Cronbach's alpha coefficient of the original CES-D was only .77 but, after removing reverse-keyed items, it increased to .88, and it was further improved to .90 with the addition of the six new items. Moreover, a principal component analysis suggested that the three factors – somatic/retarded activity, interpersonal problems, and affective mood – in the CDS were parallel to those found in the CES-D in American samples (Lin, 1989). Despite its excellent reliability, using single total scores and reporting a single alpha does not

make sense, because the CDS has a multidimensional structure rather than a unidimensional one. In addition, this inventory seems to be somewhat outdated now as it was developed in 1980s, when social trust in China was at its lowest level and interpersonal interactions were disrupted due to the Cultural Revolution (Wu, 2016). In other words, even though the additional new items improved the psychometric properties of the CDS, they may no longer apply in today's Chinese society.

Adolescent Student Self-Rating Depression Scale (ASSDS). This instrument comprises 30 items and each is rated using a 4-point scale ranging from 1 ("*never*") to 4 ("*always*"), with total scores ranging from 30 to 120 (Wang, 2007). The higher the score, the more often the respondent feel the symptoms. Wang (2007) constructed the ASSDS based on a literature review and expert suggestions, and the items on the ASSDS were drawn from the BDI, SDS, CES-D, Reynolds Adolescent Depression Scale (RADS), Kutcher Adolescent Depression Scale (KADS), and Birleson Depression Self-Rating Scale (DSRS) (Wang, 2007).

Good internal consistency (r = .89) but relatively poorer two-week test-rest reliability (r = .72) was found with a sample of 461 young adults ages 12 to 22 years (Wang, 2007). Three factors found in this scale were named depressed mood, feelings of worthlessness, and somatic symptoms, but they only accounted for 42.6% of the variance in the scale scores (Wang, 2007). Cronbach's alpha coefficients for these three dimensions were .88, .62 and .59, respectively, which were considered good, acceptable and questionable (George & Mallery, 2003). Using the SDS as a 'gold standard' or criterion, specificity was 84.3% and sensitivity was 78.2% with the recommended cut-off score of 58 (Wang, 2007), although use of another self-report scale as a criterion is highly problematic.

The results of the psychometric properties of these two measures seem to indicate that even if a depression scale is developed by Chinese researchers, it does not mean that it will be better a fit with the Chinese culture compared to translated Western measures. In fact, several reasons can be offered to explain why these Chinese-developed scales may have failed to become effective measurement tools. First, items in some inventories (e.g., Chinese Depressive Symptoms Scale) may be representative of a specific time, but they become out-of-date as society changes. Second, some papers lack details about the questionnaires. For example, researchers who created the Depression Rating Scale for the Elderly (DRSE) only mentioned that they used international and Chinese depression scales as references to generate 88 items (Zhang et al., 1992). After discussing with professionals, they modified items that were unclear in expression, eliminated repeated items, and finally reduced the number to 30 items (Zhang et al., 1992). But details about theoretical influences and why they decided to keep certain items were completely missing. As a consequence, even though it demonstrates promising reliability and validity evidence, researchers might not be confident to use it in their studies given the lack of test development information. More importantly, if we took a closer look at these measures, almost all of them were developed based on a certain Western measure or at least included items from different Western measures of depression; in this sense, they were not particularly unique to the Chinese culture.

Adaptation of a new Western depression measure

In general, different issues have undermined the reliability and validity of existing measures of depression used in Chinese studies, whether they were developed in the context of Chinese or Western cultures. Therefore, it leaves us two choices, to either (a) develop an entirely new measure based on qualitative interviews with Chinese individuals about depression, or (b)

adapt a Western measure, if we want to obtain an effective measurement tool to screen for depression in Chinese society. The former one is less preferred and practical, because conducting in-depth interviews is very time-consuming as researchers need to conduct interviews, transcribe them to texts, and summarize information, and is also costlier compared with adapting a new Western measure (Boyce & Neale, 2006). Whether researchers can gather rich and useful information relies heavily on the interviewers' skills and techniques; in other words, biases may be introduced unintentionally and the entire process may be undermined without trained interviewers (Steber, 2017). Hence, if depression does not appear to be experienced differently between Chinese and Western individuals and we already know that Chinese individuals may spontaneously tend toward reporting somatic symptoms in interview formats, then adapting a new Western measure seems to be a wiser choice considering the time and expense as well as the potential to capture depressive symptomatology.

As previously stated, both depressive symptoms and diagnostic criteria described in the Chinese diagnostic system known as the CCMD-3 are almost identical to those in the DSM-5, except for several subtle differences, which means that the conceptualization of depression is the same regardless of culture. Moreover, the tendency of reporting more somatic symptoms in Chinese people is not a reflection of different physical conditions, but a form of culturally adaptive help-seeking behavior in response to the shortage of psychiatrists and the walk-in style of care. Importantly, this tendency is significant in clinical interviews and spontaneous reports, but not in symptom inventories. Therefore, I believe that translating and using a Western screening measure is reasonable. Thus, I plan to adapt a relatively new depression scale developed in Canada and apply it to the Chinese population.

The measure for screening depressive symptoms that I plan to use is the Hubley Depression Scale for Older Adults (HDS-OA; Hubley, 1998). The HDS-OA has 16 items. The items are consistent with the DSM-5 criteria for MDD and PDD and include two additional, but non-scored, questions about the use of new medication and the presence of bereavement (Hubley, 1998). The HDS-OA items include cognitive, affective and somatic symptoms of depression. This measure was designed for use with older adults, and thus uses a dichotomous yes-no response format, large front size, and a reminder of a two-week period at the beginning of each item. Importantly, however, the symptoms of depression do not differ for adults and older adults and thus this measure may be appropriate for adults of all ages and particularly useful when the sample includes individuals across the adult age range and with differing educational levels and cognitive abilities.

Although psychometric evidence for the newer HDS-OA is comparatively limited, results from two validation studies are promising and have supported its use. For reliability, the Cronbach's alpha estimates were 0.88 in a sample of 50 elderly people aged 63-93 (Myers & Hubley, 2012) and 0.94 in a sample of 82 middle-aged and older adults aged 43-85 (Hubley et al., 2009). For convergent validity, Myers and Hubley (2012) reported that scores on the HDS-OA were strongly correlated with scores on the 30-item GDS (r = 0.89) and the 15-item GDS (r = 0.86). Discriminant evidence was demonstrated by moderate correlations with measures of physical health (r = -0.43) and mental status (r = -0.39), and a moderately strong correlation with a measure of anxiety (r = 0.67) (Myers & Hubley, 2012). Both studies yielded high sensitivity (93%, Hubley et al., 2009; 92%, Myers & Hubley, 2012) and specificity (88%, Hubley et al., 2009; 100%, Myers & Hubley, 2012), but one supported a cutoff score of 3 (Hubley et al., 2009) and the other suggested the cutoff score to be 5 (Myers & Hubley, 2012).

It can be seen that the HDS-OA is an effective case-finding tool for depression with good reliability and validity, which provides support for it being adapted into a Chinese version. In addition to the satisfying psychometric properties, the HDS-OA is shorter in length than many commonly used depression measures. For example, the BDI-II has 21 items and each item has a set of 4 statements, which is 84 statements in total. This response format increases the cognitive load and requires more administration time, which leads to higher refusal and drop-out rates in surveys (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993). Therefore, it is beneficial to use a briefer and simpler measure, the Chinese HDS-OA, to reduce the burden of respondents and also to attain adequate response rates in Chinese depression studies.

Another important reason for the applicability of the HDS-OA to the Chinese context is the dichotomous yes-no response format. According to some cross-cultural studies, apart from the emphasis on somatic symptoms, another cultural difference in depression expression is the reporting style of positive affect suppression in Asian ethnic groups. To be more specific, East Asians including Chinese, Koreans as well as Japanese, are more likely to score lower on positively-worded affect items (e.g., “I felt happy”) in depression scales compared with Americans (Iwata & Buka, 2002; Jang, Kwag, & Chiriboga, 2010; Yen, Robins, & Lin, 2000). In contrast, participants score similarly on negative-worded affect items (e.g., “I felt depressed”) across countries, which indicates that Asian ethnic groups have a tendency to suppress expression of positive affect.

This response pattern may be related to the Confucian values that have influenced East Asian countries for thousands of years. As Confucian values highlight “modesty, self-effacement, moderation, social conformity and emotional restraint”, expressing positive feelings, which is valued in Western cultures, may be regarded as immodest and frivolous in the Chinese

culture (Li & Hicks, 2010, p. 228). With the impact of such collectivistic values, positive affect suppression has become an adaptive emotion regulation strategy in East Asian countries, which indeed effectively reduced the depressive experience and physiological arousal in a Chinese sample (Yuan, Liu, Ding, & Yang, 2014). Due to the effect of expressive suppression, it seems helpful to use a yes-no format in a Chinese depression scale to diminish the influence of positive items on validity.

Furthermore, although the HDS-OA is developed to measure depression in older adults, it "has the potential to be used with adults with any age" as well (Hubley, 2014, p. 2992). Previous studies have also demonstrated that a depression scale (i.e., the GDS) designed for use in older adults can be administered to younger adults such as college students and show good reliability and validity (Brink & Niemeyer, 1992; Ferraro & Chelminski, 1996). Accordingly, with a balance of somatic and psychological items, a dichotomous response format and a shorter length, the HDS-OA appears suitable to be applied in the Chinese context.

Thus, in this thesis, I carried out a study to adapt the HDS-OA into a Chinese version using the double-back translation method and collected reliability and validity evidence with a nonclinical Chinese sample. By doing this, I attempted to obtain a well-designed, culturally sensitive measurement tool with good psychometric properties to screen for depression in Chinese adults.

Forms of translation

The cross-cultural adaptation of instruments is a complicated and challenging task, which often comprises two main stages: (1) to translate the chosen measure based on the published guidelines, and (2) to assess the psychometric properties of the translated scale. In order to

clarify these two stages, the forthcoming section will first explain the translation procedures and then describe the assessment criteria in detail.

The term *translation* is defined so broadly that it actually represents different levels of closeness of the meaning from the source language to the target language. The highest level of closeness in translation is called *adoption*, which produces exactly the same meaning as the original without adding any extra interpretation by the translator (Leong, Bartram, Cheung, Geisinger, & Iliescu, 2016). However, this cost-effective option can only be chosen when we believe no bias exists across different cultural contexts. An example would be a rendering of items collecting information about gender, age, height or weight.

Another type of translation is referred to as *adaptation*, which goes beyond the literal meaning of an item and expresses the idiomatic meaning in a particular context, in order to maximize the cultural appropriateness (Leong et al., 2016). For example, the words 'fork' and 'knife' in an item would have to be replaced with chopsticks when translated from English to Chinese, because the situation in the source culture is not commonly seen in the target culture. In this sense, adaptation amounts to a direct translation of a scale with modifications in the wording of items, which shows a moderate translatability. This term has gradually been substituted for *translation* in many publications about test development, because it indicates not only linguistically appropriate but also psychologically adequate translation. Instead of focusing on cross-cultural comparisons, researchers selecting this option emphasize "an adequate coverage of a particular construct" measured in the target culture (Hambleton, Merenda, & Spielberger, 2004, p.53).

The third form of translation, known as *assembly*, is to formulate an item in a totally different way regarding its expression, vocabulary or situation, so as to ensure that the content is

readily acceptable and appropriate to another language culture (Leong et al., 2016). As an example, being self-confident in one's abilities is valued in Western cultures, whereas it is regarded as supercilious and arrogant in the Chinese society. In this case, researchers should choose other behaviors that have similar meaning based on the cultural realities in the target language and make adjustments to that statement. The assembly option can help researchers who aim to identify cross-cultural biases in current theories of cultural psychology, but direct score comparisons across language groups will be no longer meaningful, as salient changes have been made to the instrument (Hambleton et al., 2004).

Generally, we can say that these three forms of translation lie on a continuum, with one end point called semantic translation and the other as communicative translation (Newmark, 1991). Semantic translation requires the translated version to remain loyal to the original instrument without modifying the content of items, even if it appears to be unnatural in the target language (McDermott & Palchanes, 1994). By contrast, communicative translation emphasizes that the original measure and its renditions should be "equally familiar and colloquial in content" to the source and target cultures, and thus allows for adjustments to better communicate a message (McDermott et al., 1994, p.114). That is to say, the equality of textual aspects, including the linguistic meaning, grammar usage, readability and writing style, is the focus of semantic translation, while communicative translation underlines the pragmatics of a language and the importance of laying a test in a broad socio-cultural context (Hambleton et al., 2004).

Of course, an ideal adaptation would involve attaining sematic and communicative translation at the same time. However, the criteria for these two forms of translation do not seem to lead to identical renderings in the present case considering the huge differences between Chinese and Western cultures. In such case, we have to decide whether translating a test more

semantically or communicatively is desirable, given the research goal. Typically, when one's purpose is to determine the existence of a phenomenon in a certain culture, it is suggested that getting close to the extremity of being semantic should be a guiding principle of translation (Jones, Lee, Phillips, Zhang, & Jaceldo, 2001). Contrarily, when one's study is aimed at comparing a phenomenon in different cultures, the translation is supposed to be more communicative, so as to convey the exact contextual meaning of concepts to the audience in both cultures (Jones et al., 2001). As this research aims to find a culturally sensitive measurement tool of depression for the Chinese society, I would like to adopt a position that lies in the middle of the continuum (i.e., adaptation), trying to keep a balance of being both equivalent to the English version and comprehensible to the Chinese population.

Types of equivalence

To achieve the goal of being both equivalent and comprehensible, test developers should take into account four aspects, which are language, culture, construct and measurement properties, when adapting a measure (Leong et al., 2016). From a conventional perspective, translation theorists primarily emphasize the correspondence between the source language and target language, which means that translations should be directly derived from the source language to avoid losing the essence of the original text (Nida, 1964). This type of equivalence between different language versions is called linguistic equivalence. Later on, this restricted traditional scope was broadened by Nida and De Waard (1986), who put forward functional equivalence theory. He linked translation to its functional value, because he believed that focusing on the semantic sense solely made renderings less effective in communicating the source-language message (Nida, 1986). Consequently, he proposed that translations should be both readable (regarding the language) and understandable (regarding the content) to the target

population(s) so that people from different language groups respond to the texts as proximately as possible (Zhang & Wang, 2010). This type of equivalence is known as functional equivalence.

According to Nida (1993, p. 118), the ideal degree of adequacy is that "the readers of a translated text should be able to understand and appreciate it in essentially the same manner as the original readers did". In this sense, functional equivalence is more difficult to achieve than linguistic equivalence; hence, it is suggested that aspects of culture, construct and measurement should be taken into consideration (Leong et al., 2016). For the first aspect, translators should adopt the cultural patterns of the target receptors including, but not limited to, knowledge, belief and social norms to make necessary compensations to achieve cultural equivalence (Zou, 2016). For instance, the English idiom "you can't have your cake and eat it too" figuratively means you cannot have things both ways. However, Chinese people without knowledge of English culture will not understand the real meaning of the phase, so it should be replaced with an equivalent Chinese idiom, such as "you can't have both the fish and the bear's paw at the same time" once translated into the Chinese language.

Another important aspect that needs to be considered is the construct. More specifically, it is expected that the original scale and its adapted version are assessing the same theoretical construct or, alternatively, the construct being measured has the same meaning across cultural groups, which we call conceptual equivalence (Harachi, Choi, Abbott, Catalano, & Bliesner, 2006; Leong et al., 2016). For example, the Big Five personality traits are supposed to be stable across cultures, but, in fact, the results of the adapted Big Five Inventory (BFI) did not support the openness factor in Asian countries and identified a different factor instead (Cheung, van de Vijver, & Leong, 2011). In the literature review on depression and diagnostic symptoms, it appears that the construct has the same meaning in Western and Chinese cultures.

In terms of the last aspect, measurement equivalence holds that the measurement procedure (e.g., paper-and-pencil or computer-based tests, multiple-choice or open-ended questions) must function the same way in different language versions (Leong et al., 2016). An example is that some language groups have a tendency to provide extreme answers to questions, whereas some prefer neutral answers. To deal with this problem, researchers can increase or decrease the number of categories depending on the response style of the population (Hui & Triandis, 1989). In the present study, the use of a dichotomous response format reduces the concern as does previous research, which suggests that Western and Chinese samples respond to self-report symptoms similarly in questionnaires.

These three aspects, namely, cultural, conceptual and measurement equivalence, combined with linguistic equivalence, form the theoretical framework of a new integrated approach that helps to develop a successful adapted instrument. One of the greatest strengths of this inclusive framework is stressing the textual identity with an emphasis on the contextualization of the adaptation, which means it allows for changes in the form as well as in the content to attain equivalence (Leong et al., 2016). But it is worth mentioning that the four aspects are not compensatory, so, in essence, a maximum degree of equivalence would be a combination of high levels of all four types of equivalence. However, in the case that language, culture, construct and measurement properties do not perfectly converge, it is acceptable to specify a hierarchy among the four aspects based on the need of test developers (Leong et al., 2016). In this study, cultural equivalence and linguistic identity seem to be incompatible; thus, my priority would be culture because the study purpose is to develop a culturally sensitive depression measure for the Chinese population.

ITC Guidelines for Translating and Adapting Tests

Translating tests for use in cross-cultural studies can be traced back to more than a hundred years ago, at which time researchers did test adaptions under the guidance of a plethora of scattered technical literature, such as academic journals and books. However, due to a lack of a single complete standard directly addressing the adaptation issue, not all of the translation practices have followed a rigorous procedure; as a consequence, some renderings of instruments failed to reach equivalence between the original and the adapted test. In order to improve the quality of translation and adaptation, the Council of the International Test Commission (ITC) began a project in 1992 to establish a set of practical guidelines for educational and psychological tests, with the cooperation of, and input from, seven other major international organizations – the European Association of Psychological Assessment, the European Test Publishers Group, the International Association for Cross-Cultural Psychology, the International Association of Applied Psychology, the International Association for the Evaluation of Educational Achievement, the International Language Testing Association, and the International Union of Psychological Science (Hambleton, 2001). After four years of drafting, editing and field-testing, the first edition of the *ITC Guidelines for Translating and Adapting Tests* was eventually published in 1996.

The *ITC Guidelines*, as a systematic standard, was developed using the theoretical basis of linguistic, cultural, conceptual and measurement equivalence explained in the last section. Using these four types of equivalence as its core elements, it not only conceptually highlighted the potential linguistic, psychological and cultural problems in the adaptation process, but also provided a structural framework for maximizing and assessing equivalence of the scales. As a synthesis of previous translation work, it provided a wide range of criteria for evaluating overall test quality in terms of reliability, validity and translation procedures. The latest version, released

in 2017, includes 18 guidelines, which can be classified into six categories: pre-condition (3), test development (5), confirmation (4), administration (2), scoring and interpretation (2), and documentation (2) (ITC, 2017). The 'pre-condition' category includes guidelines that need to be fulfilled at the very beginning of the translation process (e.g., obtaining the permission from the intellectual property owner of a test). The second category focuses on the adaptation process and includes everything from the qualifications of translators to translation designs for maintaining the equivalence of the measure. The third category addresses the statistical analyses that should be conducted based on empirical data to provide reliability and validity evidence for a test. And the last three categories of administration, scoring and interpretation, and documentation include guidelines about the use of adapted scales. Because we are interested in the process of adapting the HDS-OA, the five guidelines around the topic of test development are adopted and serve as the foundation of the translation process for this study. Each guideline is briefly described below.

1. TD-1 Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.

According to the *ITC Guidelines* (2017), this is one of the most crucial steps in the entire process because it has been shown that the expertise and experience of translators play an important role in the reliability and validity of the adapted test. To be more specific, it highlights the importance of using "experts" who have "sufficient combined knowledge of (1) the languages involved, (2) the cultures, (3) the content of the test, and (4) general principles of testing" (ITC, 2017, p.11). If the translator only has bilingual competence but lacks a knowledge of cultural specifics of the source and/or target culture, he is likely to produce a word-for-word translation that leads to misunderstandings in the target language group. As such, test developers

should not select bilingual translators simply because they are easily available. Furthermore, a single person cannot be expected to be an expert in all these fields. Therefore, it is highly recommended that specialists with complementary areas of expertise (e.g., translators, test experts, psychologists whose field of interest is depression) work together as a team to manage the adaptation task and accomplish professional quality translations.

2. TD-2 Use appropriate translation designs and procedures to maximize the suitability of the test adaptation in the intended populations.

This guideline suggests that it is necessary to implement a judgmental design during the adaptation process. Two of the most popular judgmental designs are called forward translation and backward translation. Briefly, forward translation is to translate the scale from the source language to target language, and backward translation is to translate the adapted version back into the source language (Brislin, 1970). The similarity between the two source language versions of the test can be seen as observable evidence of equivalence. But before using it as evidence, translators have to determine whether the original and the back-translated version can be considered as equivalent in terms of their similarity. In the process of making judgmental decisions, translators focus on the conceptual similarity rather than the literal similarity, which contributes to improving the suitability of the translated scale in the intended populations (Jeanrie & Bertrand, 1999). However, it is worth noting that both forward and backward translations have shortcomings, so additional judgmental designs ought to be supplemented to strengthen the overall quality of the adapted scale. For example, the *ITC Guidelines* (2017) suggest that researchers can check whether some important features of the source language and target language test are the same with the use of the rating scales.

3. TD-3 Provide evidence that the test instructions and item content have similar meaning for all intended populations.

The implication of this guideline is that it needs to be shown that the translated scale is neither easier nor harder for the target population to understand; otherwise, additional biases will be introduced due to the specific content, and thus impair the validity of the scale (Hambleton et al., 2004). For example, if the adapted measure uses the same stimulus material (e.g., the name of a famous movie star in Western countries), it is likely that the difficulty of the test increases for the Chinese population because they are not equally familiar with the stimulus. Likewise, if the test instructions of an adapted test require participants with different cultural backgrounds to do the same operations, the test may also become more difficult for certain populations, because people in some parts of the world have never done these specific operations before. As such, it is necessary to gather evidence to evaluate the equivalence of the test instructions and item content. To reach this aim, the *ITC Guidelines* (2017) recommend that test developers use reviewers who are native to the local language and culture to assess the equivalence of the translation. They can also recruit samples of bilingual participants to ask for their opinions regarding the similarity and difficulty of the two versions of the scale.

4. TD-4 Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.

The equivalence of item content is necessary but not sufficient for good translations. In order to guarantee the fairness of the adaption, test developers are supposed to take into account the formats, test conventions, and other procedures as well. For instance, test takers in some countries tend to fill in ovals when they select the correct answers, whereas those in other

countries are used to circling the answer. If a scale uses the same response format in different populations, it is more likely that some respondents will make mistakes when marking their answers because of the unfamiliarity or novelty of the item formats. Another example is to develop a paper-and-pencil version of a computer-administered test during the adaptation process to reduce bias caused by a lack of experience with computers in some cultural groups. In addition to the administration modes, item layouts may also introduce testing bias that can distort the results, because some populations prefer graphs and pictures appearing below the relevant text and some are more familiar with graphs and pictures appearing above the text (Hambleton, Yu, & Slater, 1999). Therefore, test developers should ensure that test conventions and procedures are clear, and that presentation modes and formats are equally familiar to all intended populations (ITC, 2017). Rating forms including questions such as "Is the item format, including physical layout, the same in the two language versions?" would be helpful for assessing this guideline (Hambleton & Zenisky, 2010).

5. TD-5 Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made.

As noted by the above-mentioned guidelines, the importance of judgmental techniques cannot be overstated. Nevertheless, we have to admit that translators may not be able to identify all possible test flaws in practice, such as low discriminating power for some items, poor variability in the total scores in the target population, and unclearly stated test instructions. As such, a pilot test is intended to detect potential flaws that cause problems, prior to a large-scale validation study. With a modest sample size, statistical techniques including item analysis, reliability assessment, and validity assessment, can be applied to provide evidence for the

psychometric quality of the preliminary version of the adaptation, and thus revisions can be made to establish a final version.

Translation models

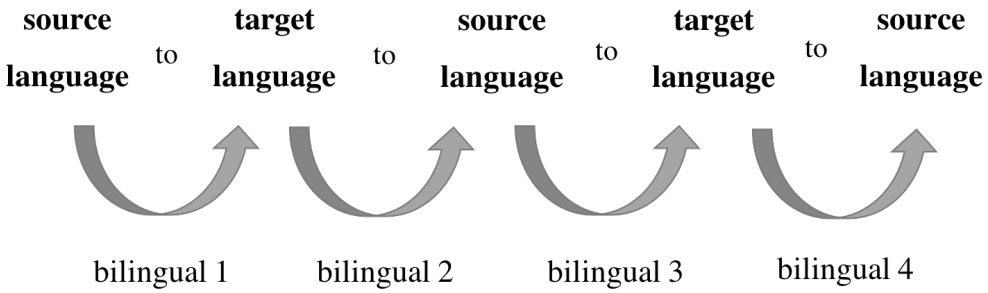
The five guidelines of test development in translation have laid the theoretical foundation of the adaptation process, so the next step would be using practical translation designs to adapt the HDS-OA as per the *ITC Guidelines*. Among translation designs, the classic back-translation model of Brislin (1970) is considered to be the most popular one for cross-cultural research (Cha, Kim, & Erlen, 2007; Hambleton et al., 2004). This model describes the process of forward and backward translation. In forward translation, one translator adapts the scale from the source language to the target language, and then two or more raters compare the two versions to see if they are equivalent. Revisions can be made to the adapted measure if any problematic items are found. However, this design only allows raters to make judgements about the equivalence of the two language versions directly, so it is possible that the raters miss some problems if they are not equally proficient in both languages (Hambleton et al., 2004). As a result, the backward translation design should follow forward translation when adapting measures. More specifically, the target language version will be blindly back translated to the source language by one translator, who does not have access to the source text (Jones et al., 2001). If two new raters determine that the original scale and the back-translated version are identical, the target language version can be considered equivalent to the source language version.

But indeed, equivalence is not necessarily guaranteed by this process even though it sounds logical. For example, if the translator doing forward translation wants to maximize the ease of backward translation, he or she is likely to use words that are closest to the source language but difficult for respondents to understand. In this case, a translator who has a thorough

knowledge in both the source language and target language would still be able to capture the intended meaning, make sense of this poorly translated scale and provide an acceptable back translation (McDermott, & Palchanes, 1994). But it does not mean that the target language version has good quality. Therefore, adaptation should be an iterative process of repeated forward and backward translation by different translators until consistency between the source and target language versions is reached (Brislin, 1986; Figure 2.1). This is, however, very time-consuming and inefficient because each translated version, regardless of Chinese or English, is required to be compared with another version by two raters and then revised, otherwise the following translator cannot move forward to the next iteration.

Figure 2.1.

Brislin's Back Translation Model.



To deal with this disadvantage, Sousa and Rojjanasrirat (2011) reviewed 47 studies focusing on the translation approaches used in cross-cultural research and developed a new guideline based on the framework of Brislin's (1970) model. This guideline is summarized as six necessary steps (Figure 2.2), the details of which are described below.

Step 1: Forward translation. The first step is to forward translate the original scale as described in Brislin's (1970) model. According to Sousa and Rojjanasrirat (2011), at least two translators are needed, and these translators must be fluent in both the source language and target language, preferably native speakers in the target language. Besides that, the two translators

should have different backgrounds (Sousa & Rojjanasrirat, 2011). One of them should be knowledgeable about the content area of the construct being studied, which helps to produce a translation that emphasizes conceptual and measurement equivalence. The other one should be a skilled translator without any clinical background, in order to provide equivalence from a more linguistic and cultural perspective. This way, two target language versions TL1 and TL2 will be obtained. It is worth noting that eligible candidates are not limited to professional translators, and student translators who receive graduate-level degrees are also acceptable for consideration (Wang, Lee, & Fetzer, 2006; Willgerodt, Kataoka-Yahiro, Kim, & Ceria, 2005).

Step 2: Synthesis of two translations. Afterwards, the investigator as well as the two translators from step 1 will compare the two forward-translated versions (TL1 and TL2) along with the original version regarding the instructions, items, and response format. The objective of this step is to resolve the ambiguities and discrepancies of meanings and arrive at a consensus on the most culturally accurate and understandable translation. This step will generate a synthesized Chinese version (TL3) of these adaptations.

Step 3: Backward translation. In step 3, two new translators back translate the TL3 blindly and independently into English, as described in Brislin's (1970) model. The requirements of the translators in this process are pretty similar to those in the forward translation process. The translators will be fluent in both the source language and target language, preferably native speakers in the source language. They should have different profiles as well, with one of them having knowledge of the constructs being examined and the other having no clinical background. This step produces two source language versions of the instrument, which are SL1 and SL2.

Step 4: Synthesis of two translations. Similar to step 2, the investigator and the translators in step 3 will discuss the ambiguities and discrepancies between both the SL1 and

SL2 and the original version after backward translation. Any item that does not retain the same meaning as the original item should be re-translated and step 1 to step 4 should be repeated to translate these items before obtaining the pre-final version of the scale in the target language.

Step 5: Pilot testing of the pre-final version. Before using the pre-final version of the Chinese HDS-OA to collect psychometric data, a pre-test will be carried out to do a final check and solicit feedback to improve the questionnaire. Based on this guideline, a sample size of 10-40 monolingual Chinese people is recommended for the pilot study. They will first be asked to complete the Chinese version of the HDS-OA and then determine whether the instructions, items, and rating scales of the instrument are clear or unclear. They will also be asked to provide comments on any aspect of the scale that they think should be improved. Any instruction or item that is rated as unclear by over 20% of the participants needs to be revised based on the comments.

Moreover, it is recommended that the 'committee approach' should be used to further examine the scale. Specifically, an expert panel that comprises 3-5 bilingual experts will be established. The experts will be asked to evaluate whether each item is translated in a clear and understandable way with appropriate words or colloquial expressions and whether it shares the same meaning in different language versions. With the committee approach, the original English version of the HDS-OA, the pre-final Chinese version, and the backward-translated English versions (SL1 and SL2) will all be provided to the committee members. After reviewing the English and Chinese statements, they will rate the items using a 3-point scale developed by Flaherty and his colleagues (1988), with a score of '1' indicating that the item has different meanings in the different versions, '2' indicating that the meanings in the different versions are almost the same, and '3' indicating that the meanings in different versions are exactly the same.

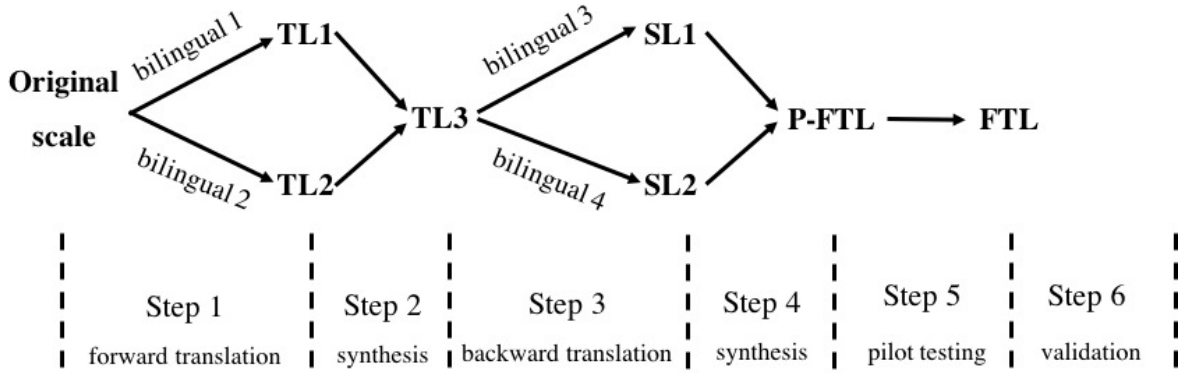
Each item will get an average score based on all of the raters and those with an average score less than '2' should be revised by the committee and rated again until all the items achieve an average score of '2' or higher (Lee, 2009).

In addition to the item content, the experts will be asked to complete a rating scale proposed by Hambleton and Zenisky (2010) and as recommended in the *ITC Guidelines* (2017). This scale lists 25 features that ought to be checked when adapting a measure. Some sample questions include "Is the language of the translated item of comparable difficulty and commonality with respect to the words in the item in the source language version?" and "If a form of word or phrase emphasis (bold, italics, underline, etc.) was used in the source language item, was that emphasis used in the translated item?" In addition to the 25-item questionnaire, the experts' suggestions and opinions about the Chinese version will also be collected. Their comments will be incorporated into the pre-final version of the Chinese HDS-OA and corresponding modifications will be made to produce the final version that fits well with the Chinese culture.

In general, Sousa and Rojjanasrirat's (2011) guidelines suggest that the adaptation process in cross-cultural research should include forward and backward translation, synthesis of different forward- and backward-translated versions, and pilot testing of the pre-final version, so as to enhance the accuracy and utility of the adapted HDS-OA. The last step in this guideline is to conduct a full psychometric test of the final version in the target population, which is described in detail in the next section. An advantage of these guidelines is that the translation process becomes more efficient and integrated, compared with Brislin's (1970) classic model. Therefore, I will follow the procedure of Sousa and Rojjanasrirat's (2011) guidelines to conduct the adaptation part.

Figure 2.2.

Sousa and Rojjanasrirat's Translation Guidelines.



Note. TL means a target language version. SL means a source language version. P-FTL means the pre-final version in the target language. FTL means the final version in the target language.

Reliability

According to Sousa and Rojjanasrirat (2011), a validity study needs to be conducted as the final step of cross-cultural adaptation, in order to demonstrate the psychometric properties of the instrument in the target language and provide evidence for its use with the target population. In this step, not only validity evidence but reliability evidence should be provided to support the use of scores, due to the fact that reliability is a necessary but not sufficient condition for validity. In other words, reliability is needed as a preliminary, albeit not sufficient, step.

As a concept denoting the repeatability of scores, reliability reflects the degree to which test scores are consistent across time, raters, or items (Price, Jhangiani, & Chiang, 2015). Correspondingly, there are three types of reliability estimates: test-retest reliability, inter-rater reliability, and internal consistency reliability. Based on the results of three reviews across disciplines, the internal consistency coefficient is the most common type of reliability estimate, with 74-94% of validity studies providing such information (Barry, Chaney, Piazza-Gardner, & Chavarria, 2014; Hogan, Benjamin, & Brezinski, 2000; Hubley, Zhu, Sasaki, & Gadermann,

2014). It can be used for measures that contain one or more subscales but should be repeated separately for any (unidimensional) subscales. Good internal consistency indicates high item homogeneity, so a single composite score can be interpreted as a reflection of all the items in the full scale or subscale (Henson, 2001).

Comparatively, other reliability estimates have been used with relatively low frequency. Only 14-20% of studied cases in the above-mentioned reviews reported test-retest reliability coefficients (Barry et al., 2014; Hogan et al., 2000; Hubley et al., 2014). One possible cause is a more cost-intensive and time-consuming process, because the measure must be administered to the same group of people at two different times. Besides, it is less applicable for constructs assumed to fluctuate over time, such as mood states, as various factors may change individuals' true construct levels and lead to a low test-retest reliability (Furr, 2017). Inter-rater reliability estimates were barely visible in these studies (0-8.6%), because they are only appropriate for tests that require raters to evaluate test takers' responses and make judgements in the scoring process (Barry et al., 2014; Hogan et al., 2000; Hubley et al., 2014). Due to the aforementioned reasons, internal consistency reliability is favored over the other reliability coefficients for this study.

There are various ways to estimate internal consistency reliability. One approach is to randomly split all items into two halves, compute a score for each half, and then examine the correlation between these two sets of scores (Price et al., 2015). This is called split half reliability. However, with different ways of dividing the items into two parts, large fluctuations are sometimes observed in the results, which makes it hard to trust any of these values (Kuder & Richardson, 1937).

To solve this problem, Cronbach (1951) proposed the coefficient alpha, also known as Cronbach's α. In essence, it is 'the average of all the possible split-half coefficients for a given test' (Cronbach, 1951, p. 300). Conceptually, Cronbach's α refers to the proportion of true score variance of the test to observed score variance (Crocker & Algina, 1986; Rios & Wells, 2014). In this sense, a Cronbach's α coefficient of .80 means that 80% of the variation in tests scores is accounted for by true scores and 20% is measurement error (Van Blerkom, 2017). Statistically, it is defined as $\alpha = \frac{K\overline{c}}{\overline{v}+(K-1)\overline{c}}$. Here, K is the number of items, $\overline{c}$ is the average covariance between item-pairs, and $\overline{v}$ is equal to the average variance of each item (Cronbach's alpha, 2019). Similarly, a standardized α is defined as $\alpha_{standardized} = \frac{K\overline{r}}{1+(K-1)\overline{r}}$, where $\overline{r}$ denotes the average correlations between item-pairs (Cronbach's alpha, 2019). The difference between these two versions of α is whether one uses the raw scores of each item or one standardizes them before computing a composite score (Falk & Savalei, 2011).

Compared with the split half reliability, Cronbach's α has been the overwhelming favorite estimate, as it was found in 87-100% of studies that provided internal consistency reliability (Barry et al., 2014; Hogan et al., 2000; Hubley et al., 2014). In spite of its popularity, Zumbo and his colleagues (2007) argued that the estimate of Cronbach's α may not be accurate when a Likert type response scale has less than five scale points. To further elaborate, alpha is built on covariance or correlation matrices. In fact, most of the time, it is the Pearson correlation matrix that is used by popular software programs, such as SPSS and SAS (Zumbo et al., 2007). However, when the assumption that the data are continuous is violated, the Pearson correlation matrix will be distorted and thus should not be used (Zumbo et al., 2007). In this case, the polychoric correlation matrix, which is designed for examining the correlation between two continuous latent variables from two observed ordinal variables, is a better choice, such as when

Likert type response scales are applied and item responses are ordinal (Polychoric correlation, 2019; Zumbo et al., 2007). Likewise, the tetrachoric correlation matrix for linear relationships between two dichotomous underlying variables are recommended when the data are binary (Polychoric correlation, 2019; Zumbo et al., 2007). The estimate that takes into account the ordinal nature of Likert scaled responses is called ordinal coefficient alpha. Because the HDS-OA produces binary data at the item level, we will conduct reliability analysis to obtain ordinal coefficient alpha for the Chinese HDS-OA.

Five sources of validity evidence

As a validation study, it is worth noting that the major focus of 'validation' here is not the test itself but specific inferences made from the test, because validity is referred to as 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests', according to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). As a further elaboration, a test will not be valid for all purposes or situations (AERA, APA, & NCME, 1974; Shaw & Crisp, 2011). For instance, a test that allows us to measure individuals' depression levels may not provide valid inferences for predicting suicidal attempts. Likewise, inferences from a test may not be equally valid for all groups of people, which means that an inventory designed to evaluate depressive symptoms in the Canadian context may not provide equally valid inferences in the Chinese context (AERA et al., 1974; Shaw & Crisp, 2011). Therefore, validity is never an inherent property of a test and the validation process should always focus on the use of test scores for an intended purpose in a target group (Sireci, 2016).

According to Messick's (1989, p. 13) unifying framework for validity, validity is "an integrated evaluative judgment" that synthesizes multiple evidence sources rather than a singular piece of evidence to support the adequacy and appropriateness of test interpretations and test uses. In his framework, construct validity, which represents the whole of validity, consists of six distinguishable aspects, including: content, substance, structure, generalizability, externality, and consequence (Messick, 1995). Resonating with Messick's unified yet multi-faceted concept, the most recent version of the *Standards* (AERA et al., 2014) suggests collecting evidence from five sources to build solid arguments for validity and justify inferences and actions based on the test scores. These five key sources of evidence are (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing (AERA et al., 2014).

The first source of validity evidence (i.e., evidence based on content) was previously described as content validity and concerns whether the content of a test reflects the measured construct and is congruent with the purposes of testing (Institute of Medicine, 2015; Sireci & Faulkner-Bond, 2014). To obtain such evidence, the test content, including "the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring", should be evaluated by experts to examine its clarity, relevancy and sufficiency (AERA, APA, & NCME, 1999, p. 11).

Evidence based on response processes examines the extent to which respondents understand the measure and answer the items in a way that corresponds with the intended defined construct (Whiston, 2009). An example against it would be that individuals give socially acceptable rather than 'true' responses in a self-report measure because of social desirability. Typically, a useful method for addressing response processes is the think/talk-aloud protocol,

which requires the respondents to make their psychological or cognitive processes as explicit as possible by describing their thoughts or actions (Think aloud protocol, 2019). Apart from respondents, the investigations of response processes can also rely on observers, judges, or raters who record and evaluate performances with certain criteria (AERA et al., 2014; Goodwin & Leech, 2003).

Another source of evidence focuses on whether the actual internal structure of responses to items on a measure matches the theoretically expected patterns (Institute of Medicine, 2015). Different types of analyses are adopted depending on the use of the test. For instance, researchers who attempt to verify if the data align with a theory may examine the interrelationships among items; namely, whether the data show a predetermined single dimension of behavior or multiple dimensions that are homogenous but distinct from each other (AERA et al., 2014). The dimensional structure of a test is usually examined via confirmatory factor analysis (CFA) or exploratory factor analysis (EFA) (Rios & Wells, 2014). Some studies are designed to examine whether people with similar abilities but from different groups, such as racial or gender subgroups, have different performance on specific test items, due to the item content (AERA et al., 2014). These kinds of differences can be reflected by the differential item functioning (DIF) technique and used as evidence of internal structure validity (Rupp & Leighton, 2016).

Evidence that reflects the relationships between obtained test scores and other criterion or more or less related constructs is also important for making claims about validity (Institute of Medicine, 2015). There are two main categories of evidence for this source: One is criterion-based evidence, which gauges to what degree a measure predicts an outcome, usually behavior or performance (Salkind, 2010). It can be divided into concurrent validity and predictive validity, depending on whether the outcome occurs or is obtained at the same time as the measure of

interest or in the future. The other one is convergent and discriminant evidence, which has been historically subsumed under construct validity (McCoach, Gable, & Madura, 2013). Convergent validity focuses on the degree to which test scores of two scales that measure the same or similar constructs are related, whereas discriminant validity focuses on the degree to which scores are related in scales measuring more distinct constructs (Reeves & Marbach-Ad, 2016). Convergent and discriminant measures can be considered as being on a continuum (Hubley & Zumbo, 2013). Statistical techniques for quantifying these external relationships include correlations, multitrait-multimethod (MTMM) matrices, and structural equation modeling (SEM) (McCoach et al., 2013).

The final source of validity evidence focuses on the intended and unintended consequences of legitimate score interpretation and use (Messick, 1998). To be more specific, what researchers should be concerned with during the validation process is not the side effects of test misuse, but "unanticipated adverse effects that are traceable to sources of test invalidity such as construct underrepresentation and construct-irrelevant difficulty" (Messick, 1998, p. 40). Although many scholars recognize the importance of this validity source, the inclusion of consequences as an aspect of validity still remains a matter of debate today (Lane, 2014). Little validation research provides evidence of this sort and only a few strategies could be found for investigating the consequences of the assessments (Kane, 1992; Lane, Parke, & Stone, 1998; Lane & Stone, 2002).

Most Common Sources of Evidence

According to four syntheses of validation practices across a variety of areas, evidence related to internal structure and relations to other variables reflected the two most frequently reported types of validity evidence, while little to no evidence from the other three sources was

presented in most validation studies (Chan, Zumbo, Darmawanti, & Mulyana, 2014; Collie & Zumbo, 2014; Hubley et al., 2014; Shear & Zumbo, 2014).

One possible explanation for the popularity of these two kinds of evidence is the increasingly accessible methods of acquiring evidence; researchers can conduct data analyses (e.g., correlations, factor analysis) by just point and click methods using user-friendly software programs (Zumbo & Chan, 2014). The acquisition of evidence related to response processes, however, is far more time-consuming. Take the think aloud protocol as an example; data need to be collected by one-on-one interviews, transcribed into written form, and coded using specific schema before any analysis can be done (Van Someren, Barnard, & Sandberg, 1994).

In addition to accessibility, the current climate in the field of measurement may also influence the frequency of different types of evidence reported in validation studies. For instance, providing evidence related to internal structure is a requirement for many journals nowadays, whereas evidence based on consequences as a part of validity is still debated or even discouraged (Zumbo & Chan, 2014). Therefore, it can be understood why reporting evidence related to internal structure and external (convergent, discriminant) relationships has gradually become a norm in this area. Given the foregoing information, this validation study will focus on these two aspects of validity. The approaches for obtaining evidence to support a test's internal structure will be described first, followed by approaches for providing evidence related to external relationships.

Evidence based on internal structure: CFA. Based on three recent synthesis, factor analysis is the most commonly used approach for evidence based on internal structure, with 81-100% of entries that examined this sort of evidence presenting this statistical technique (Cox & Owen, 2014; Gunnell, Schellenberg, et al., 2014; Gunnell, Wilson, et al., 2014). Basically, there

are two types of factor analysis: exploratory and confirmatory. As its name implies, EFA focuses on exploring the underlying internal structure of the observed variables without imposing a hypothesized pattern drawn from theory (Child, 1990; Suhr, 2006). It is a good choice when researchers are developing a new scale with little knowledge about the relationships between the measured variables (i.e., items) and latent constructs (i.e., factors) (Yong & Pearce, 2013).

There are several key steps to performing an EFA. The first step is to choose a factor extraction method, such as maximum likelihood, principal axis factoring, or principal components analysis, depending on whether the data meet the normal distribution assumption (Costello & Osborne, 2005). After extraction, researchers need to determine the optimal number of factors, most of the time in reference to the number of eigenvalues greater than one, because this method is the default in most statistical software (Nunnally, 1978). But Costello and Osborne (2005, p. 2) believed that "this is among the least accurate methods for selecting the number of factors", so they recommended use of the scree test and parallel analysis. The scree test is to look for the point when the curve shows a steep decline, which means the eigenvalues decrease dramatically in size (Cattell, 1966). In parallel analysis, a random dataset that matches the same number of observations and items in the original dataset will be created by the statistical software program (Wood, Akloubou Gnonhosou, & Bowling, 2015). Any factor that has an eigenvalue larger than the corresponding one derived from the random dataset should be retained, as it implies that this factor is not due to chance (Wood et al., 2015). In the next step, if the data indicate a multi-dimensional structure, the factor axes should be rotated using oblique rotation (assuming factors are correlated) or orthogonal rotation (assuming factors are uncorrelated). Ideally there will be no cross loadings (i.e., achieving a simple structure). Lastly,

researchers interpret and define the factors using the observed pattern of factor loadings and theory.

Contrary to the data-driven nature of EFA, CFA is more theory-driven. Instead of searching for the most meaningful solution, it focuses on evaluating the predefined internal structure of the observed variables (Suhr, 2006). Hence, it is advisable, though not required, to conduct a CFA using a different sample after an EFA to "evaluate the EFA-informed a priori theory about the measure's factor structure and psychometric properties" (Cabrera-Nguyen, 2010, p. 100; Worthington & Whittaker, 2006). In this sense, CFA can be used in cross-cultural studies to examine if the internal structure of a measure is equivalent across cultures (Watkins, 1989). Researchers can collect data in the new culture and test it with the factor model established in the original culture (Watkins, 1989). Given that Hubley, Rajlic and Zumbo (2017) conducted an EFA of the original English version of the HDS-OA and the results supported a unidimensional structure, a CFA would be appropriate for the present study to examine if a unidimensional structure holds in a Chinese sample. If the one-factor model fits the data from the Chinese version of the HDS-OA, then we have a strong test of internal structure that guides the scoring and use of this measure in the Chinese sample.

As it is 'confirmatory', CFA starts with an a priori hypothesized model derived from theoretical knowledge and/or empirical research, with the number of factors and which items load on which factor being specified in advance. Then researchers need to test if the specified model fits the observed data and several fit indices can help us determine how consistent it is with the data.

There are two broad kinds of model fit measures in CFA: absolute and incremental. Absolute fit indices evaluate the discrepancy of the a priori model from the sample data (Lei &

Wu, 2007). Common examples of this sort include the chi-square value, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) (Hooper, Coughlan, & Mullen, 2008). When a theoretical model is specified, a population covariance matrix is estimated (Schreiber, Nora, Stage, Barlow, & King, 2006). The chi-square statistic basically represents the amount of difference between the estimated and observed covariance matrices (Suhr, 2006). The closer the value is to 0, the better the model fit. A p-value that is insignificant at a .05 threshold is also an indicator of a good fit (Hooper et al., 2008). Notably, this index is very sensitive to the sample size, with large sample sizes nearly always presenting statistically significant differences between the expected model and observed data (Stapleton, 1997). Thus, sometimes it is unclear whether one has a true situation of poor model fit or if it is just the size of the sample that has led to the unwanted significance of the chi-square value.

To deal with this problem, the RMSEA was recommended as a supplementary fit index. It takes into account the sample size and the degrees of freedom when estimating the discrepancy between the hypothesized and observed covariance matrices; hence, it produces an estimation of good quality even with a large sample size (Cangur & Ercan, 2015; Tennant & Pallant, 2012). As for the cut-offs, an RMSEA value smaller than 0.05 corresponds to a convergence fit, the range of 0.05-0.08 represents a good fit, and a value falling between 0.08 and 0.10 is said to indicate a mediocre fit (Cangur & Ercan, 2015; McDonald & Ho, 2002). But Hu and Bentler (1999) also suggested another rule that an RMSEA below 0.06 was sufficient for model evaluation. Another frequently reported absolute fit measure is SRMR, which is the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance matrix (Hooper et al., 2008). A SRMR value of less than 0.05 can be considered as an indicator of a

good fit, and SRMR values ranging from 0.05-0.08 are deemed acceptable (Hu & Bentler, 1999, Kline, 2011).

Incremental fit indices, however, reflect the improvement in fit when the expected model is compared with the worst case scenario (i.e., a baseline model), in which all of the measured variables are uncorrelated (Hooper et al., 2008; Lei & Wu, 2007). Included in this kind are the normed-fit index (NFI), Tucker-Lewis index (TLI) and comparative fit index (CFI). For these incremental fit indices, the larger the values, the better fit for the model, as the hypothesized model increases the fit to a greater extent in comparison to the baseline model.

The NFI statistic compares the chi-square value of the expected model to the chi-square value of the baseline model (Hooper et al., 2008). It is suggested that the NFI value should be greater than 0.95 in a well-fitting model (Hu & Bentler, 1999). There is a disadvantage that NFI is affected by the sample size, with an underestimated fit being presented in small samples ($N < 200$) (Hooper et al., 2008; Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989). Taking into account this problem, TLI and CFI were developed as revised versions of the NFI. These two fit indices are not significantly influenced by the sample size (Schermelleh-Engel, Moosbrugger, & Müller, 2003), and CFI even performs better in studies using a small sample size (Cangur & Ercan, 2015; Hu & Bentler, 1998).The cut-off criteria for these two indices is the same as the criteria for NFI (Hu & Bentler, 1999).

Different indices assess the model fit in different ways, so reporting a combination of multiple fit measures is always recommended for model evaluation. Hu and Bentler (1999) suggested a two-index presentation strategy of reporting SRMR along with RMSEA, CFI or TLI. This combinational rule is said to result in the least sum of underrejection rates of misspecified models and overrejection rates of true-population models in most situations (Hu & Bentler,

1999). Kline (2005) proposed that four indices should be reported, including the chi-square statistic, RMSEA, CFI and SRMR. Based on these criteria, the present study will use the chi-square value, RMSEA, CFI, TLI and SRMR to evaluate the overall model fit.

Evidence based on external relationships: Convergent and discriminant validity. In addition to the expected internal structure, evidence based on relations to other variables is also necessary for making validity claims related to a measure. If the pattern of intercorrelations between the newly developed scale and other external measures is consistent with theoretical expectations, it would be convincing that this scale is measuring what it claims to measure, rather than some other competing interpretations.

In terms of this kind of validity evidence, convergent evidence was reported most frequently (83.7%), according to Hubley et al.'s (2014) validation synthesis. Comparatively, criterion-related evidence was far less common (30.2%). This is understandable because testing for criterion validity is undoubtedly more cost- and time-intensive than convergent validity. For instance, clinical diagnosis derived from a well-established psychiatric interview, such as the Structured Clinical Interview for DSM (SCID), was often used as the gold standard when assessing criterion validity of depression scales (Wancata, Alexandrowicz, Marquart, Weiss, & Friedrich, 2006). This kind of assessment must be individually administered by mental health professionals or sufficiently trained researchers (Structured Clinical Interview for DSM-5, n.d.). Due to the user's different level of clinical experience, the minimum administration time is 30 minutes and the maximum is 120 minutes (Structured Clinical Interview for DSM-5, n.d.). With limited time and money, it makes more sense that we first gather convergent evidence of the newly adapted HDS-OA and test for criterion-related validity in future studies when more resources are available.

Another important subtype of relations with other variables evidence is discriminant validity. It is disappointing that not many researchers include such evidence in their validation studies; only 30.2% of cases did so in Hubley et al.'s (2014) synthesis. In fact, discriminant evidence is as significant as convergent evidence. Instead of viewing them as two separate measurements, I find it more reasonable to understand them as a single analysis. The only difference between them is the magnitude of correlation coefficients.

Conceptually, convergent validity tests whether measures that should be theoretically related are indeed highly related, whereas discriminant validity tests whether measures that should not be theoretically related are indeed not very related (Furr & Bacharach, 2013). However, there is never an absolute standard to determine what range of values should be interpreted as convergent and discriminant evidence. Instead of a 'all-or-nothing' outcome, they are more like a matter of degree (Furr & Bacharach, 2013).

Given the blurred boundary between these two concepts, it is better to put those values on a continuum wherein correlations can be ranked based on the strength of the relationships rather than simply classify them as either convergent or discriminant validity (Hubley & Zumbo, 2013; Hubley et al., 2014). More importantly, when using this kind of evidence, rationales should be provided for why specific constructs and measures are selected and what magnitudes of coefficients are expected so others can judge how well the evidence supports the intended inferences (Hubley et al., 2014).

To choose the appropriate measures for this research, a list of convergent and discriminant measures that have been adopted by previous Chinese studies of depression measures is provided in Table 2.5. The correlation coefficients between these scales and the validated scales are also noted. All of the measures in Table 2.5 refer to the Chinese versions.

It can be seen that existing, well-established scales of the same construct have always been a popular choice when validating a depression measure, but few researchers justified the selection of the comparator instruments in their papers. In fact, using comparator instruments of high quality is the key to achieving good validation results, so it should always be considered carefully whether the comparator instrument shows evidence of reliable scores and valid inferences (Abma, Rovers, & van der Wees, 2016).

According to the available studies seen in Table 2.5, satisfactory validity evidence to support the intended inferences from depression scale scores was not always evident, especially for the BDI, CES-D-10, and GDS-15. Several depression measures did show satisfactory validity evidence, notably the BDI-II, CES-D, GDS, and PHQ-9.

Convergent validity for the BDI-II was demonstrated through the moderately strong correlations between the BDI-II and CES-D (r = 0.70, Yang, Wu, & Peng, 2012; r = 0.72-0.76, Yang et al., 2014). Chin et al. (2015) presented good convergent validity evidence for the CES-D with a relatively high correlation between the CES-D and PHQ-9 (r = 0.78). The 30-item GDS was also found to be relatively strongly correlated with its comparator instrument, the CES-D (r = 0.79, Liu, Wang, Wang, Song, & Yi, 2013; r = 0.96, Chan, 1996). Despite the high correlations, only a few validation articles can be found for these three scales, which does not provide extensive evidence.

Comparatively, the PHQ-9 has been validated more extensively with various populations, including adolescents (Hu, Zhang, Liang, Zhang, & Yang, 2014), middle aged and elderly people (Wang et al., 2014; Xu, Wu, & Xu, 2007), primary care patients (Liu et al., 2011) and patients with post-stroke depression (Zheng et al., 2013). More importantly, it was also validated with a large sample of 6028 people who were 15 years or older, which almost completely coincides

Table 2.5.

Convergent and Discriminant Measures Used in Previous Validation Studies

Measure of Focus	Comparator Measure for Construct Validity	Intended Construct of Comparator Measure	Validity Coefficients (reliability, internal structure)	Reference
BDI	HADS	Depression	0.52 (0.88)	Ye et al., 2013
	HAMD	Depression	0.66 (0.88)	Ye et al., 2013
BDI-II	CES-D	Depression	0.70 (0.85*, 2 factors); 0.72-0.76 (0.89-0.93, 2 or 3 factors)	Yang, Wu, & Peng, 2012; Yang et al., 2014
	HAMD	Depression	0.67 (0.94*, 2 factors)	Wang et al., 2011
CES-D-10	BDI-II	Depression	0.56-0.70 (0.78-0.81, 3 factors)	Xiong, 2015
CES-D-13	CMHI	Mental health	-0.69 (0.71-0.86, 3 factors)	Zhang & Li, 2011
	PSQI	Sleep quality	0.41 (0.71-0.86, 3 factors)	Zhang & Li, 2011
CES-DC	SAS-C	Anxiety	0.63 (0.57-0.82, 4 factors)	Li, Chung, & Ho, 2010
	RSES	Self-esteem	-0.52 (0.57-0.82, 4 factors)	Li et al., 2010
CES-D	PHQ-9	Depression	0.78 (0.43-0.86, 2 factors)	Chin, Choi, Chan, & Wong, 2015
	SF-12 MCS	Mental health	-0.75 (0.43-0.86, 2 factors)	Chin et al., 2015

Measure of Focus	Comparator Measure for Construct Validity	Intended Construct of Comparator Measure	Validity Coefficients (reliability, internal structure)	Reference
	TAI	Anxiety	Suicide attempters 0.46 vs comparison residents 0.58 (0.90-0.94, 3 factors); 0.67 (0.90*, 3 factors)	Yang, Jia, & Qin, 2015; Zhang et al., 2008
	BHS	Hopelessness	Suicide attempters 0.44 vs comparison residents 0.72 (0.90-0.94, 3 factors)	Yang et al., 2015
GDS-15	PHQ-9	Depression	0.50 (0.53)	Wang et al., 2014
GDS	CES-D	Depression	0.96 (0.89); 0.79 (0.85, 1 factor)	Chan, 1996; Liu, Wang, Wang, Song, & Yi, 2013
	SSRS	Social support	-0.46 (0.92, 1 factor)	He, Xiao, & Zhang, 2008
	QOLS	Quality of life	-0.45 (0.85, 1 factor)	Liu et al., 2013
HADS	BDI	Depression	0.52 (0.87)	Ye et al., 2013
	HAMD	Depression	0.85 (0.87)	Ye et al., 2013
	SDS	Depression	0.69 (0.81-0.88, 3 factors)	Sun et al., 2017
	SAS	Anxiety	0.60 (0.81-0.88, 3 factors)	Sun et al., 2017

Measure of Focus	Comparator Measure for Construct Validity	Intended Construct of Comparator Measure	Validity Coefficients (reliability, internal structure)	Reference
PHQ-9	BDI	Depression	0.77 (0.85, 1 factor); 0.80 (0.84)	Hu, Zhang, Liang, Zhang, & Yang, 2014; Zheng et al., 2013
	SDS	Depression	0.85 (0.93, 2 factors*)	Yu, Sun, & Sun, 2017
	HADS	Depression	0.79 (0.86)	Bian, He, Qian, Wu, & Li, 2009
	HAMD	Depression	0.81 (0.86); 0.68 (0.80, 1 factor); 0.75 (no reliability); 0.84 (0.84)	Bian, He, Qian, Wu, & Li, 2009; Liu et al., 2011; Yang et al., 2015; Zheng et al., 2013
	GDS-15	Depression	0.50 (0.88)	Wang et al., 2014
	CHQ-12	Mental health	0.49 (0.82, 1 factor)	Yu, Tam, Wong, Lam, & Stewart, 2012
	SF-12 MCS	Mental health	-0.60 (0.82, 1 factor)	Yu et al., 2012
	SF-12 PCS	Physical health	-0.27 (0.82, 1 factor)	Yu et al., 2012
	Happiness Scale	Happiness	-0.41 (0.82, 1 factor)	Yu et al., 2012
	Q-LES-Q SF	Quality of life	-0.49 (0.80, 1 factor)	Liu et al., 2011;

Note. BDI = Beck Depression Inventory. BDI-II = Beck Depression Inventory-II. BHS = Beck Hopelessness Scale. CES-D = Center for Epidemiologic Studies-Depression Scale = CES-D. CES-DC = CES-D for Children. CES-D-10 = 10-item CES-D. CES-D-13 = 13-item CES-D. CHQ-12 = Chinese Health Questionnaire. HADS = Hospital Anxiety and Depression Scale. CMHI = Chinese Mental Health Inventory. HAMD = Hamilton Depression Rating Scale. GDS =

Geriatric Depression Scale. GDS-15 = 15-item GDS. Q-LES-Q SF = Short Form of the Quality of Life Enjoyment and Satisfaction Questionnaire. QOLS = Quality of Life Scale. PHQ-9 = Patient Health Questionnaire-9. PSQI = Pittsburgh Sleep Quality Index. RSES = Rosenberg's Self-Esteem Scale. SAS = Zung Self-Rating Anxiety Scale. SAS-C = Short Form of the State Anxiety Scale for Children. SDS = Zung Self-Rating Depression Scale. SF-12 MCS = Short-Form 12-item Health Survey Mental Component Summary. SF-12 PCS = Short-Form 12-item Health Survey Physical Component Summary. SSRS = Social Support Rating Scale. TAI = Trait Anxiety Inventory. The reliability estimate with a * sign refers to the overall reliability of the scale and the reliability estimates of the subscales are missing in the study.

with the target population of our study (Yu, Tam, Wong, Lam, & Stewart, 2012). The results of these articles revealed that the PHQ-9 was significantly highly correlated with many depression scales, including the BDI (r = 0.77, Hu et al., 2014; r = 0.80, Zheng et al., 2013), SDS (r = 0.85, Yu, Sun, & Sun, 2017), HADS (r = 0.79, Bian, He, Qian, Wu, & Li, 2009) and HAMD (r = 0.81, Bian et al., 2009; r = 0.68, Liu et al., 2011; r = 0.75, Yang et al., 2015; r = 0.84, Zheng et al., 2013).

These results together indicate that the PHQ-9 is an appropriate convergent measure for our study, as sufficient convergent evidence was provided by the strong correlations between the PHQ-9 and different scales that measure the same construct. Therefore, the Chinese version of PHQ-9 is selected as one of the comparator instruments for the adapted HDS-OA.

In addition to depression, measures of constructs similar to depression should also be included as comparator instruments, and an anxiety scale is a must. Anxiety is defined as "anticipation of a future concern" and usually "associated with muscle tension and avoidance behavior" (American Psychiatric Association, 2017). In terms of phenomenology, these two

constructs are distinguishable, since depression is a feeling of sorrow and gloom, while anxiety is a feeling of apprehension and worry (Watson et al., 1995). However, in spite of this seeming distinctiveness, overlapping symptoms are found in the DSM-5 criteria for diagnosing anxiety disorders and major depression (e.g., fatigue, difficulty concentrating, sleep disturbance) (APA, 2013). Not surprisingly, the contents of specific items also overlap in anxiety and depression scales developed based on these criteria, thereby causing the moderately strong correlations between measures of anxiety and depression, with estimates in the range of 0.58 to 0.67 (Li, Chung, & Ho, 2010; Sun et al., 2017; Yang, Jia, & Qin, 2015; Zhang et al., 2008). Due to these shared symptoms in diagnostic criteria, it is necessary to use an anxiety scale as the comparator instrument to ensure that the Chinese version of the HDS-OA successfully discriminates depression from anxiety.

Based on previous research, the 7-item Generalized Anxiety Disorder Scale (GAD-7) has been validated more often and has shown better psychometric properties in the Chinese population (Cai, 2013; He, Li, Qian, Cui, & Wu, 2010; Qu & Sheng, 2015), compared with other commonly used measures (e.g., the State-Trait Anxiety Inventory [STAI; Chen, Cao, & Liu, 2013; Li & Qian, 1995], Zung Self-Rating Anxiety Scale [SAS; Liu, Tang, Peng, Chen, & Dai, 1995; Tao & Gao, 1994], and Hamilton Anxiety Rating Scale [HAMA; Wang et al., 2011]), so the Chinese version of the GAD-7 will be used as a comparator instrument.

Besides that, a hopelessness scale is also necessary. An important idea in Beck's cognitive theory is that, hopelessness, which refers to "negative expectations about the future", is "both a determinant and a component of the depressive condition" (Greene, 1989, p. 651). Due to a feeling of hopelessness and worthlessness, people tend to believe that their actions are barely effective for solving serious life problems and their suffering will never end (Beck, Steer,

Kovacs, & Garrison, 1985). When such negative thoughts linger for a long time, people might develop depression and perhaps even start to view suicide as the only way out of their current situation (Beck et al., 1985). Though hopelessness is a core characteristic of depression, additional symptoms are required for making a diagnosis based on the DSM-5 criteria. Consequently, hopelessness, as a narrower depression-related construct, is expected to be modestly correlated with depression.

Empirical studies using Chinese samples showed that people with higher levels of depression tended to score higher on hopelessness scales, with correlation coefficients in the 0.44 to 0.51 range (Wu, Chen, Yu, Duan, & Jiang, 2015; Yang et al., 2015). In order to differentiate these two constructs, the Beck Hopelessness Scale (BHS) is also considered as a comparator instrument for the adapted HDS-OA. It is the most frequently used measure of hopelessness and demonstrates satisfactory reliability and validity evidence in China (Kong, Zhang, Jia, & Zhou, 2007; Liu et al., 2011).

Lastly, existing studies have suggested an association between depression and health-related quality of life (Yu et al., 2012). People who are depressed are at higher risk of having stroke, diabetes and chronic diseases, because they are inclined to have an unhealthy lifestyle, such as eating poorly, exercising less, and getting less sleep (APA, 2013; Canadian Mental Health Association, n.d.). Yu et al.'s (2012) research identified that people with a higher level of depression did report a poorer physical health status, although the correlation was not very strong (r = -0.27). Comparatively, the correlation between measures of depression and mental health status was much higher (r = -0.60, Yu et al., 2012; r = -0.75, Chin et al., 2015). It is worth noting that health-related quality of life (regardless of mental or physical health) are theoretically

inversely related to depression, so the results of the studies showed negative correlation coefficients.

As convergent and discriminant validity are the two ends of a continuum, it would be better to have values on both sides as relations to external variables evidence of the adapted HDS-OA. Therefore, the 12-Item Short Form Health Survey (SF-12) is another good choice as a comparator instrument. This scale provides us with composite scores of mental and physical functioning separately, which allows us to obtain coefficients that are close to the ends of convergent as well as discriminant validity.

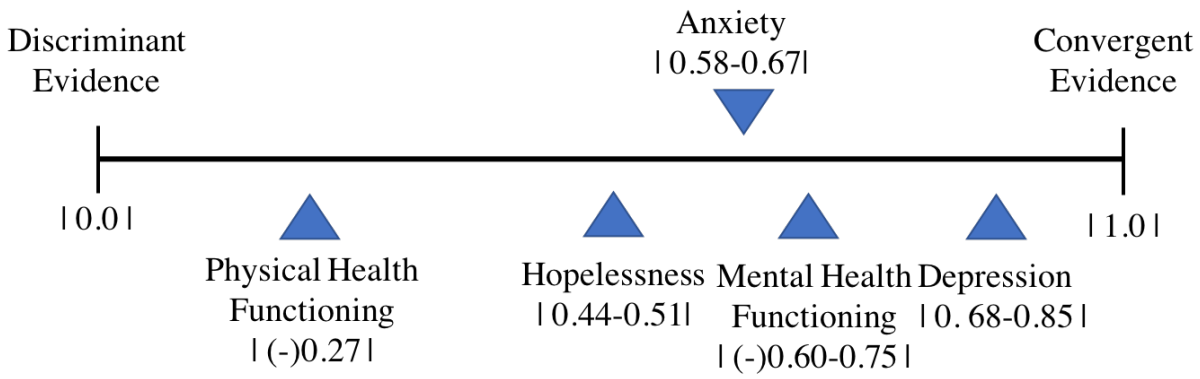
In general, this study is a cross-sectional survey; namely, the data are collected at a single time point from a sample drawn from a specified population. In order to collect evidence regarding relationships with external variables, four constructs are measured with four scales. The construct of interest is depression, which refers to symptoms such as feeling sad, losing interest in previously enjoyable activities, and being unable to carry out daily activities for a period of at least two weeks (World Health Organization, 2018). In this study, it will be assessed by the HDS-OA as the primary measure of interest and the PHQ-9, from which convergent validity evidence is obtained.

To examine whether the adapted HDS-OA can discriminate depression from related but conceptually distinct constructs, the GAD-7, BHS, and SF-12 are selected as comparator instruments. The GAD-7 is used for measuring anxiety level, with higher scores indicating that an individual has been bothered by the problems of "feeling nervous, anxious, or on edge" (Spitzer, Kroenke, Williams, & Löwe, 2006, p. 1094). The BHS is designed to quantify hopelessness, in terms of three major aspects: feelings about the future, loss of motivation, and future expectations (Beck, Weissman, Lester, & Trexler, 1974). The SF-12 is a self-report

questionnaire to measure perceived mental and physical health. Because it is self-rated but not assessing health problems with objective medical devices or health physicians, it reflects one's subjective experience of health status. If scores of the HDS-OA are associated with scores of the PHQ-9, GAD-7, BHS and SF-12 (mental health and physical health) in the expected manner (see Figure 2.3), then we have sufficient convergent and discriminant evidence to support the interpretation and use of the HDS-OA scores as reflecting level of depressive symptomatology.

Figure 2.3.

The continuum of convergent and discriminant validity.



Note. The values are correlations observed from previous studies and the (-) sign means that the scores were negatively related to depression.

Chapter 3: Manuscript

Note. This chapter is written as a free-standing manuscript and thus there is some redundancy with the Literature review and Conclusion chapters of the thesis.

An Adaptation and Validation Study of the Hubley Depression Scale for Older Adults (HDS-OA) in the General Adult Population in China

Introduction

Depression has become the single largest cause of disability worldwide (7.5%) and a major contributor to the overall global burden of disease, due to its detrimental effects on physical and mental health (World Health Organization, 2018). It not only leads to impairments of cognitive functioning, such as diminished ability for visuo-spatial processing and deficits in executive function, but also results in impairments of social functioning, such as dysfunctions in interpersonal interactions and decreased academic or work performance (Kupferberg, Bicks, & Hasler, 2016; Lam, Kennedy, McIntyre, & Khullar, 2014).

According to the widely accepted *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*; (American Psychiatric Association [APA], 2013)), major depressive disorder (MDD) is diagnosed, in part, when at least five of the following symptoms are present for nearly every day during the same two-week period: (1) depressed mood, (2) loss of interest in most activities, (3) significant change in weight (5% or more) or appetite, (4) insomnia or hypersomnia, (5) psychomotor agitation or retardation, (6) fatigue or loss of energy, (7) feeling worthless or excessively or inappropriately guilty, (8) diminished ability to think or concentrate or make decisions, and (9) thoughts of death or suicide or having a suicide plan. These nine symptoms can be categorized into two dimensions: somatic complaints and affective or cognitive disturbance. Somatic complaints include symptoms (3) to (6), while affective or

cognitive disturbance comprises symptoms (1), (2), (7), (8) and (9) (Tylee & Gandhi, 2005). Because somatic symptoms of MDD may overlap with the symptoms of other disorders, such as anxiety disorders, chronic pain, and dementia, either of the core symptoms of depression, namely, depressed mood and loss of interest, must be present when making a definitive diagnosis of MDD (APA, 2013; Ellis, Robinson, & Crawford, 2006; Kapfhammer, 2006).

In addition to the *DSM-5*, many clinicians in China also adopt a local diagnostic system called the *Chinese Classification of Mental Disorders Version 3* (*CCMD-3*). The first version of the *CCMD* was published in 1985 and the latest version, *CCMD-3*, was released in 2001 (Chinese Society of Psychiatry, 2001). It matches the fourth edition of the *DSM* (*DSM-IV*) and another popular classification standard, the *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (*ICD-10*), to a great extent because its descriptive definitions and diagnostic criteria refer to these two international systems (Dai et al., 2014). At the same time, locally salient features were also included to ensure the adherence to etiology and pathology in the Chinese context; hence, it is widely accepted by psychiatrists throughout the country (Chen, 2002).

The symptoms listed in the *CCMD-3* for diagnosis of MDD are almost identical to those in the *DSM-5*, except that the symptom (8) in the *DSM-5* is rephrased as "having difficulties in making associations and diminished ability to think" and "reduced libido" is added as a depressive symptom. The greatest difference between the *CCMD-3* and *DSM-5* is that the Chinese standard requires depressed mood and four or more of the other nine symptoms to be present for at least two weeks; namely, loss of interest is not a core symptom of depression in the *CCMD-3*. Still, the way that it defines depression does not seem to deviate significantly from the

most popular international classifications, which implies that the symptoms of depression are more or less consistent across cultures.

The lifetime prevalence of depression in 18 countries across five continents ranged from 6.5% (China) to 21.0% (France), with an average of 12.7% in low- to middle-income countries and 15.2% in high income countries (Kessler & Bromet, 2013). Although prevalence rates of depression in Western countries are much higher than that in China (Kessler & Bromet, 2013), statistics from the World Health Organization indicate that the suicide rate in China (8.0 per 100, 000 population) is quite close to those in Western countries (e.g., Netherlands: 9.6, New Zealand: 11.6, France: 12.1, United States: 13.7, per 100, 000 population) ("Suicide rate estimates", 2018). Given that depression is reported to be a main risk factor for suicide attempts and committed suicide in China, it can be inferred that many of these cases have failed to be diagnosed with depression (Liu, Contreras, Muñoz, & Leykin, 2014).

One explanation for the low prevalence of depression is that language has a direct impact on subjective psychopathological experiences, which results in different manifestations of depressive symptoms across cultures (Zhou, 2012). That is, a lack of vocabulary for affective states in the Chinese language increases difficulties in verbalizing inner feelings; hence Chinese people tend to articulate depression physically when they are asked to express their feelings (Kleinman, 1982; Zhu & Wang, 2011). However, when a scale was used to measure difficulties with clearly identifying and describing emotional states, Chinese and Euro-Canadian psychiatric outpatients did not show significant differences in scores, indicating that the variations in languages are hardly responsible for the tendency to report somatic symptoms (Ryder et al., 2008).

Another explanation is that people with depression in China are inclined to report bodily sensations, and thus are often mistakenly diagnosed with neurasthenia, which is determined by predominantly somatic symptoms such as physical fatigue, tension headache, and sleep disturbance (Kleinman, 1982; Kleinman, 1986; Lee, Kim, & Cho, 2017). The ratio between Chinese psychiatric patients who were diagnosed with depression and those identified as having neurasthenia was 1:30 (Kleinman, 1986). This explains why depression seems to be much less common than neurasthenia in China, unlike many Western countries (He, 2013; Zhou, 2012). Notably though, when reassessing 100 neurasthenia patients in China, it appeared that 87% of them should have been diagnosed with some kind of depressive disorder (Kleinman, 1982). Still different reasons have been proposed to account for the tendency to report somatic symptoms in China. Nakao and Barsky (2007) held the view that psychiatric patients experienced bodily sensations as being particularly intense and disturbing (an effect known as 'somatosensory amplification') due to the effect of psychological distress on their perception. However, the reporting of multiple unexplained somatic symptoms was shown to be common in a sample of 1146 patients with major depression in 14 countries (including China) and there were no significant variations in prevalence across countries (Simon, VonKorff, Piccinelli, Fullerton, & Ormel, 1999). This suggested a consistency in the physical conditions of people with depression in various cultures and thus, somatosensory amplification does not seem to account for the tendency for somatization in the Chinese group.

Ryder et al. (2008, p. 302) argued that Chinese people attempted to camouflage their psychological problems as physical disorders to "inhabit the sick role in their societies without bearing the burden of stigma", as mental illnesses have long been stigmatized in the Chinese culture. However, Simon et al. (1999) showed that the proportions of patients rejecting the two

most notable psychological symptoms (i.e., depressed mood and feelings of guilt or worthlessness) did not vary significantly among the sites and the overall prevalence was only 11%. Therefore, stigmatization of mental illness may not serve as the main cause of the somatization tendency.

Lastly, Yen, Robins and Lin (2000) described the tendency for somatization as a socially efficacious way to obtain health care resources in Chinese settings. Due to the shortage of psychiatrists and the unpopularity of psychotherapy, Chinese patients with depression often have to see a physician who does not specialize in psychiatry and emphasize their physical complaints in order to get medical treatment as soon as possible (Mao, 2013). Simon et al.'s (1999) research found that the tendency of reporting only somatic symptoms as the reason for seeking help was significantly more common in countries that offered walk-in care (e.g., China: 87%) than countries which offered a more personal form of primary care (e.g., France: 45%). Generally, such evidence suggests that the tendency to report somatic symptoms is not due to an inability to describe affective states, a different experience of physical conditions, or an unwillingness to express psychological distress, but a culturally effective presentation mode to seek help from health services in the Chinese context. Hence, even though people with depression in China are inclined to express depression somatically compared with those in Western countries, it appears that the construct of depression itself remains consistent across cultures.

Given that the diagnosis and experience of depression appears fairly consistent between China and other cultures, several commonly used Western depression screens have been translated into Chinese versions, including the Zung Self-rating Depression Scale (SDS), Center for Epidemiologic Studies Depression Scale (CES-D), and Beck Depression Inventory-II (BDI-II). Unfortunately, none of these measures have shown consistently strong psychometric

properties with Chinese samples. The Chinese SDS is the most widely used depression scale in China (Jin & Zhang, 2017; Yan, Xiao, & Hu, 2016), as it is one of the earliest introduced depression inventories in the country (Wang & Chi, 1984). However, Peng et al. (2013) reported Cronbach's alpha coefficients for four subscales as .35, .47, .60, and .74, and .78 overall in a sample of middle-aged women, which would be considered unacceptable, poor, questionable, and acceptable, respectively (George & Mallery, 2003). Lee et al.'s (1994) study that yielded a two-factor structure only reported an overall Cronbach's alpha of .91 in a sample of older adults. In addition, the area under the curve (AUC < 0.70) and specificity (36%) did not provide particularly strong test-criterion evidence for validity with a general community sample (Duan & Sheng, 2012).

The CES-D has been one of the most validated depression scales in China (Sun, Li, Yu, & Li, 2017). The psychometric studies indicated good internal consistency, with Cronbach's alpha between .85 and .90, in three samples (He et al., 2013; Zhang & Li, 2011; Zhang et al., 2015). But the coefficients of test-retest reliability (2 weeks, .91, Chin et al., 2015; 2-4 weeks, .64, Zhang et al., 2015) and convergent evidence for validity (r = .78 with the Patient Health Questionnaire-9, Chin et al., 2015; r = .61 with the BDI-II, Zhang & Li, 2011) have shown inconsistent results in studies using different age groups, and test-criterion validity evidence was barely acceptable (AUC = 0.75, Chin et al., 2015).

The BDI-II yielded more than one factor in both adolescents and in a general community sample, but only overall internal consistency coefficients were reported in these studies (.94, Wang et al., 2011; .89, Yang et al., 2014). This was problematic because a multidimensional structure negates the use of a total score or an alpha based on a total score. Convergent evidence for validity was not particularly strong (r = .67 with the Hamilton Rating Scale for Depression,

Wang et al., 2011; r = .76 with the CES-D, Yang et al., 2014). Besides, the BDI-II consists of 21 items and each item has a set of 4 statements, which means respondents read 84 statements in total. This response format increases the cognitive load and requires more administration time, leading to higher refusal and drop-out rates (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993).

A relatively new depression scale that might be promising for use in China is the Hubley Depression Scale for Older Adults (HDS-OA; Hubley, 1998). The HDS-OA has 16 items. The items are consistent with the *DSM-5* symptom criteria for MDD and persistent depressive disorder (PDD; also known as dysthymia) (Hubley, 1998). The HDS-OA items include cognitive, affective and somatic symptoms of depression. This relatively shorter measure was designed for use with older adults, and thus uses a yes-no response format, larger front size, and a reminder of the two-week period at the beginning of each item (Hubley, 2014). Importantly, however, the symptoms of depression do not differ for adults and older adults in diagnostic systems and thus this measure may be appropriate for adults of all ages and particularly useful when the sample includes individuals across the adult age range and with differing educational levels and cognitive abilities.

The HDS-OA has been found to be unidimensional (Hubley, Rajlic, & Zumbo, 2017) and has exhibited satisfactory internal consistency reliability in two mixed samples of depressed and non-depressed adults, with Cronbach's alphas ranging from .88 to .94 (Hubley et al., 2009; Myers & Hubley, 2012). Validity evidence was supported by strong correlations between scores on the HDS-OA with scores on the 30-item Geriatric Depression Scale (GDS) (r = 0.89) and a shorter 15-item version of the GDS (r = 0.86), a moderately strong but lower correlation with a measure of anxiety (r = 0.67), and moderate correlations with measures of physical health (r = -

0.43) and mental status (r = -0.39) (Myers & Hubley, 2012). Both studies also yielded high sensitivity (93%, Hubley et al., 2009; 92%, Myers & Hubley, 2012) and specificity (88%, Hubley et al., 2009; 100%, Myers & Hubley, 2012), although they supported different cut-off scores with different clinical samples. Thus far, the HDS-OA appears to be an effective case-finding tool for depression with satisfactory psychometric properties and a relatively short length that can reduce the burden on respondents.

Two studies are presented here. In Study 1, the English version of the HDS-OA was adapted into Chinese with the assistance of seven bilingual translators using Sousa and Rojjanasrirat's (2011) guidelines for cross-cultural translation, which include backward and forward translation. The translation procedures in their guidelines are aligned with the widely accepted *ITC Guidelines for Translating and Adapting Tests* developed by the International Test Commission (ITC, 2017). Still following Sousa and Rojjanasrirat's (2011) guidelines, Study 1 also included pilot tests conducted with a review panel of experts and a review panel from the target population of Chinese adults. In Study 2, the psychometric properties of the Chinese version of the HDS-OA were evaluated with a nonclinical sample from the general population in China, examining internal (factor) structure, internal consistency reliability, and convergent and discriminant validity.

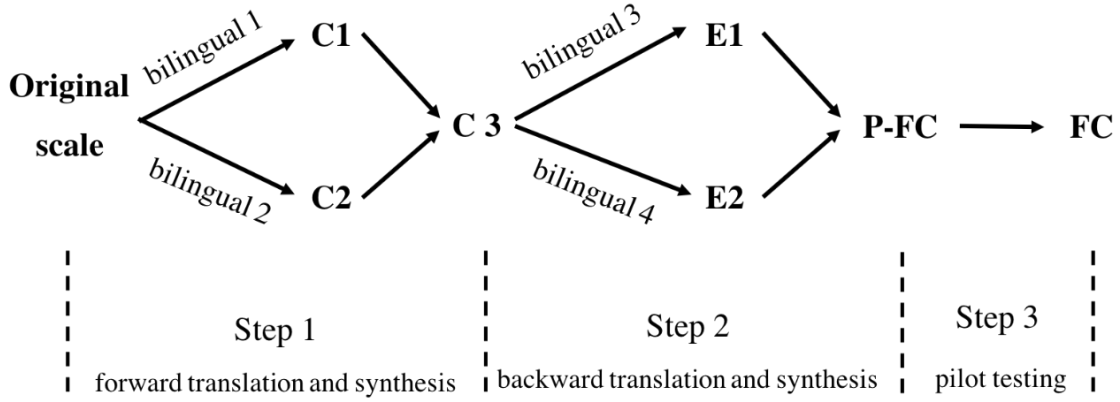
Study 1: Adaptation of the HDS-OA into Chinese

Method and Recruitment of Translators and Review Panel Members

The adaptation processes of the HDS-OA were summarized as three steps: (1) forward translation and synthesis, (2) backward translation and synthesis, and (3) pilot testing of the pre-final version (see Figure 3.1).

Figure 3.1.

Sousa and Rojjanasrirat's Translation Guideline.



Note. C1 and C2 are the target language (Chinese) versions. E1 and E2 are the source language (English) versions. P-FC is the pre-final Chinese version. FC is the final Chinese version.

Step 1: Forward translation and synthesis. The HDS-OA was adapted from English to Chinese by two translators who were fluent in both English and Chinese. We aimed to have two translators who were preferably Chinese native speakers with different backgrounds: one should be at least halfway through or already have a graduate-level degree in clinical/counselling psychology and be knowledgeable about the content area of depression, for the purpose of maximizing conceptual and measurement equivalence. According to previous cross-cultural translation studies, graduate-level training in this field is sufficient for being a translator of depression measures (Guo, Wang, & Chen, 2009; Liu, 2010; Papasavvas et al., 2016). We aimed for the other person to be a skilled translator, in order to provide equivalence from a more linguistic and cultural perspective. The inclusion criteria for this translator was: (1) having a bachelor's degree or higher in English language and literature or interpretation and translation studies; (2) either having passed one of the most authoritative and widely accepted English certificate tests (i.e., the Test for English Majors-band 8, TEM-8) in China or having the China Accreditation Test for Translators and Interpreters (CATTI) certificate, as an indicator of their

fluency of English and Chinese (Jin & Fan, 2011); and (3) having over 3 years of experience in translation.

Each translator translated the HDS-OA from English to Chinese independently. This way, two Chinese versions (C1 and C2) were obtained. Afterwards, the two translators and the first author compared the C1 and C2 versions along with the original English version regarding the instructions, items, and response format, in order to resolve any ambiguities or discrepancies and arrive at a consensus on the most culturally accurate and understandable adaptation. Thus, a synthesized Chinese version (C3) was generated.

Step 2: Backward translation and synthesis. In this step, the C3 version was blindly and independently back translated into English by two new translators, who did not have access to the English text. These translators would be fluent in both Chinese and English and preferably English native speakers[1]. These translators would also have different backgrounds as described above for the forward translation. This step produced two new English versions of the HDS-OA (E1 and E2). The translators and first author, as well as the test developer of the HDS-OA (second author), discussed any ambiguities or discrepancies between both the E1 and E2 versions and the original English version. Any item that did not retain the same meaning as the original item would be re-adapted and step 1 and 2 would be repeated to adapt these items before obtaining the pre-final Chinese version of the HDS-OA (C-HDS-OA).

Step 3: Pilot testing of the pre-final version. A pilot study on the pre-final version of the C-HDS-OA was carried out to do a final check and solicit feedback to improve the scale. We aimed to recruit 3-5 bilingual experts to form a review panel, with a goal of including experts in language (e.g., professional translators), methods (e.g., methodologists with experience in

[1] Unfortunately, we were unable to recruit English native speakers who were fluent in Chinese.

developing and translating instruments), and content (e.g., mental health professionals) (Ohrbach, Bjorner, Jezewski, John& Lobbezoo, 2013; Squires et al., 2013). Eligible candidates included individuals who were at least halfway through or already have had a graduate-level degree (i.e., master's or doctoral degree). Because the committee needed to review all of the translated versions as well as the original version of the HDS-OA, being a bilingual was required (Sousa & Rojjanasrirat, 2011). As it has been recommended that the original developer(s) of the measure be included as well during the pre-testing process (Beaton, Bombardier, Guillemin, & Ferraz, 2000; Tsang, Royse, & Terkawi, 2017), both authors discussed these results. Despite the inclusion criteria of expertise in different areas, the number of experts in such committees usually ranges from N = 1-5 in most current instrument translation studies (which sometimes includes the researcher) (Aqeel, Jami, & Ahmed, 2017; Gómez-Lugo et al., 2016; Hajebi et al., 2018; Li, Liu, Zhang, Wang, & Chen, 2011; Oishi et al., 2017; Papasavvas, Al-Amin, Ghabrash, & Micklewright, 2016). The experts were recruited online using convenience sampling and snowball sampling strategies.

The original English version of the HDS-OA, the pre-final Chinese version (C-HDS-OA), and the backward-translated English versions (E1 and E2) were provided to the panel members. After reviewing the English and Chinese statements, they rated each item using a question list developed based on Hambleton and Zenisky's (2010) Review Form, which included 25 questions around five broad topics: General Translation Questions, Item Format and Appearance, Grammar and Phrasing, Passages and Other Item-Relevant Stimulus Materials (if relevant), and Cultural Relevance or Specificity. They recommended that these translation questions should be answered for each of the items in the measures, which would make it a tedious and time-consuming task as the HDS-OA has 18 items in total. Therefore, we extracted

the main ideas from these questions and created tables under five new themes: meaning, difficulty, familiarity, cultural specificity, and format and appearance. Using these tables, the experts compared the different translated versions of the HDS-OA to the original version at the item level in terms of the first four areas below as well as overall in the fifth area below:

(1) Meaning: Experts rated the extent to which they thought each item in the Chinese version and the two backward-translated English versions have the same meaning as the original HDS-OA. In total, there were three tables for the comparison of meaning.

(2) Difficulty: Experts considered if there was anything in the Chinese adaptation (e.g., omissions, substitutions as well as additions, changes in punctuation, or modifications of the item's structure) of each item that may make it easier or more difficult to admit to negative experiences or feelings (e.g., item 1: *not* "feeling useful and needed" or item 4: "feeling sad and downhearted") than in the original English HDS-OA.

(3) Familiarity: When comparing the Chinese statements to the English statements, experts rated whether each item was equally familiar to respondents regarding the use of metaphors, idioms and colloquialisms, the concepts or constructs referred to in the item, and the grammatical structure.

(4) Cultural Specificity: For each item, experts examined if there were any Chinese cultural differences that would impact the likelihood of a response being chosen when the item was presented in the translated version, especially any phrasing or content that would be perceived as demeaning, offensive, controversial, or inflammatory.

(5) Format and Appearance (overall): Experts also compared the Chinese HDS-OA to the original HDS-OA in terms of the whole test. Specifically, the item format and appearance, including the physical layout, response format, item length, response option length and the use of phrase emphasis (bold, italics, etc.) were considered.

Altogether, the experts needed to complete seven tables. Either a 3-point response scale or a dichotomous response scale was used. Meanwhile, their feedback and suggestions about the Chinese version were collected and incorporated into the C-HDS-OA version.

In addition to the expert panel, we aimed to recruit 15-20 monolingual or bilingual Chinese people by online advertising to establish a participant panel. A sample size in the range of 5-8 (Wild et al., 2005), 5-15 (Beatty & Willis, 2007), 6-8 (Ohrbach et al., 2013;), 8-15 (Streiner, Norman, & Cairney, 2015), or 10-40 (Sousa & Rojjanasrirat, 2011) was believed to be adequate to uncover the most serious problems of the measure and evaluate the quality of the items. Hence, we aimed to have a minimum of 15 participants across gender, and different adult age groups (young adults ages 18-35 years, middle-aged adults ages 36-55 years, older adults ages 55 years and over). While it is recommended that the pre-test sample be monolingual with Chinese as their native language (Kalfoss, 2019; Sousa & Rojjanasrirat, 2011), we expected that the young groups would inevitably comprise bilingual Chinese, because English has been a compulsory subject for Chinese students from grade 3 of primary school to university since 2003 (Qi, 2016). It was expected that both the middle-aged and elderly groups would mainly consist of monolingual Chinese.

Based on Sousa and Rojjanasrirat's (2011) guidelines, another question list was developed for this panel. Specifically, the participants were asked to rate whether the items, instructions and response format and options were clear or unclear and provide suggestions if

there were any, after completing the C-HDS-OA. Instructions and items rated as unclear by over 20% of the participants would be revised based on the comments.

Results

The results were discussed in terms of each of the translation processes, followed by the pre-test. Notably, some of the steps were iterated twice to reach a satisfactory version of the C-HDS-OA. Seven translators were used in total.

Step 1: Forward Translation and Synthesis

Two bilingual graduate students in educational psychology prepared the first draft adaptation of the HDS-OA independently. Originally from Mainland China, both forward translator 1 (FT1) and forward translator 2 (FT2) had spoken Chinese for more than 20 years and had been learning English for more than 15 years. FT1 had a minor in English translation and had been certified as a translator by the National Advisory Committee for Foreign Language Teaching in China. FT2 had been living in Canada for six years and had been an instructor of Chinese at a university in eastern Canada for two years. FT1 and FT2 adapted the items of the HDS-OA from English to Chinese. After obtaining the two Chinese versions (Table 3.1), the first author had a discussion with the two translators to identify any unclear or confusing items or instructions.

While establishing a synthesized version of the C-HDS-OA, the two most problematic items were Item 8: *"Over the past two weeks, have you felt that you (or others) would be better off if you were dead?"* and Item 12: *"Over the past two weeks, have you enjoyed doing things as much as ever?"*. For Item 8, there is no direct translation of *better off* in Chinese, so both translators translated it as *"living a better life"* (i.e., "过得更好"). Thus, Item 8 became "*have you felt that you (or others) would live a better life if you were dead?*" in Chinese. This

Table 3.1.

Forward Translation of the Chinese Hubley Depression Scale for Older Adults (C-HDS-OA)

Original English Version	Forward Translation 1	Forward Translation 2	Synthesized Chinese Version
Instructions: The following questions have to do with changes that might have taken place in your life recently. For each question, please circle the answer (yes or no) that best applies to you.	指导语：请回忆您生活中最近可能发生的变化，以下问题均针对这些变化提出。请圈出您认为最符合您情况的答案（是或否）。	指导语：下列问题与您最近生活中发生的变化相关。对于各个问题，请圈出最适用您自身情况的回答（是或不是）。	指导语：以下问题均关于你生活中最近可能有发生过的变化。请圈出你认为最符合你情况的答案（是或否）。
1. Over the past two weeks, have you felt useful and needed?	在过去的两周中，您是否感觉自己有用或是被需要?	在过去两周内，你有感受到自己有用，有被需要吗?	在过去的两周内，你有没有觉得自己有用和被需要?
2. Over the past two weeks, have you noticed any changes in your appetite? (examples: you didn't feel like eating or you felt hungrier than usual)	在过去的两周中，您是否感觉您的胃口有什么变化？（例如：感觉自己不想吃东西，或是觉得比以往更加饥饿）	在过去两周内，你有发觉自己的食欲发生了什么变化吗?（例如：你不怎么想吃东西，或者你比往常感觉更饥饿）	在过去的两周内，你有没有发现你的食欲有什么变化？（例如：感觉自己不想吃东西，或是觉得比往常更饿）
3. Over the past two weeks, have you felt full of energy?	在过去的两周中，您是否感觉到充满活力?	在过去两周内，你有感觉到充满精力吗?	在过去的两周内，你有没有感到精力充沛?
4. Over the past two weeks, have you often felt sad and downhearted?	在过去的两周中，您是否经常感到伤心难过、情绪低落?	在过去两周内，你经常感觉伤心、沮丧吗?	在过去的两周内，你有没有经常感到伤心难过、心情低落?
5. Over the past two weeks, have your sleeping patterns changed? (examples: you have been waking up in the middle of the night or unusually early)	在过去的两周中，您在睡眠上是否有较大变化？（例如：突然开始在夜晚醒来，或是与往常相比醒得更早）	在过去两周内，你的睡眠规律有变化吗？（例如：你在半夜醒来或是异如往常醒得早）	在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是比往常早醒）
6. Over the past two weeks, have you had difficulty concentrating?	在过去的两周中，您是否更难集中注意力?	在过去两周内，你在集中注意力方面是否困难?	在过去的两周内，你有没有感觉难以集中注意力?
7. Over the past two weeks, have you been interested in your usual activities?	在过去的两周中，您是否仍有兴趣进行您的日常活动?	在过去两周内，你是否对你平常的活动仍有兴趣?	在过去的两周内，你还有没有兴趣进行你平常的活动?

Original English Version	Forward Translation 1	Forward Translation 2	Synthesized Chinese Version
8. Over the past two weeks, have you felt that you (or others) would be better off if you were dead?	在过去的两周中，您是否认为如果您死了，您（或者是其他人）会过得更好?	在过去两周内，你是否曾认为如果你自己死了，你（或他人）能过得更好?	在过去的两周内，你有没有觉得如果自己死了，对自己（或他人）都比较好?
9. Over the past two weeks, have you become irritated more easily than usual?	在过去的两周中，您是否比起往常更加易怒?	在过去两周内，你是否有变得比往常更易恼怒?	在过去的两周内，你有没有觉得比起往常更容易生气发火?
10. Over the past two weeks, have you felt different than you usually do? (examples: you felt unusually restless or you felt like you were moving in slow motion)	在过去的两周中，您是否感觉自己与往常有所不同? （例如：您感觉异常疲惫，或是动作迟缓）	在过去两周内，你是否有感觉到自己较之平常行为有异? （例如：你感觉非常焦躁或你感觉自己似乎在做慢动作）	在过去的两周内，你有没有觉得自己与往常有所不同? （例如：你感觉异常坐立不安，或是动作迟缓）
11. Over the past two weeks, has anyone mentioned to you that you don't look or seem your usual self?	在过去的两周中，是否有人向您提起，您好像与之前看上去不同了?	在过去两周内，是否有人向你提及你看上去不像是往常你自己的样子?	在过去的两周内，有没有人向你提起过，你看上去不像往常的自己?
12. Over the past two weeks, have you enjoyed doing things as much as ever?	在过去的两周中，您是否与之前一样享受生活的乐趣?	在过去两周内，你是否像以往一样同等程度地享受做事情?	在过去的两周内，你还能不能与之前一样从做事中得到乐趣?
13. Over the past two weeks, have you felt like everything was your fault?	在过去的两周中，您是否感到所有事都是您的错?	在过去两周内，你是否有感觉到似乎所有事情都是你的错?	在过去的两周内，你有没有觉得似乎所有事都是你的错?
14. Over the past two weeks, have you frequently felt like crying?	在过去的两周中，您是否常常想哭?	在过去两周内，你是否有经常感觉到想要哭泣?	在过去的两周内，你有没有常觉得想哭?
15. Over the past two weeks, have you found it harder than usual to make decisions?	在过去的两周中，您是否常感到与平常相比更加难以做决定?	在过去两周内，你是否发觉自己比往常更难做决定?	在过去的两周内，你有没有觉得比往常更难做决定?
16. Over the past two weeks, have you thought that the future looks hopeless?	在过去的两周中，您是否对未来感到绝望?	在过去两周内，你是否有想过未来看似无望?	在过去的两周内，你有没有觉得未来看起来没有希望?
(A) Have you started taking a new medication in the past month?	您是否在过去的一个月中开始服用之前未曾服用的药物?	你是否在最近一个月开始服用新的药品?	最近的一个月内你有没有开始服用新的药物?
(B) Are you grieving for someone who has died in the past two months?	您是否在过去的两个月里经历了某人的离世?	你是否在哀悼一位最近两个月内去世的人?	你有没有因为有人在近两个月内离世而处于哀痛之中?

adaptation seemed to be logically contradictory, because the phrase *"living a better life"* is used to describe a scenario where someone is alive, and accordingly, is in conflict with another part of the item *"if you were dead"*. For those respondents who do not believe in life after death, it makes no sense to talk about living a life if they were dead already. To resolve this issue, Item 8 was translated as "你有没有觉得如果自己死了，对自己（或他人）都比较好？" (i.e., *"have you felt that it would be better for you (or others) if you were dead?"*).

How the two translators understood the phrase *"enjoy doing things"* in Item 12 was different. FT1 adapted it as *"enjoy the pleasure of life"* (i.e., "享受生活的乐趣"), where *life* seemed to be more general than *doing things* and thus some extra interpretation was included. The translation by FT2 was somewhat more literal (i.e., "享受做事情"), which sounded neither natural in Chinese nor conveyed the meaning of the original English item. Because the translators and the first author could not come up with an appropriate and effective adaptation after discussion, a third forward translator (FT3) was added to improve the quality of the adaptation. Born in Mainland China and immigrating to Canada at the age of 7, FT3 grew up speaking English at school and Chinese at home, which made her fluent in both languages. She was a graduate student in counselling psychology.

Based on FT3's understanding, Item 12 described a main symptom of MDD called anhedonia, which refers to a loss of interest in previously enjoyable activities and a reduced ability to experience pleasure. In other words, it should not be seen as a warning sign of depression when people lose interest in activities that they did not used to enjoy. Hence, FT3 suggested inserting a modifying phrase to indicate the implied meaning of this item and translating it as *"get pleasure from doing things that you normally enjoy"* (i.e., "从喜欢的事情中得到乐趣"). Given that adding a modifier might make the adaptation less equivalent to the

original text, the researcher presented different renderings to the test developer of the HDS-OA (second author) and asked for advice. While agreeing with FT3's interpretation, the second author believed that it was more important to maintain equivalence between the HDS-OA and C-HDS-OA, so the FT3's suggestion was not followed and an alternative adaptation (i.e., "从做事中得到乐趣", *"get pleasure from doing things"*) was used instead.

FT3 was also asked to provide suggestions for the adaptations of the rest of the items. According to her feedback, the items were written in a formal style in Chinese while the English items were, in fact, more casual expressions. In order to maintain equivalence between different language versions, the wording of some Chinese items was changed without substantially changing the meaning. For example, Item 15 was slightly changed from "您是否感到过比往常更难做决定？" to "你有没有觉得比往常更难做决定？". Integrating feedback from FT3, a synthesized Chinese version of the HDS-OA was eventually developed (Table 3.1).

Step 2a: 1st Backward Translation and Synthesis

In this step, two new bilingual graduate students independently back translated the synthesized Chinese version into English. Both backward translator 1 (BT1) and backward translator 2 (BT2) were blind to the original English version. BT1 was a Chinese native speaker who lived in Mainland China for more than 20 years. She was also fluent in English as she had learned English for over 20 years and had been living in Canada for 9 years. She had a Master's degree in counselling psychology and had worked as a registered clinical counsellor for two years. In addition to her clinical background, two years of working on book translation projects on counselling and psychotherapy had also made BT1 knowledgeable about the content domain of depression. Born in Taiwan, BT2 had spoken Chinese for over 20 years. She began learning English in elementary school and had a bachelor's degree in English education.

The two back-translated versions (Table 3.2) were compared to the original English version by BT1, BT2, and both authors wherein ambiguities and discrepancies of meanings were marked. The team then reviewed and discussed these discrepancies, and they found that six items (Item 5, 7, 9, 10, 12, and 13) needed further revisions. Specifically, the identified problems can be categorized as the following three types: (1) varying levels of language intensity in different versions (Items 5 and 9); (2) distortion of the meaning of some words or phrases (Items 7 and 10); and (3) addition of extra meaning to the original text (Items 12 and 13).

In terms of the strength of words, the backward translated versions of Item 5 were less extreme than the original one, while those of Item 9 were more intense. For example, for Item 5 ("在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是比往常早醒)" *"Over the past two weeks, have your sleeping patterns changed? (examples: you have been waking up in the middle of the night or unusually early)"*), both BT1 and BT2 translated "比往常早醒" (i.e., "*unusually early"*) as *"earlier than usual"*. The problem here is that it would be considered *"unusually early"* when someone used to get up at 8:00 am but now gets up at 3:00 am, whereas *"earlier than usual"* is just getting up one or two hours earlier. In order to make the English and Chinese text at the same level of intensity, the adaptation was changed from "比往常早醒" to "异常地早醒". As for Item 9 ("在过去的两周内，你有没有觉得比起往常更容易生气发火)" *"Over the past two weeks, have you become irritated more easily than usual?"*), BT2 adapted "生气发火" (i.e., *"become irritated"*) as *"lose your temper"*. The former describes the feeling of slight anger and annoyance, while the latter means becoming very angry. To make it less extreme, the adaptation was revised as "烦躁".

Table 3.2.

Backward Translation of the Chinese Hubley Depression Scale for Older Adults (C-HDS-OA)

Original English version	Backward Translation 1	Backward Translation 2
Instructions: The following questions have to do with changes that might have taken place in your life recently. For each question, please circle the answer (yes or no) that best applies to you.	Instructions: The following questions are all about possible changes in your life recently. Please circle the answer that you think best describes your situation (yes or no).	Instructions: the following questions are about changes that possibly have happened in your life recently. Please circle the answer that best describes your situation (Yes/No).
1. Over the past two weeks, have you felt useful and needed?	Over the past two weeks, have you felt that you are useful and needed?	In the past two weeks, have you felt yourself being useful and needed?
2. Over the past two weeks, have you noticed any changes in your appetite? (examples: you didn't feel like eating or you felt hungrier than usual)	Over the past two weeks, have you noticed any changes in your appetite (for example, not feel like eating or feel hungrier than usual)?	In the past two weeks, have you found any difference in your appetite (e.g. didn't want to eat or felt hungrier than usual)?
3. Over the past two weeks, have you felt full of energy?	Over the past two weeks, have you felt energetic?	In the past two weeks, have you felt energetic?
4. Over the past two weeks, have you often felt sad and downhearted?	Over the past two weeks, have you often felt sad or low-spirited?	In the past two weeks, have you often felt sad and upset (depressed at first)?
5. Over the past two weeks, have your sleeping patterns changed? (examples: you have been waking up in the middle of the night or unusually early)	Over the past two weeks, have your sleep patterns changed (for example, you would wake up in the middle of the night or wake up earlier than usual)?	In the past two weeks, have there been any changes to your sleeping cycle (e.g. you would wake up in the middle of the night or wake up earlier than usual)?
6. Over the past two weeks, have you had difficulty concentrating?	Over the past two weeks, have you found it hard to focus?	In the past two weeks, have you felt hard to concentrate your mind?
7. Over the past two weeks, have you been interested in your usual activities?	Over the past two weeks, have you had interest in your usual activities?	In the past two weeks, have you still been interested in your normal routines?

Original English version	Backward Translation 1	Backward Translation 2
8. Over the past two weeks, have you felt that you (or others) would be better off if you were dead?	Over the past two weeks, have you felt you (or others) would be better off if you were dead?	In the past two weeks, have you felt that if you were dead, it would be better for yourself (or others)?
9. Over the past two weeks, have you become irritated more easily than usual?	Over the past two weeks, have you found it easier to be irritated than usual?	In the past two weeks, have you felt easier to lose your temper than usual?
10. Over the past two weeks, have you felt different than you usually do? (examples: you felt unusually restless or you felt like you were moving in slow motion)	Over the past two weeks, have you felt that you are different than usual (for example, you feel particularly unsettled or sluggish)?	In the past two weeks, have you felt yourself different than usual (e.g. you would feel unexpectedly anxious or acting slow)?
11. Over the past two weeks, has anyone mentioned to you that you don't look or seem your usual self?	Over the past two weeks, has anyone mentioned to you that you don't seem your usual self?	In the past two weeks, has anyone mentioned to you that you didn't look like yourself?
12. Over the past two weeks, have you enjoyed doing things as much as ever?	Over the past two weeks, have you been getting pleasure from doing things as usual?	In the past two weeks, have you still been able to find pleasure from things as you did before?
13. Over the past two weeks, have you felt like everything was your fault?	Over the past two weeks, have you felt that everything seems to be your fault?	In the past two weeks, have you felt that it seemed like everything is your fault?
14. Over the past two weeks, have you frequently felt like crying?	Over the past two weeks, have you often felt like crying?	In the past two weeks, have you frequently felt wanting to cry?
15. Over the past two weeks, have you found it harder than usual to make decisions?	Over the past two weeks, have you found it more difficult to make decisions than usual?	In the past two weeks, have you felt harder to make decisions than usual?
16. Over the past two weeks, have you thought that the future looks hopeless?	Over the past two weeks, have you felt that the future seems hopeless?	In the past two weeks, have you felt your future seemed hopeless?
(A) Have you started taking a new medication in the past month?	Did you start to take any new medication over the past month?	Have you started to take new drugs in the past one month?
(B) Are you grieving for someone who has died in the past two months?	Are you grieving for someone's death over the past two months?	Have you been grieving due to anyone passing away in the past two months?

The second issue is that the meaning of some words or phrases was distorted in the back-translated versions. For instance, for Item 7 (“在过去的两周内，你还有没有兴趣进行你平常的活动？” *“Over the past two weeks, have you been interested in your usual activities?”*), BT2 interpreted “平常的活动” (i.e., *“usual activities”*) as *“normal routines”*, which basically refers to everyday activities, such as brushing teeth, taking a shower or having dinner. However, *“usual activities”* are more than basic activities of daily living. It could be visiting friends or relatives, going to church, playing sports, and so forth. In fact, the difference between these two concepts was already discussed during the meeting with the forward translators. To address this point, the synthesized Chinese version avoided the phrase “日常活动”, which means daily activities, and chose the phrase “平常的活动” instead. After comparing the synthesized Chinese version to the two forward translated versions, BT2 agreed that the current adaptation of *“usual activities”* was appropriate. Though no changes were made to Item 7, the team believed that a second backward translation was still necessary to ensure its quality.

The renderings of Item 10 (“在过去的两周内，你有没有觉得自己与往常有所不同？（例如：你感觉异常坐立不安，或是动作迟缓）” *“Over the past two weeks, have you felt different than you usually do? (examples: you felt unusually restless or you felt like you were moving in slow motion)”*) were also questionable. Specifically, “动作迟缓” (i.e., *“moving in slow motion”*) was translated as *“sluggish”* by BT1 and *“acting slow”* by BT2. According to the test developer (second author), the symptom of moving in slow motion means that people suffering from depression may feel like everything takes extra effort, they feel weighed down, or they are slowed down due to lack of energy. To be more specific, it is a state of being that manifests in people’s thoughts, sensations, and behaviours. In this sense, *“sluggish”* is close to

the original text, but *"acting slow"* is too specific, and thus is not an accurate adaptation. To emphasize this feeling as a result of lacking energy, the adaptation was changed from "动作迟缓" to "提不起劲做事" (i.e., *"lack the energy in doing things"*).

Another translation issue has to do with adding extra meaning. Although neither of the backward translations seemed to deviate from the original text, Item 12 ("在过去的两周内，你还能不能与之前一样从做事中得到乐趣?" *"Over the past two weeks, have you enjoyed doing things as much as ever?"*) was revised. As mentioned earlier in the forward translation step, FT3 suggested translating *"enjoy doing things"* as "从喜欢的事情中得到乐趣" (i.e., *"get pleasure from doing things that you normally enjoy"*). Both BT1 and BT2 agreed that embedding an adjective clause (i.e., *"that you normally enjoy"*) made this item clearer in Chinese and did not distort the original intent, though it seemingly added extra meaning. Given that the premise of measuring depression is the comprehensibility of the translated instrument, the authors eventually decided to trade off equivalence for sentence clarity and accepted FT3's adaptation. Another item that needed further revision was Item 13 ("在过去的两周内，你有没有觉得似乎所有事都是你的错？" *"Over the past two weeks, have you felt like everything was your fault?"*). Both backward translators used *"seem"* in their adaptations as a result of the Chinese word "似乎" (BT1: *"Everything seems to be your fault"*; BT2: *"It seemed like everything is your fault"*). Because the original English item does not include *"seem"*, "似乎" was deleted from the Chinese statement.

Step 2b: 2nd Backward translation and synthesis.

In step 2a, by comparing the backward translated versions to the original English scale, changes were made to six items (Items 5, 7, 9, 10, 12, and 13) in the C-HDS-OA. Accordingly, a

second backward translation was necessary to determine if the suggested revisions were appropriate. Two translators (i.e., BT3 and BT4) who were fluent in English and Chinese read the six newly revised Chinese statements and back translated them into English. BT3 was a Chinese native speaker who had been learning English for over 10 years and had a minor in English. He was a graduate student in special education. Originally from Taiwan, BT4 was a bilingual Chinese faculty member in education and had been living in Canada for over 10 years.

The revised Chinese version and new back-translated versions of the six items are summarized in Table 3.3. BT3, BT4, and both authors had a discussion to evaluate the equivalence of the different English versions and finalize the C-HDS-OA. Only two items (Items 5 and 7) were revised in this step. For Item 5 (“在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是异常地早醒)”） *“Over the past two weeks, have your sleeping patterns changed? (examples: you have been waking up in the middle of the night or unusually early)”*), both BT3 and BT4’s adaptations were close to the original statement. However, BT3 thought that “异常” (i.e., *“unusually”*) might not be a good choice of word as it might be mistakenly perceived as *“abnormal”* (i.e., another meaning of “异常” in Chinese), and thus how people answer it might be biased due to the stigma of mental illness. To solve the ambiguity, “异常早醒” was changed to “比往常早醒许多”. With respect to Item 7 (“在过去的两周内，你还有没有兴趣进行你平常的活动？” *“Over the past two weeks, have you been interested in your usual activities?”*), “平常的活动” (i.e., *“usual activities”*) was similarly depicted by BT2 and BT3 as daily routines, which distorted the meaning of the original text. Hence, the team decided to revise it as “平常喜欢的活动” (i.e., *“activities that you would normally enjoy”*) based on BT4’s adaptation, in order to distinguish it from everyday activities.

Table 3.3.

Second Backward Translation of Six Chinese Hubley Depression Scale for Older Adults (C-HDS-OA) Items

1st Revised Chinese version	Backward Translation 3	Backward Translation 4	2nd Revised Chinese version
5. 在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是异常地早醒）	In the past two weeks, have you experienced any change in your sleep pattern? (For example, you woke up in the midnight or extremely early)	In the past two weeks, has your sleep pattern changed? (For example, you woke up in the middle of the night or woke up unusually early)	在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是比平常早醒许多）
7. 在过去的两周内，你还有没有兴趣进行你平常的活动?	In the past two weeks, have you had any interest in proceeding your daily activities?	In the past two weeks, did you lose interest in activities that you would normally enjoy?	在过去的两周内，你还有没有兴趣进行你平常喜欢的活动?
9. 在过去的两周内，你有没有变得比起往常更容易烦躁?	In the past two weeks, have you been more easily dysphoric than usual?	In the past two weeks, have you become more irritable?	在过去的两周内，你有没有变得比起往常更容易烦躁?
10. 在过去的两周内，你有没有觉得自己与往常有所不同?（例如：你感觉异常坐立不安，或是提不起劲做事）	In the past two weeks, have you found yourself different from usual? (For example, you felt restless or had no energy to do anything.)	In the past two weeks, did you feel that you were a different person? (For example, you felt restless or lost interest in everything)	在过去的两周内，你有没有觉得自己与往常有所不同？（例如：你感觉异常坐立不安，或是提不起劲做事）
12. 在过去的两周内，你还能不能与之前一样从喜欢的事情中得到乐趣?	In the past two weeks, have you still had fun doing your preferred activities as usual?	In the past two weeks, were you still able to gain fun from doing things that you normally would enjoy to do?	在过去的两周内，你还能不能与之前一样从喜欢的事情中得到乐趣?
13. 在过去的两周内，你有没有觉得所有事都是你的错?	In the past two weeks, have you felt it was all your fault?	In the past two weeks, did you feel that everything is your fault?	在过去的两周内，你有没有觉得所有事都是你的错?

No changes were made to the rest of the items, as the back-translated statements were considered equivalent to the original version. Notably, for Item 10 ("在过去的两周内，你有没有觉得自己与往常有所不同？（例如：你感觉异常坐立不安，或是提不起劲做事"）*"Over the past two weeks, have you felt different than you usually do? (examples: you felt unusually restless or you felt like you were moving in slow motion)"*), BT4 translated "提不起劲做事" (i.e., *"lack the energy in doing things"*) as *"lost interest in everything"*, which seemed to be more extreme. However, BT4 mentioned that it was not because of the adaptation itself but his over-interpretation of this item, hence no revisions were needed. A reconciled version of the six items is presented in Table 3.3.

Step 3a: Pilot Testing of the Pre-final Version with an Expert Panel

In order to do a final check and collect feedback to improve the adaptation, an expert panel and a participant panel were recruited to rate the quality of the items. As suggested by Ohrbach et al. (2013) and Squires et al. (2013), the first panel consisted of bilingual experts in content (i.e., expert 1 and 2), methods (i.e., expert 3), as well as language (i.e., expert 4). Specifically, expert 1 (E1) and expert 2 (E2) were mental health professionals. E1 received both his bachelor's degree in Social Work and his master's degree in Criminology in Hong Kong. He had been working as a youth and family counsellor in the lower mainland of British Columbia for 22 years. E2 was a Master's student in counselling psychology. The third expert (E3) was a psychometrician from a local educational testing company with extensive experience in cross-cultural and large-scale assessments. She received her Ph.D. degree in the measurement field. The last expert (E4) was a professional translator with four years of experience. He received his Bachelor's degree in English and his Master's degree in Translation and Interpreting from a Chinese university, and had passed both the TEM-8 and CATTI.

Table 3.4 reports the extent to which the experts thought the items in the Chinese version had the same meaning, difficulty and familiarity as the English items and whether the adaptations showed satisfactory attention to cultural differences or issues. With regard to 'meaning', the expert panel was in full agreement with 12 of the translated items, which were Items 1, 2, 5, 6, 8, 9, 11, 12, 13, 14, 16 and (A). Notably, E3 recommended subtle changes of grammatical particles to Items 8 and 9, and E2 suggested adding a comma in the Item (A), in order to create a smoother reading experience for respondents. Their suggestions were adopted.

In most cases (12 of the translated items: Items 1, 2, 5, 6, 8, 9, 11, 12, 13, 14, 16 and (A)), the meaning of the C-HDS-OA and original English HDS-OA items were viewed as the same by all of the experts. For four items, (Items 3, 4, 15 and (B)), three out of the four experts felt that the meaning was the same in different versions and one was unsure. For Item 3, one of the panel members (E2) suggested that *"full of energy"* should be translated as "充满活力", which was exactly how FT1 translated it. Although this translation has the same meaning as the English phrase, it was ruled out because the translators in steps 1 and 2 agreed that people rarely identify themselves as being full of energy even for those who do not have depression. Instead, the current adaptation "精力充沛" (i.e., *"energetic"*) captures the intended meaning more than the literal translation "充满活力", and thus was kept. In terms of Item 4, E3 recommended the words "时常" or "常常" to replace "经常" (i.e., *"often"*) in this statement, because she felt that its meaning was stronger than *"often*". But, in fact, the difference among these three words is very subtle and "经常" is one of the most common translations of *"often"* in Chinese, so we decided to keep it as it is. E3 also suggested adding "自己" (i.e., *"yourself"*) in Item 15, as she felt that the subject was not explicitly stated here so that the meaning of this sentence was not perfectly clear. Given that it is grammatically correct in Chinese, and the other three experts did not

Table 3.4.

Expert Panel (N=4) Feedback on Meaning, Difficulty, Familiarity, and Cultural Specificity for the C-HDS-OA and Original English HDS-OA

Item	Meaning	Difficulty	Familiarity	Cultural Specificity
1	Same = 100%	Same = 100%	Same = 100%	Same = 100%
2	Same = 100%	Same = 100%	Same = 100%	Same = 100%
3	**Same = 75%; Unsure = 25%[a]**	Same = 100%	Same = 100%	Same = 100%
4	**Same = 75%; Unsure = 25%[b]**	**Same = 75%; More Difficult = 25% [b]**	Same = 100%	Same = 100%
5	Same = 100%	Same = 100%	Same = 100%	Same = 100%
6	Same = 100%	Same = 100%	Same = 100%	Same = 100%
7	**Same = 50%; Unsure = 25%[c]; Not the Same = 25%[d]**	**Same = 50%; More Difficult = 25%[c]; Easier = 25%[d]**	Same = 100%	Same = 100%
8	Same = 100%[e]	**Same = 75%; More Difficult = 25%[f]**	**Same = 75%;Not the Same = 25%[g]**	Same = 100%
9	Same = 100%[h]	Same = 100%	Same = 100%	Same = 100%
10	**Same = 25%; Unsure = 50%[i,j]; Not the same = 25%[k]**	**Same = 50%; More Difficult = 25%[i]; Easier = 25% [k]**	Same = 100%	**Same = 75%; Not the Same = 25% [i]**
11	Same = 100%	Same = 100%	Same = 100%	Same = 100%
12	Same = 100%	Same = 100%	Same = 100%	Same = 100%
13	Same = 100%	Same = 100%	Same = 100%	Same = 100%
14	Same = 100%	Same = 100%	Same = 100%	Same = 100%
15	**Same = 75%; Unsure = 25%[l]**	Same = 100%	Same = 100%	Same = 100%
16	Same = 100%	Same = 100%	Same = 100%	Same = 100%
(A)	Same = 100%[m]	Same = 100%	Same = 100%	Same = 100%
(B)	**Same = 75%; Unsure = 25%[n]**	**Same = 75%; More Difficult = 25% [n]**	Same = 100%	**Same = 75%; Not the Same = 25% [n]**

Note: Meaning: rated as Same, Unsure, or Not the Same; Difficulty: rated as Same, Easier or More Difficult; Familiarity rated as Same or Not the Same; Cultural specificity rated as Same or

Not the Same. Cases without Same = 100% are bolded. Feedback from experts on each item is identified below.

[a] E2: Perhaps consider 充满活力 (i.e., full of energy) instead?)

[b] E3: To me, 经常 means *frequently* and implies regularity, which makes its meaning stronger than *often* (i.e., many times in the past two weeks). 常常 or 时常 might be a better choice to resemble the meaning of *often*.

[c] E1: *Usual activities* doesn't refer only to 平常喜欢的活动 (i.e., activities you like to do). It can include 平常会做但并不一定喜欢的活动 (i.e., activities you normally do but necessarily like to do

[d] E3: To me*, usual activities* means daily routines but the meaning of 平常喜欢的活动 in the Chinese version is closer to favorite activities. Given the context, *interested in* might be better understood as the feeling of wanting to give attention to rather than the thought that something is interesting.

[e] E3: Maybe change 都 (in the sentence "对自己（或他人）*都*比较好") to 会. It doesn't change the meaning of the sentence but grammatically fits better.

[f] E2: Perhaps consider 走了(i.e., gone) or 消失了(i.e., disappeared). Those words might be gentler but implies death.

[g] E3: Based on my experience, in Chinese culture, people usually do not talk about death (i.e., use the word 死) directly and publicly. When it is mentioned, it's common to be referred using other words, such as *has gone*. So what is asked in this question might be something less familiar to Chinese respondents.

[h] E3: Could remove this word 起 from the sentence 变得比起往常更容易烦躁.

[i] E1: *Felt like you were moving in slow motion* is not translated literally. It can be translated as you did (i.e., 提不起劲做事) if there is understanding that it is used in a depression scale. If there is no prior knowledge of that, it can be translated as 生活节奏放慢.

[j] E3: To me, 提不起劲做事 emphasizes unmotivated rather than slow. I'm not sure what the focus of the original English version is. The translated version well reflects the general meaning of this example.

[k] E4: There are many ways to explain the phrase *you were felt like moving in slow motion*. It can be you felt like the days seem to go by slowly, or you felt tired or sluggish and didn't want to do anything. 提不起劲做某事 only reflects the latter explanation. Maybe translate it as 反应迟钝.

[l] E3: Consider adding a phrase and change the translation to 你有没有觉得[自己]比往常更难做[出]决定. The current translated version could be interpreted in different ways.

[m] E2: Consider a comma after 最近的一個月內.

[n] E1: It depends on the degree of grieving in the translation, it can be just 哀悼 and not necessarily be 处于哀痛之中.

interpret the adaptation in any other different ways without the subject being included, the current version was kept. Regarding Item (B), E1 translated *"grieving"* as FT2 did using the word "哀悼", which he believed denoted a more neutral tone than the current phrase "处于哀痛之中". However, "哀悼", which describes the temporary action of showing deep sorrow, is not common in everyday use, but mostly found in formal texts, such as government notices and press releases. Considering the present progressive tense in the original item, it makes more sense to keep the current rendering, which indicates the ongoing continuity of a state of feeling sorrow and is more widely used in daily life.

The most controversial adaptations involved Items 7 and 10. For Item 7, two of the experts agreed that the meaning was the same, one was unsure and the other one felt that the meaning was different. Due to the revision made previously, the experts rating 1 (*not the same*) and 2 (*unsure*) commented that the adaptation of *"usual activities"* (i.e., "平常喜欢的活动") represented pleasurable activities, to which the English phrase was not necessarily restricted. According to the test developer (second author), this phrase can actually be understood in either way in English, but there is no similar adaptation that indicates both enjoyable activities and daily routines. One possible solution is to include two phrases to cover both meanings, which would be somehow redundant in Chinese. In this case, using the current adaptation would be preferable, as we intend to distinguish *"usual activities"* from everyday activities.

Regarding Item 10, one expert felt that the meaning was the same, one chose the opposite, and the other two were unsure. It caused uncertainty and disagreement because there is no literal translation of *"moving in slow motion"* in Chinese. Even in English, this phrase can be explained in different ways. It not only describes the feeling of time dragging by, but also denotes the resulting effects, such as delayed responsiveness, slowed thought processes, and

diminished body movements. Not one single phrase in Chinese covers all the meanings above, so the current Chinese version used “提不起劲做事”, which the translators believed captured the general meaning (i.e., lack of energy) as much as possible. However, the experts in the panel had different opinions about the intended focus of the English phrase. E1’s suggestion was “生活节奏放慢” (i.e., *“having a slower pace of life”*), but it described a negative experience in a positive tone and thus conveyed the message inaccurately. E4 tended to translate it as “反应迟钝” (i.e., *“slowed physical and emotional reactions”*), which emphasized the psychomotor retardation symptom. In this situation, another round of forward and backward translation would not help to obtain a version approved of by all the experts.

Given that this item is asking whether people have felt different than they normally do and this phrase is just used as an example for demonstrating *feeling different*, the HDS-OA test developer (second author) suggested using a new example to replace the current one. In fact, there are two examples in this item: feeling *“unusually restless”* and feeling like you were *“moving in slow motion”*. Because the second one needs to be replaced, the point would be to find an alternative phrase that is the opposite of being “unusually restless”. The HDS-OA developer (second author) provided the phrase *“unusually listless/lethargic”*, which the researcher translated as “异常萎靡不振” in Chinese. The old version, the new adaptation, as well as E4’s recommendation, were sent to the four experts in the panel and two translators with counselling background in previous steps (i.e., FT3 and BT1) for comparison. Five of them responded. Among them, all except one, agreed that keeping both E4’s adaptation and the new example in the Chinese statement was necessary, because the two phrases represent different things, but both can reflect the opposite of being unusually restless. However, three of them proposed that the adaptation of *“listless/lethargic”* (i.e., “萎靡不振”) should be replaced with a

synonym “无精打采”, as the latter is more commonly seen in the target language. Only BT1 held a different view that the new example was good enough to lead the respondents towards the main point of the Item 10. Therefore, the final version included both “反应迟钝” and “无精打采” as suggested by most experts.

In addition to meaning, experts also evaluated item quality in terms of difficulty, familiarity and cultural specificity. For these three aspects, 13 of the translated items (Items 1, 2, 3, 5, 6, 9, 11, 12, 13, 14, 15, 16 and (A)) received unanimous approval of the expert panel. For the rest of the items, Items 4 and (B) were said to be more difficult than the original English version by one out of the four experts. For Items 7 and 10, two experts said they had the same level of difficulty as the English items, one chose easier and one chose more difficult. Meanwhile, the cultural specificity of Items 10 and (B) was rated as not satisfactory by one expert. As can be seen in the notes to Table 3.4, all the explanations that the experts gave were the same as the comments in the second column (i.e., meaning), so we would not repeat them here. Lastly, Item 8 was said to be more difficult by E2 and not of equal familiarity by E3, compared to the original text. They both provided the same reason that respondents might feel it too blunt as this statement used the word “死了” (i.e., *“dead”*) instead of saying it in a gentler way, such as “走了” (i.e., *“have gone”*) or “消失了” (i.e., *“disappear”*). However, these euphemisms for death are at a higher risk of being misunderstood or misinterpreted, so the current adaptation was kept, given that it was not identified as offensive in the Chinese culture.

With respect to the overall quality of the whole measure, all experts agreed that the item format and appearance, including the physical layout, item length, response option length and the use of phrase emphasis (bold, italics, etc.) was the same in the two language versions. As shown in Table 3.5, the only concern was about the response format, as E3 suggested a slight change to

how the statements are worded in the C-HDS-OA (from “有没有” to “是否”/“是不是”) to make them better correspond with the response options (i.e., 是/否). In fact, this was how FT1 translated the items at the very beginning but, later, FT3 recommended another way to phrase the statements in the adaptation so that the two language versions had an equally informal writing style. Because the current adaptation was acceptable to the other experts, it was not revised.

Table 3.5.

Expert Panel (N = 4) Feedback on Item Format & Appearance for the C-HDS-OA & Original HDS-OA

Questions	Proportions and feedback
1. Is the item format, including the physical layout, the same in the two language versions?	Same = 100%
2. Is the type or format of the test (i.e., use of item statements with dichotomous ‘yes’ vs. ‘no’ response options) equally familiar to respondents in the two language versions?	**Same = 75%; Not Same = 25%** E3: Maybe consider changing (some of) the phrases “有没有” in the item stems to “是不是”/”是否” which correspond better with the response options. I’m not sure, but it might be confusing for some Chinese participants to respond questions asking “有没有” using “是/否”.
3. Is the item length and response option length about the same in the two language versions?	Same = 100%
4. Is the use of word or phrase emphasis (bold, italics, underline, etc.) the same in the two language versions?	Same = 100%

Note: Panel response options were “Same” or “Not Same”. Cases without Same = 100% are bolded.

Step 3b: Pilot Testing of the Pre-final Version with a Participant Panel

Demographic characteristics for this sample of 26 Chinese adults who formed the participant panel can be found in Table 3.6. Just over half (53.8%) of the panel identified as

female. The participants ranged in age from 25-67 years (M = 45.46, SD = 15.07). We aimed to include roughly equal numbers of young adults (aged 18-35 years), middle-aged adults (aged 36-55 years), and older adults (aged 55-85 years). The panel included individuals with a range of education levels over the minimum high school requirement and a range of monolingual, bilingual, and multilingual speakers.

Table 3.6.

Demographics of the Participant Panel (N = 26)

Variables		N (%)
Age	Young: 25-35 years	9 (34.6%)
	Middle-aged: 36-55 years	8 (30.8%)
	Older: 56-67 years	9 (34.6%)
Gender	Female	14 (53.8%)
	Male	11 (42.3%)
	Prefer not to answer	1 (3.9%)
Highest Education	High School	10 (38.5%)
	Trade School/College Diploma	7 (26.9%)
	Undergraduate Degree	9 (34.6%)
Number of Languages Spoken	Monolingual (1)	10 (38.5%)
	Bilingual (2)	10 (38.5%)
	Multilingual (3 or more)	6 (23%)

Table 3.7 summarizes the percentage of participants who rated the adaptations as clear or unclear, and their feedback. All of the participants agreed that the rendering of the response format and Items 7, 10, 12, 13, 14 and (A) was clear. As for Items 4, 6, 9, 11 and 15, the inter-rater agreement among the sample was 96%. With regard to the instructions and Items 1, 2, 5, 8, 16 and (B), the agreement of raters was 92%. Three participants rated Item 3 as unclear. The panel's feedback on various items mostly addressed two things. First, they thought that some of

Table 3.7.

Participant Panel's (N=26) Ratings Regarding the Clarity of the C-HDS-OA

Item	Y	N	Explanation
1	24 (92%)	2 (8%)	个人认为很难定义有用和被需要，感觉改成无用和不被需要更好理解 (It's hard to define useful and needed. It might be better to change it to useless and not needed.) 是在哪一方面（单位、家等）有用和被需要，可能在不同处境有不同的感受 (Useful and needed in which way? At home or workplace? People might have different feelings in different situations.)
2	24 (92%)	2 (8%)	如果是生病也会受影响吧？是否需要更具体说明食欲变化是因为"生理上"还是"心理上"的某一维度？还是都囊括？(Appetite might also be influenced when people get sick. You might want to specify if changes in appetite is a result of their physical condition or mental condition, or both.)
3	22 (85%)	4 (15%)	一直精力充沛，大多数时候还是有一天都行？(Are you asking feeling full of energy for a day or for most of the time?) sort of overall 感觉换成是否疲倦乏力更能判断 (It'd be better to change it to 'feel tired') 精力充沛有时候会误认为躁狂(Sometimes full of energy might be mistaken as mania)
4	25 (96%)	1 (4%)	心情低落换成情绪低落好点 (It might be better to change "心情" to "情绪")
5	24 (92%)	2 (8%)	例如可加失眠多梦？(You can consider adding insomnia as an example.) 如果一直半夜睡算规律有变化吗？(If some people stay up late every day and don't go to bed until midnight, does it mean that their sleeping patterns changed?)
6	25 (96%)	1 (4%)	在集中注意力后面加个工作之类的词？玩游戏也可以集中的(Sometimes people can concentrate on playing games but not work. I think it might be better to say 'concentrating on something', like work.)
7	26 (100%)	0 (0%)	

Item	Y	N	Explanation
8	24 (92%)	2 (8%)	
9	25 (96%)	1 (4%)	
10	26 (100%)	0 (0%)	
11	25 (96%)	1 (4%)	
12	26 (100%)	0 (0%)	
13	26 (100%)	0 (0%)	
14	26 (100%)	0 (0%)	
15	25 (96%)	1 (4%)	
16	24 (92%)	2 (8%)	这个未来指的是什么未来（上班还是更加深层次的东西？） (What does ‘future’ refer to here? Career or something else?)
(A)	26 (100%)	0 (0%)	
(B)	24 (92%)	2 (8%)	
Instructions	24 (92%)	2 (8%)	前半句就觉得不太通顺 (The first sentence sounds a bit weird.)
Response Format & Options	26 (100%)	0 (0%)	

the items needed to be more detailed. For example, some participants suggested specifying in what way people have felt useful and needed in Item 1, and in what aspect they have difficulty concentrating in Item 6. However, these items were meant to be generic, otherwise it would limit the generalizability of the items to the general adult population. Second, some panel participants mentioned that the statements would be easier to understand if we replaced some of the adjectives with their antonyms. An instance would be using "*useless"* and *"not needed"* instead of *"useful"* and *"needed"* in Item 1, and *"tired"* instead of *"full of energy"* in Item 3. But such changes might make the Chinese version less equivalent to the English version, especially given that positive and negative wording/keying are known to sometimes be understood or processed differently (e.g., Chen, 2017; Coleman, 2013; Hubley, Zhu, & Zhang, 2019).

Overall, none of the adaptations were found to be unclear by over 20% of the participants, which indicated good quality according to Sousa and Rojjanasrirat's (2011) guidelines. Therefore, no further revisions were made, except that two characters were deleted from the first sentence of the instructions to make it sound more natural. The final version of the C-HDS-OA is shown in Table 3.8.

Discussion

The purpose of Study 1 was to adapt the HDS-OA from English to Chinese under the theoretical framework of the *ITC Guidelines* (ITC, 2017) and following Sousa and Rojjanasrirat's (2011) procedures for cross-cultural translation. The adaptation procedures included forward translation and two rounds of backward translation because six of the adapted items (i.e., Items 5, 7, 9, 10, 12, and 13) were revised after the first back-translation and thus needed to be re-examined. A pilot study was then performed as a final check with a four-person expert panel and a participant panel comprising 26 Chinese adults, which evaluated linguistic,

Table 3.8.

Final Version of the Chinese Hubley Depression Scale for Older Adults (C-HDS-OA)

Instructions: The following questions have to do with changes that might have taken place in your life recently. For each question, please circle the answer (yes or no) that best applies to you. 指导语：以下问题均关于你生活中最近可能发生的变化。针对每个问题请圈出你认为最符合你情况的答案（是或否）。
1. Over the past two weeks, have you felt useful and needed? 在过去的两周内，你有没有觉得自己有用和被需要？
2. Over the past two weeks, have you noticed any changes in your appetite? (examples: you didn't feel like eating or you felt hungrier than usual) 在过去的两周内，你有没有发现你的食欲有什么变化？（例如：感觉自己不想吃东西，或是觉得比往常更饿）
3. Over the past two weeks, have you felt full of energy? 在过去的两周内，你有没有感到精力充沛？
4. Over the past two weeks, have you often felt sad and downhearted? 在过去的两周内，你有没有经常感到伤心难过、心情低落？
5. Over the past two weeks, have your sleeping patterns changed? (examples: you have been waking up in the middle of the night or unusually early) 在过去的两周内，你的睡眠规律有没有变化？（例如：你会在半夜醒过来，或是比平常早醒许多）
6. Over the past two weeks, have you had difficulty concentrating? 在过去的两周内，你有没有感觉难以集中注意力？
7. Over the past two weeks, have you been interested in your usual activities? 在过去的两周内，你还有没有兴趣进行你平常喜欢的活动？
8. Over the past two weeks, have you felt that you (or others) would be better off if you were dead? 在过去的两周内，你有没有觉得如果自己死了，对自己（或他人）会比较好？

9. Over the past two weeks, have you become irritated more easily than usual?

在过去的两周内，你有没有变得比往常更容易烦躁？

10. Over the past two weeks, have you felt different than you usually do? (examples: you felt unusually restless or you felt like you were moving in slow motion)

在过去的两周内，你有没有觉得自己与往常有所不同？（例如：你感觉异常坐立不安，或是异常无精打采或反应迟钝）

11. Over the past two weeks, has anyone mentioned to you that you don't look or seem your usual self?

在过去的两周内，有没有人向你提起过，你看上去不像往常的自己？

12. Over the past two weeks, have you enjoyed doing things as much as ever?

在过去的两周内，你还能不能与之前一样从喜欢的事情中得到乐趣?

13. Over the past two weeks, have you felt like everything was your fault?

在过去的两周内，你有没有觉得所有事都是你的错？

14. Over the past two weeks, have you frequently felt like crying?

在过去的两周内，你有没有常觉得想哭？

15. Over the past two weeks, have you found it harder than usual to make decisions?

在过去的两周内，你有没有觉得比往常更难做决定？

16. Over the past two weeks, have you thought that the future looks hopeless?

在过去的两周内，你有没有觉得未来看起来没有希望？

(A) Have you started taking a new medication in the past month?

最近的一个月内，你有没有开始服用新的药物？

(B) Are you grieving for someone who has died in the past two months?

你有没有因为有人在近两个月内离世而处于哀痛之中？

Response Format & Options: Yes/No 是/否

cultural, conceptual and measurement equivalence between the English and Chinese versions. The translators and experts showed a high level of agreement to most adapted items except Items 5, 7, 9, 10, and 12.

How to adapt fixed expressions (for Items 7, 10 and 12) and select words with the same level of strength in the target language (for Items 5 and 9) remained the greatest challenges in adapting these items. In fact, many fixed expressions are language-specific, and the pragmatic meaning is not deducible from that of the literal words, which makes it difficult to translate. For example, the phrase *moving in slow motion* in Item 10, which refers to a slowing of physical and mental activity, implies a feeling of lack of energy. It was first translated word by word as "动作迟缓" (i.e., *"moving slowly"*), which seemed very vague in Chinese. The adaptation was then redrafted as "提不起劲做事" (i.e., *"lacking the energy in doing things"*), but some translators thought that it only covered part of the meaning. To include as much content as we could, the final version was revised as "异常无精打采或反应迟钝" (i.e., *"being unusually lethargic or having slowed physical and emotional reactions"*). Clearly, when it comes to this item, direct equivalents are hard to find in the target language; however, similar feelings and behaviors are still observed in the Chinese culture. Therefore, translators can find appropriate expressions that are tangible and easy-to-understand based on their knowledge and experience, or even coin their own phrases that transfer the same messages.

Another challenge is obtaining equivalence in word intensity between the source and target languages, given the fact that there are many synonyms that have the same meaning but different strengths in both languages. For instance, the word *irritated* in Item 9 was first adapted as "生气发火" and then revised as "烦躁". Although both words in Chinese are on a one-dimensional spectrum of anger, the former one indicates a higher level, which upgraded the

intensity of the original expression and led to a back-translation *lose your temper* in the first round. To detect the differences in strengths of adjectives between two language versions, blind backward translation is always essential as a quality control. It should be noted, however, that even with a proper forward translation, back translators are likely to choose a synonym with a different level of intensity in the source language. An example was that *irritated* was back translated as *dysphoric* in the second round by one translator, because he could not think of a better word at that moment, but he agreed the current adaptation "烦躁" was appropriate when he saw the original text. Therefore, group discussions with translators for clarification is strongly recommended as a subsequent step.

Study Strengths

The key strengths of Study 1 are the use of strong methods that follow well-defined translation guidelines and the careful choice of translators and experts. Although Chinese researchers have adapted multiple commonly used Western depression measures into Chinese, most adaptations of these measures were conducted before a single complete standard set of guidelines for adapting tests was released and thus they did not necessarily follow a rigorous and systematic procedure in doing their work (Yan, Xiao, & Hu, 2016). For example, Yu and Li's (2000) study only included one forward translated version, one backward translated version, and a pre-test. Without different versions in each step for comparison, discrepancies between the Chinese and original English version would be hard to identify. Therefore, in the present study, two translators with distinct backgrounds (one with content knowledge and one skilled at translating) generated independent adaptations in each of the forward and backward translation steps. A synthesis of these versions provided more nuanced adaptations that emphasized both technical accuracy and common usage (Ohrbach et al., 2013).

In addition, we conducted pilot testing with an expert panel and a participant panel. Though previous studies also employed a committee approach, some just invited content experts to evaluate the cultural relevancy of the translated items and/or bilingual researchers to rate the linguistic equivalence between the original and back-translated versions (Li et al., 2011; Wang et al., 2008). In our study, various types of experts were included to address language (e.g., professional translators), methods (e.g., methodologists with experience in developing and translating instruments), and content (e.g., mental health professionals). This helped us incorporate perspectives from diverse disciplines to improve the translated instrument. Moreover, we provided each expert with a list of questions focusing on comparisons of five aspects, including meaning, difficulty, familiarity, cultural specificity, and item format and appearance, to collect sufficient evidence.

Study Limitations

There remain a few potential limitations of Study 1. First, from a cultural perspective, this study only recruited individuals from Mainland China in the participant panel, and thus did not obtain ratings and feedback on the C-HDS-OA from subgroups living in Hong Kong and Taiwan. However, two of the translators were from Taiwan and the expert panel included individuals originally from Hong Kong and Taiwan, so there has been some linguistic and cultural considerations of these subgroups. Second, for the participant panel pre-test, we sent everyone an electronic copy of the C-HDS-OA and a question list for assessing clarity at the item level and collecting feedback. However, only some of the participants gave suggestions as requested for rewriting the statement when they rated it as unclear, so we lacked more specific information to revise the wording of possibly problematic items. A feasible solution in future research would be to use the same question list but administer it through a personal interview

format rather than the original impersonal (e.g., through an e-mail) format, as the former was reported to detect errors more effectively when the same materials were given (Reynolds & Diamantopoulos, 1998).

Despite the limitations, Study 1 appears to have resulted in a satisfactory translated version of the HDS-OA in Chinese. It is worth noting that, although this measure was initially designed with older adults, the content is appropriate for adults of any age because the criteria for depression are the same for adults of all ages. In addition to the development of the C-HDS-OA, this study also contributed a robust and effective methodological framework for cross-cultural translation, especially for the adaptation of depression instruments to the Chinese language.

Study 2: Validation of Intended Inferences from the C-HDS-OA

According to Sousa and Rojjanasrirat (2011), a full psychometric evaluation of the final C-HDS-OA version with a Chinese sample should be conducted as the final step of cross-cultural adaptation. In this study, several psychometric properties of the C-HDS-OA were evaluated, including internal (factor) structure, internal consistency, and convergent and discriminant evidence for validity.

Sample Recruitment

We aimed to recruit 200 to 400 adults over the age of 18 years (Charter, 1999; Frost, Reeve, Liepa, Stauffer, & Hays, 2007) who were raised in China and acquired the Chinese language as their primary language, regardless of their cultural background or nationality. Using convenience sampling, we recruited participants by putting an advertisement on the messenger application WeChat, which is the largest social media platform among Chinese people today,

functioning as a combination of Facebook, Messenger, and Twitter ("WeChat User & Business Ecosystem Report 2017 · TechNode", 2017).

Method

Measures. The following measures were completed by the study participants.

Chinese Hubley Depression Scale for Older Adults (C-HDS-OA). The Chinese adaptation of the16-item HDS-OA (C-HDS-OA) prepared in Study 1 was used in this study. The C-HDS-OA screens for depressive symptoms. It uses a dichotomous response format. Answering "*yes*" on all items except 1, 3, 7, and 12 (which are reversed scored) indicates a depressive response and is scored "1", while answering "*no*" indicates a non-depressive response and is scored "0". By summing the scores across the items, total scores range from 0 to 16, with higher scores representing higher level of depressive symptomatology. There are two additional questions that ask about the use of new medication and the existence of bereavement, but are not scored.

Chinese Patient Health Questionnaire-9 (C-PHQ-9). The PHQ-9 (Kroenke, Spitzer, & Williams, 2001) is a self-administered instrument designed to assist health care professionals in assessing and monitoring the severity of depression. It consists of nine items that correspond to the nine symptoms described in the DSM-IV for major depressive disorder. Each item generates a score of 0, 1, 2 or 3 based on the response options of "*not at all*", "*several days*", "*more than half the days*" or "nearly every day", so the total score range is 0-27. Higher scores indicate greater severity of depression. For the Chinese population, psychometric studies on the Chinese version of the PHQ-9 (Bian, He, Qian, Wu, & Li, 2009) reported internal consistency alpha coefficients of .82 and a two-month test-retest reliability of .76 in the general population (Yu et al., 2012). Evidence of validity was provided with moderately strong correlations with the mental

component scores from the SF-12 and Chinese Health Questionnaire, a moderate correlation with the Happiness Scale, and a weak correlation with the physical component scores from the SF-12 (Yu et al., 2012).

Chinese General Anxiety Disorder-7 (C-GAD-7). The GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006) is a 7-item scale for evaluating the severity of generalized anxiety disorder. Similar to the PHQ-9, all items in the GAD-7 are scored on a scale of 0 ("*not at all*") through 3 ("*nearly every day*"), resulting in a total score ranging from 0 to 21. Higher scores indicate greater severity of anxiety. The Chinese version of the GAD-7 (He, Li, Qian, Cui, & Wu, 2010) has demonstrated good reliability and validity in the initial validation study, as indicated by a Cronbach's alpha of .90, a strong correlation with another anxiety measure (r = .82), and high sensitivity (86.2%) and specificity (95.5%).

Chinese Beck Hopelessness Scale (C-BHS). The BHS is a 20-item scale for "measuring the extent of negative attitudes about the future" with regard to three aspects: feelings about the future, loss of motivation, and loss of expectations (Beck & Steer, 1988, p. 1). Nine items are negatively keyed and 11 items are positively keyed. A yes/no response format is used, so the total score ranges from 0-20, with higher scores representing higher levels of hopelessness. Psychometric studies support use of the Chinese translation of the BHS (Kong, Zhang, Jia, & Zhou, 2007) in China by presenting a Cronbach's alpha of .85 and a moderate correlation with a depression measure (r = .55) in a Chinese sample in rural areas (Han, 2008).

Chinese 12-Item Short Form Health Survey (C-SF-12). The SF-12 (Ware, Kosinski, & Keller, 1996) is a 12-item questionnaire for assessing mental and physical functional health status, which are indicated by physical (PCS) and mental (MCS) component scores, respectively. All 12 items are used for calculations of PCS and MCS, but are weighted differently. Each item

is answered using either a dichotomous yes/no format or a Likert-type response format with a three-, five-, or six-point scale. It is recommended that total scores for PCS and MCS should be calculated with the QualityMetric Health Outcomes Scoring Software and transformed to fit a 0-100 scale, so that higher scores represent higher level of health (Saris-Baglama et al., 2007). Shou et al.'s (2016) study of the Chinese SF-12 (Lam, Eileen, & Gandek, 2005) using a Chinese elderly sample demonstrated satisfactory internal consistency reliability with a Cronbach's alpha of 0.91. The SF-12 component scores showed high correlations with those from the original SF-36 in the general population (MCS: r = .91, PCS: r = .80; Lam, Lam, Fong, & Huang, 2013).

Demographic information. Demographic variables included the participants' gender, age, first language, primary language, education level, marital status, employment status, country of birth, length of time living in China, and length of time living abroad.

Procedures. The data were collected using an anonymous survey hosted by software called Wen Juan Xing, which is commonly used by Chinese universities. This free software offers ready-for-use survey templates. After different components of the survey were set up, a link was created by Wen Juan Xing, which was attached in the advertisement for recruitment. Participants clicked the link created by Wen Juan Xing and then completed the questionnaires online. Upon completing the study, participants received the equivalent of $1 CAD as a reward through an account set up on Wen Juan Xing.

Ethical considerations. Ethics approval for this study was obtained from the Behavioral Research Ethics Board at the University of British Columbia. One common ethical concern is whether there will be any harm in answering questionnaires measuring depression. However, Siu and the U.S. Preventive Service Task Force (2016, p. 380) claimed that "the magnitude of harms of screening for depression in adults is small to none". Even still, mental health and counselling

resources, such as crisis lines and websites for counselling therapists, were provided at the end of the survey, with the purpose of providing educational resources and also to help individuals who may be experiencing depression and hoping for support or assistance.

Data Analysis. The data analyses conducted to examine the psychometric properties of the C-HDS-A are described below.

Internal structure. Given that Hubley, Rajlic and Zumbo (2017) conducted an exploratory factor analysis of the original English version of the HDS-OA and the results supported a unidimensional structure, this study used a confirmatory factor analysis (CFA) to examine the factor structure of the C-HDS-OA. It was expected that unidimensionality of the C-HDS-OA would be confirmed in the Chinese sample. The chi-square value, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), Tucker-Lewis index (TLI) and comparative fit index (CFI) fit indices were used to evaluate the overall model fit (Hu & Bentler, 1999; Kline, 2005). A chi-square statistic close to 0 with a p-value greater than the significance level (.05) is an indicator of a good fit (Hooper, Coughlan, & Mullen, 2008), although it is well-recognized that the statistical power that comes with the larger sample sizes sought for CFA often results in a statistically significant finding. RMSEA < .06, SRMR < .05, TLI > .95 and CFI > .95 are key indicators of good model fit (Hu & Bentler, 1999). The estimation method was Robust Diagonally Weighted Least Squares (robust DWLS), which could provide more accurate parameter estimates with ordinal data (Mîndrilã, 2010). The variance of the latent factor was fixed to be 1.0 so that the factor loadings of all the indicators were freely estimated.

Internal consistency estimate of reliability. The internal consistency reliability of the items of the C-HDS-OA was evaluated by estimating inter-item tetrachoric correlation

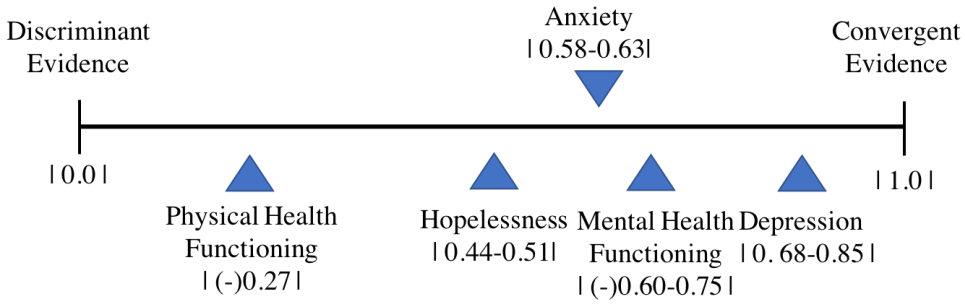
coefficients and using ordinal coefficient alpha (Zumbo, Gadermann, & Zeisser, 2007). Given that the scale uses a dichotomous response format, the data produced at the item level were binary and did not show a normal distribution. Hence, the ordinal version of Cronbach's alpha (i.e., ordinal coefficient alpha) was recommended to improve the accuracy of the estimate. However, Cronbach's alpha coefficient using the Pearson correlation matrix was estimated as well, so as to compare the results from this study with the previous ones. Coefficients equal to or greater than .80 are considered satisfactory. The same reliability analyses were conducted for the other scales using polychoric (ordinal responses) or tetrachoric (dichotomous responses) correlation matrices, as appropriate.

Convergent and discriminant evidence for validity. Convergent and discriminant evidence for validity was provided by calculating the Pearson's correlation coefficients between the C-HDS-OA and other external measures, including the C-PHQ-9, C-GAD-7, C-BHS, and C-SF-12 PCS and MCS. There is no absolute standard to classify measures as convergent or discriminant so it is useful to consider them along a continuum (Hubley & Zumbo, 2013). It was expected that the scores of the C-HDS-OA and C-PHQ-9 would show the strongest correlation, because they both measured the same construct of depression. A moderately strong, but lower, correlation was expected between the C-HDS-OA and C-GAD-7, due to the overlapping symptoms in the *DSM-5* for diagnosing major depression and anxiety disorders (e.g., fatigue, difficulty concentrating, sleep disturbance) and overlapping items in anxiety and depression scales developed based on these criteria (APA, 2013). The C-BHS was expected to be modestly correlated with the C-HDS-OA, as hopelessness was said to be "both a determinant and a component of the depressive condition" (Greene, 1989, p. 651). When feelings of hopelessness and negative thoughts linger for a long time, people might develop depression (Beck, Steer,

Kovacs, & Garrison, 1985). Though hopelessness is a core characteristic of PDD, other symptoms are required for making a diagnosis of MDD based on the *DSM-5* criteria, which makes the construct of depression distinct from hopelessness. Given previous studies suggesting an association between depression and health-related quality of life (Yu et al., 2012) and health conditions (APA, 2013; Canadian Mental Health Association, n.d.), it was expected that C-SF-12 MCS scores would show moderate to strong correlations with C-HDS-OA scores of a similar or high magnitude to that found with anxiety, whereas C-SF-12 PCS scores would be weakly correlated with C-HDS-OA scores. Figure 3.2 showed how these correlations were expected to be ranked on a theoretical continuum from low (more discriminant) to high (more convergent) relatedness to depression.

Figure 3.2.

The continuum of convergent and discriminant evidence for validity.



Note. The values are correlations observed from previous studies (Wu, Chen, Yu, Duan, & Jiang, 2015; Yang, Jia, & Qin, 2015; Yu et al., 2012; Yu, Sun, & Sun, 2017) using other measures of depression than the C-HDS-OA and the (-) sign means that the scores were negatively related to depression.

Results

Demographic Characteristics

A total of 380 adults who are Chinese speakers completed the questionnaires using Wen Juan Xing. Sixteen participants were excluded from data analysis because of abnormally short response time (i.e., less than 3 minutes) or careless responding. Given the overlapping content in the C-HDS-OA and C-PHQ-9, careless responding was identified by comparing each participant's responses to four pairs of C-PHQ:C-HDS-OA items: Items 4:3 (loss of energy), 5:2 (change in appetite), 7:6 (diminished ability to concentrate), 8:10 (psychomotor agitation or retardation), and 9:8 (thoughts of death). At the same time, responses to the Item 16 in the C-HDS-OA were compared to the total scores of the C-BHS, as they both were supposed to measure hopelessness. Any cases that showed inconsistency in more than two pairs of comparisons were dropped. Of the 364 remaining responses, there were no missing data points regarding the C-HDS-OA and the missing rate of the other four measures was from 0.04% to 0.67%. For the C-PHQ-9 and C-GAD-7, missing values were replaced with the mean score across the scale for the participant (i.e., case mean replacement) when only one item was missing in the measures. Total scores were not computed when two or more items were missing. As to the C-BHS, two missing items were allowed for case mean replacement because it has 20 items in total. No total scores were calculated if there was any missing value in the C-SF-12.

Demographic characteristics of the analyzed sample are summarized in Table 3.9. Overall, participants ranged in age from 18 to 81 years (M = 37.56, SD = 13.17), with just over half (51.4%) being in the younger age group (ages 18-35 years). Over half the sample (62.6%) were women. The sample tended to be well-educated, with 83.5% having more than a high school education. Most of the sample was married (55.5%) or single/unmarried (39.6%) and thus, not surprisingly, most did not live alone (89.8%). Most described themselves as employed full-time (55.2%), self-employed (9.1%), retired (11.8%), or a student (14.3%). The majority of

Table 3.9.

Demographics of the Participant (N = 364)

Variables		N (%)
Age	Young: 18-35 years	187 (51.4%)
	Middle-aged: 36-55 years	134 (36.8%)
	Older: 56-81 years	43 (11.8%)
Gender	Female	228 (62.6%)
	Male	133 (36.5%)
	Prefer not to answer	3 (0.9%)
Highest Education	Elementary School	7 (1.9%)
	High School	53 (14.6%)
	Trade School/College Diploma	88 (24.2%)
	Undergraduate Degree	144 (39.6%)
	Master's Degree	62 (17.0%)
	Doctoral Degree	7 (1.9%)
	Other	3 (0.8%)
Number of Languages Spoken	Monolingual (1)	104 (28.6%)
	Bilingual (2)	191 (52.5%)
	Multilingual (3 or more)	69 (18.9%)
Experience of Living Abroad	Yes	62 (17.0%)
	No	302 (83.0%)
Marital Status	Single/Unmarried	144 (39.6%)
	Married	202 (55.5%)
	Divorced	13 (3.6%)
	Widowed	3 (0.8%)
	Other	2 (0.5%)
Living Situation	Living alone	37 (10.2%)
	Living with others (e.g.., with a romantic partner, friend, family, or roommate)	327 (89.8%)
Employment Status	Employed-full time	201 (55.2%)
	Employed-part time	13 (3.6%)
	Self-employed	33 (9.1%)
	Student	52 (14.3%)
	Retired	43 (11.8%)

Not employed	17 (4.7%)
Other	5 (1.3%)

this sample was born in China (98.1%), and use Chinese as their primary language (97.3%), and speak more than one language (71.4%).

Internal Structure of the C-HDS-OA

After reverse coding Items 1, 3, 7 and 12, a CFA was conducted to determine if the C-HDS-OA has a unidimensional factor structure as has been found for the English version. The results are presented in Table 3.10. The one-factor model showed satisfactory goodness of fit indices except the chi-square and SRMR value: $\chi^2(104) = 171.61;\ p < .001$; CFI = .98; TLI = .97; RMSEA = .04; SRMR = .09.

Table 3.10.

Factor Loadings of the Unidimensional Model

Items	Factor loadings
1	0.510*
2	0.603*
3	0.484*
4	0.807*
5	0.652*
6	0.726*
7	0.408*
8	0.801*
9	0.877*
10	0.887*
11	0.659*
12	0.417*
13	0.840*
14	0.855*
15	0.826*
16	0.733*

* $p < .001$

Descriptive Results and Internal Consistency Reliability for Study Measures

The descriptive results and internal consistency reliability values for all measures used in the study are presented in Table 3.11. A wide range of scores was obtained for all measures. The internal consistency reliability for the C-HDS-OA was satisfactory, as seen with the obtained Cronbach's alpha of .85 and the more appropriate ordinal alpha of .93. In addition, all of the other measures yielded good internal consistency reliability. Because of the way that the C-SF-12 MCS and PCS scores are obtained (all items are used for both but weighted differently), it is not possible to compute meaningful internal consistency reliability coefficients for these scores.

Table 3.11.

Descriptive results and internal consistency reliability for all measures

	Possible Range	Obtained Range	Mean (SD)	Skewness (SE)	Kurtosis (SE)	Cronbach's Alpha	Ordinal Alpha
C-HDS-OA	0-16	0-16	3.53 (3.56)	1.24 (0.13)	0.90 (0.26)	0.85	0.93
C-PHQ-9	0-27	0-27	5.38 (4.61)	1.30 (0.13)	2.85 (0.26)	0.90	0.94
C-BHS	0-20	0-20	4.27 (4.31)	1.39 (0.13)	1.50 (0.26)	0.88	0.94
C-GAD-7	0-21	0-21	4.26 (4.07)	1.34 (0.13)	2.50 (0.26)	0.93	0.95
C-SF-12 MCS	0-100	15.27-61.65	45.73 (9.51)	-0.79 (0.13)	0.47 (0.26)	-	-
C-SF-12 PCS	0-100	22.36-64.65	51.02 (7.22)	-0.65 (0.13)	-0.38 (0.26)	-	-

Note. C-HDS-OA = Chinese Hubley Depression Scale for Older Adults; C-PHQ-9 = Chinese Patient Health Questionnaire-9 Items; C-BHS = Chinese Beck Hopelessness Scale; C-GAD-7 = Chinese General Anxiety Disorder-7 Items; C-SF-12 MCS = Short Form–12 Items Health Survey Mental Composite Score; C-SF-12 PCS = Short Form–12 Items Health Survey Physical Composite Score.

Convergent and Discriminant Evidence for Validity of C-HDS-OA Inferences

Convergent and discriminant validity evidence was examined along a theoretically and empirically expected continuum from convergent to discriminant (Hubley & Zumbo, 2013). Based on findings from validation research with other depression measures in the literature, Pearson's correlation coefficients were expected in the following approximate order: depression (~|0.68| to |0.85|), mental health functioning (~|0.60| to |0.75|), anxiety (~|0.58| to |0.67|), hopelessness (~|0.44| to |0.51|), and physical health functioning (~|0.27|). All validity coefficients showed the expected positive or negative signs based on how the measure was scored. The obtained validity coefficients in this study are presented in Table 3.12.

Table 3.12.

Pearson's correlation coefficients between the C-HDS-OA and convergent/ discriminant scales

	Depression (C-PHQ-9)	Hopelessness (C-BHS)	Anxiety (C-GAD-7)	Mental Health (C-SF-12 MCS)	Physical Health (C-SF-12 PCS)
C-HDS-OA	|0.75|*	|0.68|*	|0.67|*	|0.63|*	|0.28|*

Note. C-HDS-OA = Chinese Hubley Depression Scale for Older Adults; C-PHQ-9 = Chinese Patient Health Questionnaire-9 Items; C-BHS = Chinese Beck Hopelessness Scale; C-GAD-7 = Chinese General Anxiety Disorder-7 Items; C-SF-12 MCS = Short Form–12 Items Health Survey Mental Composite Score; C-SF-12 PCS = Short Form–12 Items Health Survey Physical Composite Score.

*$p < .001$

Discussion

Study 2 provided the first report on the psychometric properties of the C-HDS-OA in the general adult population in China. The CFA results supported a unidimensional structure of the C-HDS-OA, which was consistent with the original HDS-OA (Hubley et al., 2017). With respect to reliability, the results demonstrated good internal consistency with satisfactory ordinal alpha ($\alpha = .93$) and Cronbach's alpha ($\alpha = .85$), the latter of which was somewhat lower than the

Cronbach's alpha of the original HDS-OA (α = .88, Myers & Hubley, 2012; α = .94, Hubley et al., 2009).

Correlation analyses indicated appropriate convergent and discriminant evidence for validity of inferences from the C-HDS-OA. Validity coefficients were generally as expected except for a notably higher correlation with hopelessness. A moderately strong correlation was found between the C-HDS-OA and C-PHQ-9 (r = |0.75|). As theoretically expected, this was the highest validity coefficient obtained. The magnitude of the coefficient is consistent with previous validation research conducted with other Chinese measures of depression (r = |0.75|, Yang et al., 2015; r = |0.77|, Hu et al., 2014; r = |0.79|, Bian et al., 2009). The C-HDS-OA was also moderately correlated with anxiety (r = |0.67|) and mental health functioning (r = |0.63|). The magnitude of these coefficients are in line with Zhang et al.'s (2008) and Yu et al.'s (2012)' results, respectively, with other Chinese depression measures. More importantly, these correlations were notably lower than the one between the C-HDS-OA and C-PHQ-9, supporting that the C-HDS-OA appears to measure depression more than anxiety or mental health functioning, although these concepts are theoretically related to depression. The correlation between the C-HDS-OA and the C-BHS (r = |0.68|) was quite a bit higher than expected based on extant research with other Chinese depression measures (r = |0.44|, Yang et al., 2015; r = |0.51|, Wu et al., 2015), but given that hopelessness is "both a determinant and a component of the depressive condition" (Greene, 1989, p. 651), this finding is still supportive in terms of the validity of inferences made from the C-HDS-OA. And the correlation with physical health functioning (r = |0.28|) was considerably lower than the other correlations, as expected, and very similar to that found in previous research with other Chinese depression measures (r = |0.27|, Yu

et al., 2012). Generally, these results provided strong convergent and discriminant evidence for the interpretation and use of the C-HDS-OA in the Chinese adult population.

Study Strengths

Besides providing the first report on the psychometric properties of the C-HDS-OA in the general adult population in China, Study 2 has two key strengths. First is the use of multiple measures that fall along the theoretically expected convergent/discriminant continuum. Most validation studies of existing depression scales only selected measures of the same construct (i.e., depression) to examine construct validity (Chin et al., 2015; Li, Liu, Zhang, Wang, & Chen, 2011; Wang et al., 2014), which is not a strong test of the validity of the inferences to be made from the test scores. Without the inclusion of measures of constructs that may serve as alternative interpretations of test scores that can be compared to one another, it is difficult to critically evaluate such results. It is important to demonstrate that depression is a more likely interpretation of the C-HDS-OA scores than similar constructs such as anxiety or mental health functioning. Therefore, in Study 2, scores from measures of four additional constructs that are theoretically related to depression to different extents along a convergent/discriminant continuum were chosen for inclusion. Finding that the magnitudes of validity coefficients between the scores of the C-HDS-OA and these measures vary in the expected way, provides stronger evidence to support that level of depressive symptomatology is the more likely inference to be made from C-HDS-OA scores rather than inferences about anxiety, mental health functioning, hopelessness, and physical health functioning.

Another strength of Study 2 is the use of a more appropriate internal consistency reliability estimate. The ordinal alpha coefficient was adopted to estimate the internal consistency reliability of the instruments, taking into account the ordinal nature of the Likert-

type responses. This estimate is specifically designed for binary or ordinal data at the item level, and thus is more accurate than the frequently used Cronbach's alpha.

Study Limitations and Future directions

There are a few limitations to this study. First, we did not collect personal information on participants' regional identity, which means we cannot examine if there is any difference among Chinese speakers from different regions, such as Mainland China, Hong Kong and Taiwan. We recruited participants through online social media platforms that are widely used by Chinese speakers all over the world, so it would have been ideal to include an item in the demographic questionnaire to identify people from different regions of Chinese culture. Future research should collect such data, which would allow for measurement invariance to be examined and other reliability and validity evidence to be reported separately by regional group.

Second, the sample shows an unbalanced distribution of ages. Because participant recruitment and data collection were both conducted online and we know that many older adults may not regularly use the Internet, our sample is primarily comprised of young (51.4%) and middle-aged (36.8%) adults. Hence, the findings might be more relevant to young and middle-aged adults in China and caution should be taken when generalizing them to the older population. This study was conducted during a period of COVID-19 pandemic restrictions (including university research ethics restrictions) that prevented in-person recruitment and data collection, particularly with older adults. The use of mixed data collection methods (i.e., paper-and-pencil and Internet-based survey) in future research may address this shortcoming.

Last but not least, this study only reports validity evidence from three out of five sources suggested by the *Standards* (AERA, APA, & NCME, 2014). As the first attempt to validate inferences made from C-HDS-OA scores in a sample of Chinese adults, Study 2 provides

preliminary psychometric evidence for its use in the general population in China. However, validation of an instrument is an ongoing process. Reliability and validity are "not established by any single study but by the pattern of results across multiple studies" (Price et al., 2015, p 92). Most importantly, future studies that provide test-criterion evidence of validity are needed. An external reference standard (usually a clinical interview) should be administered, so that sensitivity and specificity can be obtained and an optimal cut-off score for identifying depressed and non-depressed individuals can be decided. It would also be important to explore response processes in forthcoming research, which allows us to understand how Chinese test takers interpret and answer the statements on the C-HDS-OA. Comparisons of responses from those who take the translated version and original English version may help to explain the tendency of Chinese individuals to report somatic symptoms of depression as has been reported in the clinical literature (Ren, Li, & Wang, 2001; Ryder et al., 2008). Finally, still more work is needed to investigate convergent and discriminant evidence of validity. The more adequately the relationship to other variables is examined, the more confident we can be about what the construct of the C-HDS-OA entails. Thus, future work should continue to explore how C-HDS-OA scores relate to a broad array of other variables, by using different measures of depression, anxiety, or hopelessness, or measures of some other constructs, such as optimism, distress, happiness or well-being. This could provide further support for the psychometric properties of the C-HDS-OA.

Concluding Remarks

In summary, support was found for the interpretation of total scores from the 16-item C-HDS-OA to be measuring current level of depressive symptomatology in the general adult population in China. In Study 1, the HDS-OA was adapted from English to Chinese following

well-defined translation guidelines and standards, which have attended to its cultural appropriateness. Study 2 provides strong psychometric evidence for the C-HDS-OA as a consistent unidimensional factor structure was found with the C-HDS-OA and the HDS-OA and both satisfactory internal consistency reliability and adequate convergent and discriminant evidence for validity were obtained. Still, more studies need to be conducted to evaluate the reliability and validity of the C-HDS-OA, especially to examine test-criterion evidence, before using it as a screen for depression in the Chinese population.

Chapter 4: Conclusion

According to epidemiological data, China has a low prevalence rate of depression, which is only one third of the estimates in Western countries such as United States, New Zealand and Netherlands (Kessler & Bromet, 2013). Despite the large difference in depression prevalence, the suicide rates in China and those countries are quite close (China: 8.0, Netherlands: 9.6, New Zealand: 11.6, United States: 13.7, per 100, 000 population) ("Suicide rate estimates", 2018). Given that depression is a leading cause of suicide in China, this suggests poor identification of depression (Liu et al, 2014).

A lack of effective screens may explain the failure of diagnosis. The results of two systematic reviews indicate that the SDS, HAMD, SCL-90, CES-D and HADS, which were all adapted from Western measures, are the most commonly used instruments in China (Jin & Zhang, 2017; Yan et al., 2016). However, their quality remains questionable because most of these scales were adapted before a single complete standard for adapting tests was established and thus the adaptation procedures might have been less than rigorous. Indeed, poor reliability and validity evidence are often obtained when these measures are used in different Chinese samples (Duan & Sheng, 2012; Peng et al., 2013; Zhang et al., 2013). In addition to directly translated scales, another choice is depression measures specifically designed for the Chinese population, such as the SSDA, CDSS, ASSR and DRSE, though none of them are as popular as the Western ones. In fact, most of these instruments lack consistently strong psychometric evidence except the initial scale development and validation study. Apart from being overly long and dated, there is a greater concern for using these Chinese-developed scales as little is known about how the items were collected and selected and what theoretical assumptions the test developers had regarding the construct of depression. Use of these existing measures with

unknown quality leads to a high risk of misdiagnosis, and a more effective screening tool for depression is thus needed for the Chinese population.

The HDS-OA was chosen to be adapted from English to Chinese because it was developed based on the latest diagnostic criteria, is relatively short (i.e., 16 items), and has demonstrated satisfactory psychometric properties in both depressed and non-depressed groups. We prepared a translated version of the HDS-OA in Study 1 referring to the *ITC Guidelines* (ITC, 2017), Sousa and Rojjanasrirat's (2011) recommendations, as well as Hambleton and Zenisky's (2010) work. The rigorous and systematic adaptation processes, which included forward and backward translation and pilot testing, minimize errors in adaptation, and are a strength of this study. To be more specific, direct comparisons of the original and back-translated items allowed us to detect any discrepancies between the Chinese and English versions, and group discussions with translators led to a consensus on the best adaptation. Another strength of this study was the careful choice of translators and experts. Translators from diverse backgrounds and a committee of experts from the fields of language, method and content can provide a mix of perspectives, which helped secure both semantic equivalence and cultural appropriateness of the adaptation.

In Study 2, we examined the psychometric properties of the Chinese version of the HDS-OA with a sample from the general adult population in China. The data supported a unidimensional factor structure of the C-HDS-OA and yielded good internal consistency reliability. Not only was the traditional Cronbach's α calculated, but also the ordinal α designed for Likert-type item response data was adopted, which helps generate a more accurate estimate of the reliability coefficient. A strong pattern of convergent and discriminant evidence for validity was also provided, as the correlations between the C-HDS-OA and the other four

measures of constructs that are supposed to be related to depression fell along the theoretically expected convergent/discriminant continuum. It is essential to use multiple convergent and discriminant measures in any validation study, so as to determine whether the intended construct is a more likely interpretation of the obtained scores than other similar constructs.

Despite satisfactory preliminary psychometric evidence, further studies should be performed to provide more support for the C-HDS-OA. First, future studies may consider conducting clinical diagnostic interviews along with the survey to examine test-criterion validity. With sensitivity and specificity for different scores reported, an optimal cut-off can be decided. Second, response processes are another important source of validity evidence and how Chinese understand and respond to items in the C-HDS-OA is also in need of investigation, given the cultural variations in somatic symptom presentation. Third, related constructs in the nomological network of depression are not limited to anxiety, hopelessness, and mental and physical health. Using measures of other constructs, such as happiness, quality of life, and subjective well-being, can provide additional convergent and discriminant evidence. Fourth, although the factor structure of the HDS-OA was found to be consistent in the English and Chinese versions, assumptions cannot be made that all samples or subsamples approach the items in the same way. Thus, measurement invariance should be tested in future research before making any comparisons across groups (e.g., female vs. male) or cultures (e.g., Chinese vs. Canadian). Lastly, future directions can include test-retest reliability and even item response theory analyses to examine reliability across the latent variable continuum (i.e., theta) and item analyses.

Thoughts about the Adaptation Process

Adaptation quality has always been my top concern from day one of this study. Now, at the very end of this project, I would like to share what I have learned from my experiences in

adapting the C-HDS-OA. There are three important things that we should consider before starting any adaptation work: preparation, documentation and pre-testing.

First and most importantly, more is less. Spending more time on up-front preparation actually shortens the time needed to produce a high-quality and ready-to-use instrument. I recommend that researchers have a number of possible translators in mind, and while waiting for their responses, think about which task that each person would be a good fit for (e.g., forward or backward translation?). What we need to consider is their various strengths and limitations, which include not only their language skills in the original and target languages but their regional and geographic differences regarding dialect and culture. For instance, we wanted to assign one translator from Mainland China and one Taiwanese translator for each translation step. Additionally, age and gender might also be considerations depending on the intended construct of the instrument. It is also important to prepare support materials to provide clarification to the translators about the measure of focus. For example, specify what the translated measure is used for, who the target population is, and what the tone of the text should be like (e.g., formal or casual). With this information provided, the translators can understand their task more and perform better. Notably, up-front preparation is also important for the expert panel and the participant panel (if included), especially creating the sheets for rating the adaptation quality. In fact, the original Review Form for quality evaluation from Hambleton and Zenisky (2010) lists 25 questions. I could have directly used it in Study 1, but it would be very time-consuming and perhaps overwhelming for the panel members to answer each of these questions for every item and for researchers to process and analyze these data. So spending a fair bit of time to carefully organize some sheets to cover groups of questions together seemed a prudent solution.

Second, keep clear documentation at every step during the adaptation, not only of the decision outcomes but also of the processes. In other words, one should keep a record of each of the forward and backward translations along with a description of any problems experienced by the translators, and any decisions made about the items in the intermediate and final versions. To be more specific, the documentation should contain: (1) any difficult-to-translate words or phrases and why, (2) points of disagreement and suggestions from the translators and experts, and (3) any revisions made and why these changes were made. This part is extremely important because instrument adaptation is an iterative process and one often has to go back and check their work at a later stage. Finally, it is also important to document the qualifications and experience of each translator and expert.

Last but not least, one should always conduct a pre-test before data collection. This allows one to check whether the translated instrument is working as intended and make appropriate adjustments before administering it to a large sample. This has been long suggested in test translation work (e.g., Brislin, 1970) but, given its significance, it might be more accurate to say that pre-testing is "an integral and necessary part of the translation process", rather than simply a possible option among various techniques (Pan & de La Puente, 2005, p. 15). This step requires special attention because pre-testing without well-designed approaches might lead to nothing. An example was that participants in Study 1 did not give much useful feedback on how to improve the translated measure even though they were given a question list that asked them to do so. Reynolds and Diamantopoulos's (1998) study indicated that the pretest methods actually had a great impact on error detection. They found that the detection rate was significantly higher when the pretest was carried out through personal interviews than when participants were left alone to complete the materials (i.e., impersonal administration). Therefore, it would have been

advisable for me to still use the same forms for pilot testing but through personal administration by an interviewer. Besides the administration format, the knowledge level of a participant on the survey topic is another factor that may influence the results of pre-testing (Reynolds & Diamantopoulos, 1998). Knowledgeable participants detect errors more effectively than randomly chosen participants with no knowledge of the measured construct, so an expert panel is essential. However, a participant panel from the target population is also needed to ensure that individuals from the intended group understand the instrument. As a critical step in cross-cultural translation, researchers should consider how pre-testing could be undertaken, especially what administration methods should be chosen and who should be the respondents.

Nowadays, various standardized guidelines that synthesize a wide spectrum of translation practices have been established, such as the *ITC Guidelines* (2017), Sousa and Rojjanasrirat's (2011) guidelines, and Hambleton and Zenisky's (2010) Item Translation and Adaptation Review Form for quality evaluation. These guidelines and recommendations provide systematic and methodological instructions on the adaptation processes, which helps to reduce the risk of omission, avoid errors in adaptation, and maximize cultural, conceptual, measurement and linguistic equivalence between the source language and target language versions. It is strongly recommended that practical and effective guidelines should be carefully chosen and rigorously followed when conducting adaptation work. More importantly, careful attention should be paid to the three final points in this section, which can further lead to better quality translated measures that may be more applicable to a wider target audience.

References

Abma, I. L., Rovers, M., & van der Wees, P. J. (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. *BMC Research Notes*, *9*(1), 226.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. Washington, DC, US: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC, US: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC, US: American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA, US: American Psychiatric Publishing.

American Psychiatric Association. (2017, January). What are anxiety disorders? Retrieved from https://www. psychiatry.org/patients-families/anxiety-disorders/what-are-anxiety-disorders

Anderson J. E., Michalak, E, E., & Lam, W. R. (2002). Depression in primary care: Tools for screening, diagnosis, and measuring response to treatment. *BC Medical Journal*, *44*(8), 415-419.

Aqeel, M., Jami, H., & Ahmed, A. (2017). Translation, adaptation, and cross-language validation of student: Thinking about my homework scale (STP). *International Journal of Human Rights in Healthcare*, *10*(5), 296-309.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*(1), 12-18.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Beck, A. T., Steer, R. A., Kovacs, M., & Garrison, B. (1985). Hopelessness and eventual suicide: A 10-year prospective study of patients hospitalized with suicidal ideation. *American Journal of Psychiatry, 142*(5), 559-563.

Beck, A. T., & Steer, R. A. (1988). *Beck Hopelessness Scale manual*. San Antonio, TX, US: The Psychological Corporation.

Beck, A. T., Weissman, A., Lester, D., & Trexler, R. (1974). The measurement of pessimism: The Beck hopelessness scale. *Journal of Consulting and Clinical Psychology, 42*(6), 861-865.

Bian, C. D., He, X. Y., Qian, J., Wu, W. Y., & Li, C. B. (2009). 患者健康问卷抑郁症状群量表在综合性医院中的应用研究[Reliability and validity of the Patient Health Questionnaire for screening depressive syndrome in general hospital outpatients]. *同济大学学报(医学版), 30*(5), 136-140.

Boyce, C., & Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input.* Watertown, MA, US: Pathfinder International.

Brink, T. L., & Niemeyer, L. (1992). Assessment of depression in college students: Geriatric depression scale versus center for epidemiological studies depression scale. *Psychological Reports*, *71*(1), 163-166.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology*, *1*(3), 185-216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Cross-cultural research and methodology series, Vol. 8. Field methods in cross-cultural research* (pp. 137-164). Thousand Oaks, CA, US: Sage Publications, Inc.

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99-103.

Cai, C. J. (2013). *广泛性焦虑障碍量表在基层医疗中应用的信度和效度* [Reliability and validity of a generalized anxiety disorder scale in primary care outpatients]. (Doctoral dissertation). 复旦大学.

Canadian Mental Health Association. (n.d.). The Relationship between mental health, mental illness and chronic physical conditions. Retrieved from https://ontario.cmha.ca/documents/the-relationship-between-mental-health-mental-illness-and-chronic-physical-conditions/

Cangur, S., & Ercan, I. (2015). Comparison of model fit indices used in structural equation modeling under multivariate normality. *Journal of Modern Applied Statistical Methods*, *14*(1), 152-167.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276.

Cha, E. S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing*, *58*(4), 386-395.

Chan, A. C. M. (1996). Clinical validation of the Geriatric Depression Scale (GDS) Chinese version. *Journal of Aging and Health*, *8*(2), 238-253.

Chan, E. K. H., Zumbo, B. D., Darmawanti, I., & Mulyana, O. P. (2014). Reporting of validity evidence in the field of health care: A focus on papers published in *Value in Health*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 257–265). New York, NY, US: Springer.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559-566.

Chen, Y. F. (2002). Chinese classification of mental disorders (CCMD-3): Towards integration in international classification. *Psychopathology*, *35*(2-3), 171-175.

Chen, Y. F., Cao, Y., & Liu, Z. K. (2013). 状态-特质焦虑量表中文修订版在流动儿童中的应用 [The application of a revised Chinese version of the State-Trait Anxiety Inventory in migrant children]. *中华行为医学与脑科学杂志, 22*(8), 755-757.

Chen, Y. M. (2017). *On the impact of negatively keyed items on the assessment of the unidimensionality of psychological tests and measures*. (Unpublished doctoral dissertation). University of British Columbia, Vancouver, BC, Canada.

Cheung, F. M., van de Vijver, F. J., & Leong, F. T. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *66*(7), 593-603.

Child, D. (1990). *The essentials of factor analysis*. London, NY, US: Cassell Educational.

Chin, W. Y., Choi, E. P., Chan, K. T., & Wong, C. K. (2015). The psychometric properties of the Center for Epidemiologic Studies Depression Scale in Chinese primary care patients: Factor structure, construct validity, reliability, sensitivity and responsiveness. *PloS One*, *10*(8), e0135131.

China Tech Insights. (2017, April 24). WeChat user & business ecosystem report 2017. Retrieved from https://technode.com/2017/04/24/wechat-user-business-ecosystem-report-2017/

Chinese Society of Psychiatry. (2001). *The Chinese classification and diagnostic criteria of mental disorders version 3 (CCMD-3)*. Jinan, Shandong, CHN: Science and Technology Press of Shandong Province.

Coleman, C. M. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study* (Unpublished doctoral dissertation). James Madison University, Virginia, U.S.A.

Collie, R. J., & Zumbo, B. D. (2014). Validity evidence in the *Journal of Educational Psychology*: Documenting current practice and a comparison with earlier practice. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 113-135). New York, NY, US: Springer.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9.

Cox, D. W., & Owen, J. J. (2014). Validity evidence for a perceived social support measure in a population health context. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 229-241). New York, NY, US: Springer.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto, ON, Canada: Holt, RineHart, and Winston, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach's alpha. (2019, August 19). Retrieved from https://en.wikipedia.org/wiki/Cronbach's_alpha

Dai, Y., Yu, X., Xiao, Z., Xu, Y., Zhao, M., Correia, J. M., ... & Reed, G. M. (2014). Comparison of Chinese and international psychiatrists' views on classification of mental disorders. *Asia-Pacific Psychiatry*, *6*(3), 267-273.

Duan, Q. Q., & Sheng, L. (2012). 焦虑及抑郁自评量表的临床效度 [Validity of SAS and SDS among psychiatric non-psychotic outpatients and their partners]. *中国心理卫生杂志, 26*(9), 676-679.

Ellis, G. K., Robinson, J. A., & Crawford, G. B. (2006). When symptoms of disease overlap with symptoms of depression. *Australian Family Physician, 35*(8), 647-649.

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment*, *93*(5), 445-453.

Ferraro, F. R., & Chelminski, I. (1996). Preliminary normative data on the Geriatric Depression Scale-Short Form (GDS-SF) in a young adult sample. *Journal of Clinical Psychology*, *52*(4), 443-447.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous and Mental Disease, 176*(5), 257-263.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health*, *10*, S94-S105.

Furr, R. M. (2017). *Psychometrics: An introduction*. Thousand Oaks, CA, US: Sage Publications.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction*. Thousand Oaks, CA, US: Sage Publications.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston, MA, US: Allyn & Bacon.

Gómez-Lugo, M., Espada, J. P., Morales, A., Marchal-Bertrand, L., Soler, F., & Vallejo-Medina, P. (2016). Adaptation, validation, reliability and factorial equivalence of the Rosenberg self-esteem scale in Colombian and Spanish population. *The Spanish Journal of Psychology*, *19*, 1-12.

Goodwin, G. M. (2006). Depression and associated physical diseases and symptoms. *Dialogues in Clinical Neuroscience, 8*(2), 259-265.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development*, *36*(3), 181-191.

Greene, S. M. (1989). The relationship between depression and hopelessness: Implications for current theories of depression. *The British Journal of Psychiatry*, *154*(5), 650-659.

Guerra, M., Prina, A. M., Ferri, C. P., Acosta, D., Gallardo, S., Huang, Y., et al (2016). A comparative cross-cultural study of the prevalence of late life depression in low and middle income countries. *Journal of Affective Disorders*, *190*, 362-368.

Gunnell, K. E., Schellenberg, B. J. I., Wilson, P. M., Crocker, P. R. E., Mack, D. E., & Zumbo, B. D. (2014). A review of validity evidence presented in the *Journal of Sport and Exercise Psychology* (2002–2012): Misconceptions and recommendations for validation research. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 137–156). New York, NY, US: Springer.

Gunnell, K. E., Wilson, P. M., Zumbo, B. D., Crocker, P. R. E., Mack, D. E., & Schellenberg, B. J. I. (2014). Validity theory and validity evidence for scores derived from the Behavioural Regulation in Exercise Questionnaire. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 175–191). New York, NY, US: Springer.

Chen, Y. F., Cao, Y., & Liu, Z. K. (2013). 状态-特质焦虑量表中文修订版在流动儿童中的应用 [The application of a revised Chinese version of the State-Trait Anxiety Inventory in migrant children]. *中华行为医学与脑科学杂志, 22*(8), 755-757.

Guo, X. J., Wang, Y. Q., & Chen, J. (2009). 状态-特质焦虑量表中文修订版在流动儿童中的应用 爱丁堡产后抑郁量表在成都地区产妇中应用的效能研究[A study on the efficacy of the Edinburgh Postnatal Depression Scale in puerperas in Chengdu]. *中国实用护理杂志*, *25*(1), 4-6.

Hajebi, A., Motevalian, A., Amin-Esmaeili, M., Rahimi-Movaghar, A., Sharifi, V., Hoseini, L., et al. (2018). Adaptation and validation of short scales for assessment of psychological distress in Iran: The Persian K10 and K6. *International Journal of Methods in Psychiatric Research*, *27*(3), e1726.

Hales, R. E. (2008). *The American psychiatric publishing textbook of psychiatry*. Washington, DC, US: American Psychiatric Publishing.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*(3), 164–172.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ, US: Lawrence Erlbaum Associates Inc.

Hambleton, R. K., Yu, J., & Slater, S. C. (1999). Field-test of the ITC Guidelines for adapting educational and psychological tests. *European Journal of Psychological Assessment*, *15*(3), 270-276.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver, *Cross-cultural research methods* (pp. 46-74). New York, NY, US; Cambridge University Press.

Han, Y. (2008). Beck 绝望量表中文版在农村使用的信度与效度研究[Reliability and validity of the revised Chinese version of the Beck Hopelessness Scale in rural areas]. (Doctoral dissertation). 大连医科大学.

Harachi, T. W., Choi, Y., Abbott, R. D., Catalano, R. F., & Bliesner, S. L. (2006). Examining equivalence of concepts and measures in diverse samples. *Prevention Science*, *7*(4), 359-368.

He, L. L. (2013). *神经衰弱和抑郁症概念发展中的文化分歧* [Cultural differences in the development of the concept of neurasthenia and depression]. (Doctoral dissertation). 南开大学.

He, J., Chen, Z. Y., Guo, F., Zhang, J., Yang, Y. P., & Wang, Q. (2013). 流调中心抑郁量表中文简版的编制[A short Chinese version of the Center for Epidemiologic Studies Depression Scale]. *中华行为医学与脑科学杂志, 22*(12), 1133-1136.

He, X. Y., Li, C. B., Qian, J., Cui, H. S., & Wu, W. Y. (2010). 广泛性焦虑量表在综合性医院的信度和效度研究 [Reliability and validity of a generalized anxiety disorder scale in general hospital outpatients]. *上海精神医学, 22*(4), 200-203.

He, X. Y., Xiao, S. Y., & Zhang, D. X. (2008). 老年抑郁量表在中国农村社区老年人中的信度和效度 [Reliability and validity of the Chinese version of the Geriatric Depression Scale: A study in the population of Chinese rural community-dwelling elderly]. *中国临床心理学杂志, 16*(5), 473-475.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. (Methods, plainly speaking). *Measurement and Evaluation in Counseling and Development*, *34*(3), 177-190.

History of depression. (2019, August 9). Retrieved from https://en.wikipedia.org/wiki/History_of_depression

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523-531.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hu, X. C., Zhang, Y. L., Liang, W., Zhang, H. M., & Yang, S. C. (2014). 病人健康问卷抑郁量表(PHQ-9)在青少年中应用的信效度检验[Reliability and validity of the Patient Health Questionnaire-9 in Chinese adolescents]. *四川精神卫生, 27*(4), 357-360.

Hubley, A. M. (1998). Hubley Depression Scale for Older Adults (HDS-OA). Unpublished manuscript. University of Northern British Columbia, Prince George, BC.

Hubley, A. M. (2014). Hubley Depression Scale for Older Adults (HDS-OA). *Encyclopedia of Quality of Life and Well-Being Research*, 2992-2995.

Hubley, A. M., Mangaoang, M., Burke, S., Ho, C., Ang, P., Myers, S., & Chiu, D. (2009). *Validation of a new screen for depression in older adults.* Presented at the 117th annual meeting of the American Psychological Association (APA), Toronto, ON, Canada.

Hubley, A. M., Rajlic, G., & Zumbo, B. D. (2017). *Hubley Depression Scale for Older Adults (HDS-OA): Factor structure, scoring, and reliability evidence*. Presented at the annual conference of the Association for Psychological Science (APS), Boston, MA, U.S.A.

Hubley, A. M., Zhu, M., Sasaki, A., & Gadermann, A. (2014). A synthesis of validation practices in the journals *Psychological Assessment* and *European Journal of Psychological Assessment*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193-213). New York, NY, US: Springer.

Hubley, A. M., Zhu, S. M., & Zhang, W. Q. (2019). *What response processes can tell us about positively and negatively worded and keyed items.* Presented at the 26th annual conference of International Society for Quality of Life Research (ISOQOL), San Diego, CA, USA.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in*

psychology (Vol. 1, pp. 3-19). Washington, DC, US: American Psychological Association Press.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-cultural Psychology*, *20*(3), 296-309.

Institute of Medicine. (2015). *Psychological testing in the service of disability determination*. Washington, DC, US: National Academies Press.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests (second edition).* Retrieved from https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

List of countries by suicide rate. (2019, August 23). Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate

Iwata, N., & Buka, S. (2002). Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America. *Social Science & Medicine*, *55*(12), 2243-2252.

Jang, Y., Kwag, K. H., & Chiriboga, D. A. (2010). Not saying I am happy does not mean I am not: Cultural influences on responses to positive affect items in the CES-D. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *65*(6), 684-690.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. *European Journal of Psychological Assessment*, *15*(3), 277.

Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing*, *28*(4), 589-596.

Jin, T., & Zhang, L. J. (2017). 我国常用的抑郁自评量表介绍及应用[Introduction and application of self-rating depression scales in China]. *神经疾病与精神卫生, 17*(5), 366-369.

Jones, P. S., Lee, J. W., Phillips, L. R., Zhang, X. E., & Jaceldo, K. B. (2001). An adaptation of Brislin's translation model for cross-cultural research. *Nursing Research*, *50*(5), 300-304.

Kalfoss, M. (2019). Translation and adaption of questionnaires: A nursing challenge. *SAGE Open Nursing*, *5*, 1-13.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

Kapfhammer, H. P. (2006). Somatic symptoms in depression. *Dialogues in Clinical Neuroscience*, *8*(2), 227–239.

Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy and Research*, *11*(3), 289-299.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health*, *34*, 119-138.

Kleinman, A. (1982). Neurasthenia and depression: a study of somatization and culture in China. *Culture, Medicine and Psychiatry*, *6*(2), 117-190.

Kleinman, A. (1986). *Social origins of disease and distress: Depression, neurasthenia, and pain in modern China.* New Haven, CT, US: Yale University Press.

Kline, R.B. (2005). *Principles and practice of structural equation modeling. 2nd edition*. New York, NY, US: The Guilford Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling. 3rd edition*. New York, NY, US: The Guilford Press.

Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the CES-D depression symptoms index. *Journal of Aging and Health*, *5*(2), 179-193.

Kong, Y. Y., Zhang, J., Jia, S. H., & Zhou, L. (2007). Beck 绝望量表中文版在青少年中使用的信度和效度[Reliability and validity of the Beck Hopelessness Scale in Chinese adolescents]. *中国心理卫生杂志, 21*(10), 686-689.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606-613.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, *69*, 313-332.

Lam, C. L., Eileen, Y. Y., & Gandek, B. (2005). Is the standard SF-12 health survey valid and equivalent for a Chinese population? *Quality of Life Research*, *14*(2), 539-547.

Lam, R. W., Kennedy, S. H., McIntyre, R. S., & Khullar, A. (2014). Cognitive dysfunction in major depressive disorder: Effects on psychosocial functioning and implications for treatment. *The Canadian Journal of Psychiatry*, *59*(12), 649-654.

Lam, E. T., Lam, C. L., Fong, D. Y., & Huang, W. W. (2013). Is the SF-12 version 2 Health Survey a valid and equivalent substitute for the SF-36 version 2 Health Survey for the Chinese? *Journal of Evaluation in Clinical Practice*, *19*(1), 200-208.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127-135.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 24-28.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23-30.

Lee, S. (1999). Diagnosis postponed: Shenjing shuairuo and the transformation of psychiatry in post-Mao China. *Culture, Medicine and Psychiatry*, *23*(3), 349-380.

Lee, E. A. (2009). *Achieving semantic equivalence between the Chinese and English versions of the Postpartum Depression Screening Scale*. (Doctoral dissertation). Retrieved from https://search.proquest.com/openview/d9da355b809c6310b82f44c7e0671201/1?pq-origsite=gscholar&cbl=18750&diss=y

Lee, H. C., Chiu, H. F. K., Wing, Y. K., Leung, C. M., Kwong, P. K., & Chung, D. W. S. (1994). The Zung self-rating depression scale: Screening for depression among the Hong Kong Chinese elderly. *Journal of Geriatric Psychiatry and Neurology*, *7*(4), 216-220.

Lee, M. H., Kim, Y., & Cho, S. H. (2017). Review on diagnostic criteria of neurasthenia: Suggesting pathway of culture-bound diseases. *Journal of Pharmacopuncture*, *20*(4), 230-234.

Lee, E. J., Kim, J. B., Shin, I. H., Lim, K. H., Lee, S. H., Cho, G. A., et al. (2010). Current use of depression rating scales in mental health setting. *Psychiatry Investigation*, *7*(3), 170-176.

Lei, L. Y., He, X. Y., Cao, R. F., Zhao, G. Q., Wang, Y. G., & Wang W. D. (2013). 社区老年人抑郁筛查量表的初步编制及信度效度检验[Construction and reliability and validity test of the depression screening scale for community elderly]. *全科医学临床与教育*, *11*(5), 492-495.

Lei, P. W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and practice*, *26*(3), 33-43.

Leong, F. T., Bartram, D., Cheung, F., Geisinger, K. F., & Iliescu, D. (2016). *The ITC international handbook of testing and assessment*. New York, NY, US: Oxford University Press.

Li, W., Cheung, H., Chung, O. K. J., & Ho, K. Y. (2010). Center for Epidemiologic Studies Depression Scale for Children: Psychometric testing of the Chinese version. *Journal of Advanced Nursing*, *66*(11), 2582-2591.

Li, Z., & Hicks, M. H. R. (2010). The CES-D in Chinese American women: Construct validity, diagnostic validity for major depression, and cultural response bias. *Psychiatry Research*, *175*(3), 227-232.

Li, L., Liu, F., Zhang, H., Wang, L., & Chen, X. (2011). Chinese version of the Postpartum Depression Screening Scale: Translation and validation. *Nursing Research*, *60*(4), 231-239.

Li, W. L., & Qian, M. Y. (1995). 状态特质焦虑量表中国大学生常模修订[The application of the State-Trait Anxiety Inventory in Chinese college students]. *北京大学学报：自然科学版, 31*(1), 108-114.

Li, Z. J., Qiu, B. W., & Wang, J. S. (2001). 青少年归因风格及其与心理健康水平关系的研究[Attributional style and its correlation with mental health in adolescents]. *中国心理卫生杂志, 15*(1), 6-8.

Lin, N. (1989). Measuring depressive symptomatology in China. *Journal of Nervous and Mental Disease, 177*(3), 121-131.

Lipovac, M., Chedraui, P., Gruenhut, C., Gocan, A., Stammler, M., & Imhof, M. (2010). Improvement of postmenopausal depressive and anxiety symptoms after treatment with isoflavones derived from red clover extracts. *Maturitas*, *65*(3), 258-261.

Liu, F. (2010). 农村产后抑郁症筛查量表（PDSS）中文版修订及在长沙市产妇中的应用研究[Translation and validation of the Chinese version of the Postpartum Depression Screening Scale in mothers of Changsha]. (Doctoral dissertation). 中南大学.

Liu, N. H., Contreras, O., Muñoz, R. F., & Leykin, Y. (2014). Assessing suicide attempts and depression among Chinese speakers over the internet. *Crisis, 35*(5), 322-329.

Liu, H., Jia, C. X., Xu, A. Q., Qiu, H. M., Lu, C. F., & Wang, L. L. (2011). Beck 绝望量表在农村自杀死亡研究中的信效度[Reliability and validity of the Beck Hopelessness Scale in a rural suicide study]. *中国心理卫生杂志, 25*(11), 867-871.

Liu, X. C., Tang, M. Q., Peng, X. G., Chen, K., & Dai, Z. S. (1995). 焦虑自评量表 SAS 的因子分析[Factor analysis of the Zung's Self-rating Anxiety Scale]. *中国神经精神疾病杂志,* 6, 359-360.

Liu, J., Wang, Y., Wang, X. H., Song, R. H., & Yi, X. H. (2013). 中文版老年抑郁量表在城市社区老年人群中应用的信效度研究[Reliability and validity of the Chinese version of the Geriatric Depression Scale among Chinese urban community-dwelling elderly population]. *中国临床心理学杂志, 21*(1), 39-41.

Liu, S. I., Yeh, Z. T., Huang, H. C., Sun, F. J., Tjung, J. J., & Hwang, L. C., et al. (2011). Validation of patient health questionnaire for depression screening among primary care patients in Taiwan. *Comprehensive Psychiatry, 52*(1), 96-101.

Ma, Y. G., Zhao, J. J., Li, S. Y., Wang, L. P., & Rong, P. J. (2014). 耳穴电针治疗心脾两虚型抑郁症患者 23 例随机单盲试验[Auricular electro-acupuncture for 23 depression patients with heart-spleen deficiency syndrome: A random single-blind clinical trial]. *中医杂志, 55*(17), 1484-1486.

Mao, J. L. (2013). 综合医院非专科心理障碍诊治的现状、困难及对策[The status quo, difficulties, and strategies of psychological disorders as a psychological non-specialist in general hospital]. *医学与哲学(B), 34*(2), 8-12.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence based on relations to other variables: Bolstering the empirical validity arguments for constructs. In D. B. McCoach, R. K. Gable, & J. P. Madura, *Instrument development in the affective domain* (pp. 209-248). New York, NY, US: Springer.

McDermott, M. A. N., & Palchanes, K. (1994). A literature review of the critical elements in translation theory. *Image: The Journal of Nursing Scholarship*, *26*(2), 113-118.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82.

Meng, Z. L. (2005). *情绪心理学* [Psychology of emotion]. 北京: 北京大学出版社.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washington, DC, US: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1-3), 35-44.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*(4), 283-298.

Mîndrilã, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, *1*(1), 60-66.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430-445.

Myers, S. L., & Hubley, A. M. (2012). *Hubley Depression Scale for Older Adults (HDS-OA): Reliability, validity, and a comparison to the Geriatric Depression Scale.* Presented at the 40thannual meeting of the International Neuropsychological Society (INS), Montréal, QC, Canada.

Nakao, M., & Barsky, A. J. (2007). Clinical application of somatosensory amplification in psychosomatic medicine. *BioPsychoSocial Medicine*, *1*(1), 1-17.

Newmark, P. (1991). *About Translation: Multilingual Matters (Series) 74*. Clevedon, Philadelphia, Adelaide: Multilingual Matters.

Nida, E. A. (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Leiden, Netherlands: Brill Archive.

Nida, E. A., & De Waard, J. (1986). *From one language to another: Functional equivalence in Bible translation*. Nashville, TN, US: Thomas Nelson.

Nida, E. A. (1993). *Language, culture, and translating*. Shanghai, CHN: Shanghai Foreign Language Education Press.

Nunnally, J. C. (1978). *Psychometric theory, 2nd edition*. New York, NY, US: McGraw-Hill.

Ohrbach, R., Bjorner, J., Jezewski, M. A., John, M. T., & Lobbezoo, F. (2013). *Guidelines for establishing cultural equivalency of instruments*. New York, NY, US: University of Bufalo.

Oishi, A., Haruta, J., Yoshimi, K., Goto, M., Yoshida, K., & Yoshimoto, H. (2017). Cross-cultural adaptation of the professional version of the Readiness for Interprofessional Learning Scale (RIPLS) in Japanese. *Journal of Interprofessional Care*, *31*(1), 85-90.

Pan, Y., & de La Puente, M. (2005). Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed. *Survey Methodology*, *6*.

Papasavvas, T., Al-Amin, H., Ghabrash, H. F., & Micklewright, D. (2016). Translation and validation of the Cardiac Depression Scale to Arabic. *Asian Journal of Psychiatry*, *22*, 60-66.

Parker, G., Cheah, Y. C., & Roy, K. (2001). Do the Chinese somatize depression? A cross-cultural study. *Social Psychiatry and Psychiatric Epidemiology*, *36*(6), 287-293.

Peng, H., Zhang, Y. Y., Ji, Y., Tang, W. Q., Li, Q., Yan, X. L., & Zhuang, Q. (2013). 农村地区女性自评抑郁量表中文版的信度效度分析[Analysis of reliability and validity of the Chinese version SDS scale in women in rural areas]. *上海医药, 34*(14), 20-23.

Polychoric correlation. (2019, August 22). Retrieved from https://en.wikipedia.org/wiki/Polychoric_correlation

Postpartum depression facts. (n.d.). Retrieved from https://www.nimh.nih.gov/health/publications/postpartum-depression-facts/index.shtml

Price, P. C., Jhangiani, R., & Chiang, I. C. A. (2015). *Research Methods in Psychology*. BC Campus. Retrieved from https://opentextbc.ca/researchmethods/

Qi, G. Y. (2016). The importance of English in primary school education in China: Perceptions of students. *Multilingual Education*, *6*(1), 1-18.

Qu, S., & Sheng, L. (2015). 广泛性焦虑量表在综合医院心理科门诊筛查广泛性焦虑障碍的诊断试验[A diagnostic test for screening generalized anxiety disorders in general hospital psychological department with the GAD-7]. *中国心理卫生杂志, 29*(12), 939-944.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, *1*(3), 385-401.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*(1), 1-9.

Ren, Q. T., Li, G., & Wang, Y. L. (2001). 非精神科门诊抑郁症的躯体化研究[Investigation on Somatization of Depression in Non-psychiatric Department]. *中国行为医学科学*, *10*(6), 580-581.

Reynolds, N., & Diamantopoulos, A. (1998). The effect of pretest method on error detection rates. *European Journal of Marketing, 32*(5/6), 480-498.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108-116.

Robinson, J. P., Shaver, P. R., Wrightsman, L. S., Wang, D. Y., Wang, D. F., & Yang, Y. Y. (1997). *性格与社会心理测量总览* [Measures of personality and social psychological constructs: An overview]. 台湾远流出版事业股份有限公司.

Rowan, N., & Wulff, D. (2007). Using qualitative methods to inform scale development. *The Qualitative Report*, *12*(3), 450-466.

Ruf, M., & Morgan, O., & Mackenzie, K. (2017). Differences between screening and diagnostic tests and case finding. Retrieved from https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2c-diagnosis-screening/screening-diagnostic-case-finding

Rupp, A. A., & Leighton, J. P. (2016). *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*. Malden, MA, US: John Wiley & Sons.

Ryder, A. G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S. J., & Bagby, R. M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, *117*(2), 300-313.

Sahni, A., & Agius, M. (2017). The use of the PHQ-9 self-rating scale to assess depression within primary care. *Psychiatria Danubina* , *29*(3), 615-618.

Salkind, N. J. (2010). *Encyclopedia of research design.* Thousand Oaks, CA, US: Sage.

Saris-Baglama, R., Dewey, C., Chisholm, G., Plumb, E., Kosinski, M., Bjorner, J. B., & Ware, J. E. (2007). QualityMetric Health Outcomes™ scoring software 2.0 user's guide. *Lincoln, RI: QualityMetric Incorporated*.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338.

Shaw, S., & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Research Matters*, *11*, 14-19.

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in *Educational and Psychological Measurement*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91–111). New York, NY, US: Springer.

Shi, Z. K., Liang, X. J. & Gao, C. Y. (2016). 丹栀逍遥散联合电针及重复经颅磁刺激治疗产后抑郁症 33 例[*Danzhi Xiaoyao* Power combined with electro-acupuncture and repetitive transcranial magnetic stimulation for 33 cases of postpartum depression]. *辽宁中医杂志*, *43*(6), 1220-1223.

Shou, J., Ren, L., Wang, H., Yan, F., Cao, X., Wang, H., et al. (2016). Reliability and validity of 12-item Short-Form health survey (SF-12) for the health status of Chinese community elderly population in Xujiahui district of Shanghai. *Aging Clinical and Experimental Research*, *28*(2), 339-346.

Simon, G. E., VonKorff, M., Piccinelli, M., Fullerton, C., & Ormel, J. (1999). An international study of the relation between somatic symptoms and depression. *New England Journal of Medicine*, *341*(18), 1329-1335.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 226-235.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100-107.

Siu, A. L., Bibbins-Domingo, K., Grossman, D. C., Baumann, L. C., Davidson, K. W., Ebell, M., et al. (2016). Screening for depression in adults: US Preventive Services Task Force recommendation statement. *Journal of the American Medical Association*, *315*(4), 380-387.

Song, L. L., & Liu, L. (2007). 大学生抑郁问卷的编制和信效度研究[The development of the College Student Self-rating Depression Scale and the test on its reliability and construct validity]. *教育理论与实践, 27*(2), 108-109.

Sousa, V. D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice*, *17*(2), 268-274.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092-1097.

Squires, A., Aiken, L. H., van den Heede, K., Sermeus, W., Bruyneel, L., Lindqvist, R., et al. (2013). A systematic survey instrument translation process for multi-country, comparative health workforce studies. *International Journal of Nursing Studies*, *50*(2), 264-273.

Srivatsan, S., Guduguntla, V., Young, K. Z., Arastu, A., Strong, C. R., Cassidy, R., & Ghaferi, A. A. (2018). Clinical versus patient-reported measures of depression in bariatric surgery. *Surgical Endoscopy*, *32*(8), 3683-3690.

Stapleton, C. D. (1997). *Basic concepts and procedures of confirmatory factor analysis*. Presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, U.S.A.

Steber, C. (2017, January 23). In-depth interviews: Data collection advantages and disadvantages. Retrieved from https://www.cfrinc.net/cfrblog/in-depth-interviewing

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. New York, NY, US: Oxford University Press.

Structured Clinical Interview for DSM-5 (SCID-5). (n.d.). Retrieved from https://www.appi.org/products/structured-clinical-interview-for-dsm-5-scid-5

Suhr, D. (2006). Exploratory or confirmatory factor analysis. *SAS Users Group International Conference* (pp. 1-17). Cary, NC, US: SAS Institute, Inc.

Suicide rate estimates. (2018, July 17). Retrieved from https://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en

Sun, X. Y., Li, Y. X., Yu, C. Q., & Li, L. M. (2017). 中文版抑郁量表信效度研究的系统综述 [Reliability and validity of the Chinese versions of depression scales: A systematic review]. *中华流行病学杂志, 38*(1), 110-116.

Sun, Z. X., Liu, H. X., Jiao, L. Y., Zhou, T., Yang, L. N., Fan, J. Y. (2017). 医院焦虑抑郁量表的信度及效度研究 [Reliability and validity of the Hospital Anxiety and Depression Scale]. *中华临床医师杂志(电子版), 11*(2), 198-201.

Tao, M., & Gao, J. F. (1994). 修订焦虑自评量表（SAS-R）的信度及效度[Reliability and validity of the revised Zung's Self-rating Anxiety Scale]. *中国神经精神疾病杂志*(5), 301-303.

Tennant, A., & Pallant, J. F. (2012). The root mean square error of approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions*, *25*(4), 1348-1349.

Think aloud protocol. (2019, August 28). Retrieved from https://en.wikipedia.org/wiki/Think_aloud_protocol

Tian, D. C., Zhang, X. Q., Jiang, F. Y., Zhou, J. S., Fang, X. H., & Min, B. Q. (2011). 社区抑郁状态电话筛查问卷的编制及其信度和结构效度检验 [The development of phone

screening questionnaire and the test on its reliability and construct validity]. *中国神经精神疾病杂志, 37*(9), 559-564.

Tian, D. C., Zhang, X. Q., Zhou, J. S., Jiang, F. Y., Liu, X. N., & Fang, X. H. (2011). 自行编制的社区抑郁状态电话筛查问卷与常用抑郁和焦虑评定量表的一致性研究 [The study on the correlation of community depression status phone screening questionnaire with conventional depression and anxiety scales]. *中国神经精神疾病杂志, 37*(10), 618-620.

Tsang, S., Royse, C. F., & Terkawi, A. S. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia*, *11*(Suppl 1), S80-S89.

Tylee, A., & Gandhi, P. (2005). The importance of somatic symptoms in depression in primary care. *Primary Care Companion to the Journal of Clinical Psychiatry*, *7*(4), 167-176.

Van Blerkom, M. L. (2017). *Measurement and statistics for teachers*. New York, NY, US: Routledge.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical approach to modelling cognitive*. London, UK: Academic Press.

Wancata, J., Alexandrowicz, R., Marquart, B., Weiss, M., & Friedrich, F. (2006). The criterion validity of the Geriatric Depression Scale: A systematic review. *Acta Psychiatrica Scandinavica*, *114*(6), 398-410.

Wang, X. G. (2007). *青少年学生抑郁自评量表的初步编制* [A primary design of the Adolescent Student Self-rating Depression Scale]. (Doctoral dissertation). 西南大学.

Wang, W., Bian, Q., Zhao, Y., Li, X., Wang, W., & Du, J., et al. (2014). Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population. *General Hospital Psychiatry, 36*(5), 539-544.

Wang, Z. Y., & Chi, Y. F. (1984). Self-rating depression scale (SDS). *Shanghai Archives of Psychiatry*, 2, 71-72.

Wang, C., Chu, Y. M., Zhang, Y. L., Zhang, N., Zhang, J., & Yang, H., et al. (2011). 汉密尔顿焦虑量表的因素结构研究[The factorial structure of the Hamilton Anxiety Rating Scale]. *临床精神医学杂志*(5), 299-301.

Wang, J. S., & Ding, X. H. (2003). 初中生主观幸福感与人格特征的关系研究[The relationship between subjective well-being and personality traits of secondary school students]. *中国临床心理学杂志, 11*(2), 96-98.

Wang, Y. X., Guo, F., Liu, Y. N., & Chen, Z. Y. (2019). 青年科技工作者的心理健康状况及影响因素[Mental health status of Chinese young scientific and technological professionals and its influencing factors]. *科技导报*, *37*(11), 35-44.

Wang, Z., Hu, S. Y., Chen, Z. Q., & Chen, Z. G. (2005). 简明抑郁症中医证候自评量表初步编制[Preliminary development of the concision depression symptoms rating scale of traditional Chinese medicine]. *中华行为医学与脑科学杂志, 14*(10), 945-947.

Wang, W. L., Lee, H. L., & Fetzer, S. J. (2006). Challenges and strategies of instrument translation. *Western Journal of Nursing Research*, *28*(3), 310-321.

Wang, C., Qian, W., Liu, C. T., Sheng, X. C., Fu, H., & Dai, J. M. (2014). PHQ-9 与 GDS-15 应用于上海市某社区中老年人抑郁评估的信效度比较[Comparison of reliability and validity between the PHQ-9 and GDS-15 in screening depression for middle aged and elderly in some community in Shanghai]. *复旦学报：医学版, 41*(2), 168-173.

Wang, J. S., Qiu, B, W., & He, E. S. (1997). 中学生抑郁量表的编制及其标准化[The development and validation of the Self-rating Scale for Depression in Adolescents]. *社会心理科学, 45*(3), 1-3.

Wang, W., Thompson, D. R., Chair, S. Y., & Hare, D. L. (2008). A psychometric evaluation of a Chinese version of the Cardiac Depression Scale. *Journal of psychosomatic research*, *65*(2), 123-129.

Wang, L. E., & Wang, J. S. (2005). 医学院校大学生具体生活事件与抑郁状态的相关分析 [Association between specific life events and depression in students]. *中国临床康复, 9*(16), 55-57.

Wang, D. R., Wang, Y. X., & Chen, Z. Y. (2019). 中国科协所属学会专职人员职业心理状况 [The occupational psychological state of the full-time staff from the China Association for Science and Technology]. *科技导报*, *37*(11), 64-70.

Wang, Z., Yuan, C. M., Huang, J., Li, Z. Z., Chen, Y., & Zhang, H. Y. (2011). 贝克抑郁量表第 2 版中文版在抑郁症患者中的信效度[Reliability and validity of the Chinese version of the Beck Depression Inventory-II among depression patients]. *中国心理卫生杂志, 25*(6), 476-480.

Ware Jr, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical care*, *34*(3), 220-233.

Watkins, D. (1989). The role of confirmatory factor analysis in cross-cultural research. *International Journal of Psychology*, *24*(6), 685-701.

Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology*, *104*(1), 3-14.

Whiston, S. (2009). *Principles and applications of assessment in counseling (3rd edition)*. Belmont, CA, US: Brooks/Cole, Cengage Learning.

WHO China Office fact sheet-depression. (2017, March). Retrieved from http://www.wpro.who.int/china/topics/mental_health/1703mentalhealthfactsheet.pdf

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94-104.

Willgerodt, M. A., Kataoka-Yahiro, M., Kim, E., & Ceria, C. (2005). Issues of instrument translation in research on Asian immigrant populations. *Journal of Professional Nursing*, *21*(4), 231-239.

Wood, N. D., Akloubou Gnonhosou, D. C., & Bowling, J. W. (2015). Combining parallel and exploratory factor analysis in identifying relationship scales in secondary data. *Marriage & Family Review*, *51*(5), 385-395.

World Health Organization. (2018, March 22). Depression. Retrieved from https://www.who.int/news-room/fact-sheets/detail/depression

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838.

Wu, L. (2016). *The invisible wound: The long-term impact of China's Cultural Revolution on trust*. Presented at the 67th annual meeting of the American Economic Association (AEA), San Francisco, CA, U.S.A.

Wu, C. Z., Chen, Z. Z., Yu, L. X., Duan, W. T., & Jiang, G. R. (2015). 抑郁、绝望对自杀意念的影响:心理痛苦的中介作用[Effects of depression and hopelessness on suicide ideation: The mediation effect of psychache]. *中国临床心理学杂志, 23*(6), 1040-1043.

Xiong, G. (2015). *简版流调中心用抑郁量表在我国青少年中的效度* [Validity of short forms of the Center for Epidemiologic Studies Depression Scale in Chinese adolescents]. (Doctoral dissertation). 湖南师范大学.

Xu, Y., Wu, H. S., & Xu, Y. F. (2007). 病人健康问卷抑郁量表(PHQ-9)在社区老年人群中的应用——信度与效度分析 [Reliability and validity of the Patient Health Questionnaire (PHQ-9) in Chinese elderly]. *上海精神医学, 19*(5), 257-259.

Yan, M. Q., Xiao, S. Y., & Hu, M. (2016). 我国一些抑郁量表的中文翻译与信效度问题 [Translation issues and reliability and validity issues in some Chinese depression scales]. *中国心理卫生杂志, 30*(7), 501-505.

Yang, L., Jia, C. X., & Qin, P. (2015). Reliability and validity of the Center for Epidemiologic Studies Depression Scale (CES-D) among suicide attempters and comparison residents in rural China. *BMC Psychiatry*, *15*(1), 15:76.

Yang, W. H., Liu, S. L., Zhou, T., Peng, F., Liu, X. M., & Li, L. (2014). 贝克抑郁量表第 2 版中文版在青少年中的信效度[Reliability and validity of the Chinese version of the Beck Depression Inventory-II in Chinese adolescents]. *中国临床心理学杂志, 22*(2), 240-245.

Yang, W. H., Wu, D. J., & Peng, F. (2012). 贝克抑郁量表第 2 版中文版在大一学生中的试用[Application of the Chinese version of the Beck Depression Inventory-II in Chinese first-year college students]. *中国临床心理学杂志, 20*(6), 762-764.

Yang, H., Yan, D. M., Li, X. B., Zhang, F. F., Ren, Y., & Wang, C. Y. (2015). 患者健康问卷抑郁量表在综合医院心身疾病门诊的应用[Application of the Patient Health Questionnaire-9 in psychosomatic disease outpatients in a general hospital]. *中华行为医学与脑科学杂志, 24*(5), 473-476.

Ye, R. F., Geng, Q. S., Chen, J., Ou, L. M., Zhang, M. L., & Dong, C. L., et al. (2013). 医院焦虑抑郁量表与 beck 抑郁问卷在综合医院门诊病人中评定抑郁的比较[Comparison of the HADS and BDI for detecting depression in general hospital outpatients]. *中国临床心理学杂志, 21*(1), 48-50.

Yen, S., Robins, C. J., & Lin, N. (2000). A cross-cultural comparison of depressive symptom manifestation: China and the United States. *Journal of Consulting and Clinical Psychology*, *68*(6), 993-999.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79-94.

Yu, D. W., & Li, X. (2000). 儿童抑郁量表 (CDI) 在中国儿童中的初步运用[Preliminary use of the Children Depression Inventory in China]. *中国心理卫生杂志*, *14*(4), 225-227.

Yu, X. F., Sun, Y. X., & Sun, Z. X. (2017). 患者健康问卷抑郁量表在颈椎病患者中的信度和效度研究[Reliability and validity of the Patient Health Questionnaire-9 (PHQ-9) in cervical spondylosis patients]. *中华临床医师杂志(电子版), 11*(6), 905-908.

Yu, X., Tam, W. W., Wong, P. T., Lam, T. H., & Stewart, S. M. (2012). The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Comprehensive Psychiatry*, *53*(1), 95-102.

Yuan, J., Liu, Y., Ding, N., & Yang, J. (2014). The regulation of induced depression during a frustrating situation: Benefits of expressive suppression in Chinese individuals. *PloS One*, *9*(5), e97420.

Zhang, M. Y. (1993). *精神科评定量表手册* [Psychiatric rating scales manual]. 湖南科学技术出版社.

Zhang, C. (2011). *青少年压力知觉、孤独感和抑郁情绪关系研究* [A study of adolescent depression and its relation with stress and loneliness] (Doctoral dissertation). 曲阜师范大学.

Zhang, Y., Jia, X. C., Fan, Z. L., Li, H. F., Zhang, W., & Han, X., et al. (2008). 山东省农村居民抑郁状况及量表评价[A study on depression prevalence as well as reliability and validity of the CES-D used in rural residents in Shandong Province]. *中国公共卫生, 24*(11), 1376-1378.

Zhang, B. S., & Li, J. (2011). 简版流调中心抑郁量表在全国成年人群中的信效度 [Reliability and validity of the Center for Epidemiologic Studies Depression Scale in national adult population]. *中国心理卫生杂志, 25*(7), 506-511.

Zhang, J., & Ma, B. Z. (2008). 氟哌噻吨美利曲辛片治疗冠心病伴心理障碍患者的临床观察 [Clinical observation on flupentixol and melitracen in the treatment of patients with coronary heart disease and mental disorders.]. *临床和实验医学杂志, 7*(2), 84-85.

Zhang, D., Shen, Y. C., & Zhang, Y. S. (1992). 老年抑郁评定量表的编制与试测[The development and preliminary testing of the Depression Rating Scale for the Elderly]. *中国心理卫生杂志, 6*(2), 53-55.

Zhang, Y., Ting, R., Lam, M., Lam, J., Nan, H., Yeung, R., ... & Sartorius, N. (2013). Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes. *Journal of Affective Disorders*, *151*(2), 660-666.

Zhang, Y., Ting, R. Z., Lam, M. H., Lam, S. P., Yeung, R. O., Nan, H., ... & Sartorius, N. (2015). Measuring depression with CES-D in Chinese patients with type 2 diabetes: the validity and its comparison to PHQ-9. *BMC psychiatry*, *15*(1), 198.

Zhang, Q., & Wang, J. (2010). Application of functional equivalence theory in English translation of Chinese idioms. *Journal of Language Teaching & Research*, *1*(6), 80-888.

Zhao, H., & Gu, X. (2016). China accreditation test for translators and interpreters (CATTI): Test review based on the language pairing of English and Chinese. *Language Testing*, *33*(3), 439-446.

Zheng, Y. P. (2009). 抑郁自评量表的编制[The development of the depression inventory]. *临床荟萃, 23*(5), 275-279.

Zheng, T., Shi, Y. Z., Zhang, N., Zhu, M. F., Li, J. J., & Wang, S., et al. (2013). 病人健康问卷-9在卒中抑郁患者中的信度和效度研究[Reliability and validity of the PHQ-9 in patients with post-stroke depression]. *北京医学, 35*(5), 352-356.

Zhou, X. L. (2012). *中韩躯体化机制的跨文化研究* [Unpacking somatization: A cross-cultural study among Chinese and South Koreans] (Doctoral dissertation). 湖南师范大学.

Zhu, Y. L., & Wang, X. J. (2011). 躯体化: 苦痛表达的文化习惯用语 [Somatization: Cultural idioms for expressing pain]. *东北大学学报 (社会科学版)*, *13*(3), 273-277.

Zou, S. (2016). *Research on the cultural equivalence in translation*. Presented at the 2016 6th International Conference on Management, Education, Information and Control (MEICI 2016), Shenyang, Liaoning, China.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and validation in social, behavioral, and health sciences*. New York, NY, US: Springer.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21-29.

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry*, *12*(1), 63-70.