VALIDITY AND RELIABILITY OF THE PATIENT HEALTH QUESTIONNAIRE-9 FOR UNIVERSITY STUDENTS

by

WEN QIAN ZHANG

B.Sc., The University of British Columbia, 2015

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

MASTER OF ARTS

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Counselling Psychology)

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:

Validity and Reliability Evidence for the Patient Health Questionnaire-9 for University Students

submitted by Wen Qian Zhang in partial fulfillment of the requirements for

the degree of Master of Arts

in Counselling Psychology

Examining Committee:

Anita M. Hubley, Measurement, Evaluation, and Research Methodology/ Counselling Psychology
Supervisor

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

William A. Borgen, Counselling Psychology
Supervisory Committee Member

Abstract

Depression is ranked as one of the most burdensome health conditions and is more prevalent in university students compared to the general population. Effective screening measures are an important aspect of detection and progress monitoring, as well as being key in depression research. However, current widely used measures, such as the Beck Depression Inventory-II (BDI-II) and the Center of Epidemiological Studies Depression Scale (CES-D), can be lengthy. For research participants who have to complete a battery of questionnaires and especially for depressed clients who are low on energy, it can be daunting to complete lengthy measures. Additionally, the BDI-II can be costly for research, counselling centers, and university clinics. Therefore, there is a need for a short, effective depression screen for both clinical and research purposes. The Patient Health Questionnaire-9 (PHQ-9) is an ideal candidate. It is only nine items, freely accessible, well established, and reflects the current diagnostic criteria. While there is an abundance of validity evidence for the PHQ-9 in primary and secondary care, there is a gap in validity evidence related to its use in the university population. To date, validity evidence for the inferences made from the PHQ-9 in the university population is limited to China, Japan and Nigeria, as well as a single secondary analysis in the U.S. The present study assessed internal structure, internal consistency reliability, and convergent and discriminant evidence to support the inferences made from the English version of the PHQ-9 as a depression screen for university students in Canada.

A total of 204 university students completed an online survey. Results supported a unidimensional structure and high internal consistency of the PHQ-9 scores. The PHQ-9 also demonstrated the expected pattern of convergent and discriminant validity coefficients with scores on depression, anxiety, mental health functioning, and physical health functioning

measures. Based on the psychometric results from this study, the PHQ-9 is tentatively recommended for use with university students.

Lay Summary

Depression is a serious mental health concern on university campuses, and efficient screening for university students is a key part of detection and monitoring. One of the most commonly used depression measures, the Patient Health Questionnaire-9 (PHQ-9), is brief, cost-effective, and based on the current diagnostic criteria. However, currently, there is little evidence for its use with university students. A group of 204 university undergraduate and graduate students from a number of study disciplines completed an online survey aimed to investigate the use of this measure with students. We confirmed that the use of the total score is appropriate, reliable, and assesses depressive symptomology. The results provided initial evidence to support the use of the PHQ-9 as a depression measure with university students in Canada.

Preface

This thesis is the original and independent work by the author, Wen Qian Zhang. Study set up and data collection was conducted by the author. Data analyses was conducted by the author, except for factor analyses, which was conducted with the help of Xuyan Tang, in consultation with Dr. Bruno Zumbo and Dr. Anita Hubley. Dr. Anita Hubley provided significant feedback and guidance throughout the study design, analysis, and writing of this research project. Ethical approval was granted by University of British Columbia's Behavioural Research Ethics Board (certificate number H19-02746).

Table of Contents

List of Tables

List of Figures

Acknowledgements

I wish to express my deep gratitude to my advisor, Dr. Anita Hubley, for her mentorship and generosity with her knowledge and time throughout this entire project. A thank you goes to my thesis committee member, Dr. Bruno Zumbo, for his guidance, insight, and questions that challenged me to push my limits. Thank you to Dr. Bill Borgen for joining the committee and providing valuable support and steadiness during my practicum experience.

Many thanks for Xuyan, Raly, Heejin, and many other friends who supported, distracted, and encouraged me throughout this journey. Thank you for keeping me going, I feel very blessed to have you in my life.

And finally, to my parents. Their unending amount of patience and support make this possible. Thank you.

Chapter 1: Overview of Research Study

Introduction

Depression is a highly salient issue in university students due to the higher level of prevalence compared to the general population and its debilitating effects. In a survey of the depression prevalence research, Ibrahim and colleagues (2013) reported that prevalence in university students ranged from 10% to 85%. University students are at the age where the first episode of depression often occurs, between the ages of 15-24 (Blazer, Kessler, McGonagle, & Swartz, 1994). University students are in a transitional development stage marked by identity development, which can lead to lower self-esteem and reduced social support. This impact is compounded by the sharp increase in stress associated with being in higher education (e.g., relationship challenges, academic stress, changes in living arrangements), as well as other factors such as reduced sleep. Depression in university students has been linked to a plethora of negative consequences including relationship instability (Whitton & Whisman, 2010), lower self-esteem (Conti, Adams, & Kisler, 2014), lower work performance (Harvey et al., 2011), self-medication (Ford & Schroeder, 2009), and suicide (Furr et al., 2001). In the long run, depression early in life can lead to a cascade of negative outcomes through its impact on educational attainment, financial and career outcomes, and social relationships. Thus, early detection and effective treatment is vital. Universities are well positioned to provide prevention and treatment since they already encompass many aspects of students' lives, such as social networks, residence, health care, and academics (Mowbray et al., 2006).

Quickly administered and effective depression screening measures for university students can assist mental health clinicians and health care professionals to accurately identify depression, which can facilitate more immediate treatment, as well as to monitor progress. Two of the most

commonly used depression screening tools with the university population are the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996) and the Center for Epidemiologic Studies – Depression scale (CES-D; Radloff, 1977), which are moderately long at 21 items and 20 items, respectively. Lengthy measures add additional burden for depressed clients in clinics and can be daunting for research participants who have to complete a battery of questionnaires. Furthermore, the BDI-II can be costly for university counselling centers and clinics as it is not freely available for use.

The Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001) has been developed specifically for use as a depression screening tool in fast paced clinics. At 9 items, it is only half the length of the BDI-II and the CES-D. It can be completed in a few minutes and scored rapidly, which reduces the burden on respondents and lessens time constraints for busy clinics. It is also freely available, which further eases the financial burden on institutions with limited resources. Finally, it is well established and reflects the diagnostic criteria for depression based on the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5; American Psychiatric Association, 2013). It has become a popular measure for use with university students for both research and clinical purposes. Since its publication, validity evidence for use of the PHQ-9 has been studied in many medical populations in both primary and secondary care settings. However, validity evidence for the inferences made from the PHQ-9 in the university population is scarce. To date, translated versions of the PHQ-9 have been studied with university students in China, Japan, and Nigeria, but validity evidence for the English version for use with university students is limited to a single secondary analysis in the U.S. There has yet to be a validity study that is purposefully designed to provide evidence to support the use of the English version of the PHQ-9 with university students.

Purpose of the Study

This study will examine the psychometric properties of the PHQ-9 and assess the validity evidence for the inferences made from the English version of the PHQ-9 as a depression screen for university students in Canada. To provide evidence to support its use, this study will examine: (1) internal structure by testing a number of measurement models of the PHQ-9 identified in the literature, (2) reliability of the PHQ-9 scores, and (3) convergent and discriminant validity evidence of the PHQ-9 through correlations with a measure of the same construct (i.e., depression) and correlations with measures of more or less related constructs (i.e., anxiety, mental health, physical health). Validation of the inferences made from this study will provide evidence to support use of the English version PHQ-9 in research and clinical practice with university students.

Overview

The following thesis provide a literature review of depression, including depression in university students, as well as a review of commonly used self-report depression measures with university students. This is followed by a review of sources of validity evidence and methods of estimating reliability. The literature review concludes with past validation studies of the PHQ-9 as well as other depression measures, as well the proposed study and hypotheses. This review is followed by a manuscript chapter, which includes a brief introduction followed by an outline of the research design including participant recruitment, research procedure, ethical considerations, measures used, and data analyses. The results of the study are included in the manuscript, followed by a discussion of numerical results, study strengths and limitations, and future directions. Finally, the conclusion chapter summarizes the study findings, identifies areas for future research, and discusses counselling implications.

Chapter 2: Literature Review

Introduction to Literature Review

A literature review is conducted to provide information regarding the topic of depression and measurement of depression in university students, as well as to guide the research methodology. I begin broadly with an overview of depression, leading to a focus on depression in university students. An overview of commonly used depression measures for university students will be presented, followed by a summary of the current literature on conducting validity studies, as well as calculating reliability. This will guide our understanding of validity and provide the foundation for understanding validation studies. Next, I will review validation studies of the PHQ-9. Finally, I will examine validation studies conducted with other depression measures commonly used with university students to inform the method used in the proposed study.

Overview of Depression

Symptoms and diagnostic criteria. Depression is generally characterized by a lack of positive affect, accompanied by a range of emotional, cognitive, behavioural, and physical symptoms (NICE, 2009). Symptoms can manifest in a variety of forms, including sleep disturbance, poor concentration, and lack of self-care or interest (NICE, 2009). There are several forms of depressive disorders, and the two most common are major depressive disorder (MDD) and dysthymic disorder or persistent depressive disorder (PDD) (Doris, Ebmeier, & Shajahan, 1999; Uher, Payne, Pavlova & Perilis, 2014). MDD is identified by its increased persistence, intensity, and impairment to daily life, but distinguishing MDD and PDD remains ambiguous (Lewinsohn et al., 2000).

Two widely used diagnostic systems are the International Statistical Classification of Diseases and Related Health Problems-10th Edition (ICD-10), produced by the World Health Organization (WHO) and the Diagnostic and Statistical Manual of Mental Disorders-5th Edition (DSM-5), produced by the American Psychiatric Association (APA). The ICD is a classification system for all diseases, with a section concerning psychiatric disorders, called "Mental and Behavioural Disorders." It is the official world classification system and is intended for all health practitioners, with a focus on primary care in low- and middle-income countries (Tyrer, 2014). The DSM-5, on the other hand, is a U.S. classification system, commonly used in North America and Australia, and primarily used by psychiatrists (Tyrer, 2014).

MDD in the DSM-5 is defined by having one or more major depressive episodes (MDE) that are not explained by a psychotic disorder or attributed to substance use or medication (Uher et al., 2014). As well, symptoms must cause significant impairment and distress in daily functioning, and there must be a lifetime absence of mania or hypomania. Having five out of the nine symptoms over the same two-week period meets the criteria of having an MDE. The nine symptoms include depressed mood, loss of interest or pleasure, change in weight or appetite, insomnia or hypersomnia, psychomotor retardation or agitation, loss of energy or fatigue, worthlessness or guilt, impaired concentration or indecisiveness, and thoughts of death or suicidal ideation or attempt. Additionally, one of these symptoms must be depressed mood or anhedonia. In the DSM-5, MDE can be further specified based on the number of symptoms and the level of impairment into mild, moderate, or severe.

PDD is a new diagnosis in the DSM-5, with criteria similar to dysthymic disorder from the DSM-IV (Uher et al., 2014). It is another common form of depressive disorder and possesses less severe and disabling symptoms (Doris et al., 1999). In the DSM-5, it is not clear if PDD and

MDD can be concurrent if both criteria are fulfilled, given that they are not listed as exclusion criteria for each other (Uher et al., 2014). PDD is a less severe form of depression that is diagnosed with the presence of depressed mood for most days, lasting two or more years. Additionally, two of the five symptoms need to be present over the same period. They include poor appetite or overeating, insomnia or hypersomnia, low energy or fatigue, low self-esteem, impaired concentration or indecisiveness, and hopelessness.

Unlike the DSM, which has a single manual, the ICD has two manuals. First, the Clinical Description and Diagnostic Guideline is designed for use in clinical settings, with descriptive definitions that are not operationalized (Lopez Ibor, Frances, & Jones, 1994). The second is the Research Criteria, which is intended for research purposes, and provides more detailed and operationalized formula, much like in the DSM (Lopez Ibor, Frances, & Jones, 1994). The ICD-10 lists 10 symptoms that have considerable overlap with the DSM. A small difference is that the ICD-10 lists loss of confidence or self-esteem and inappropriate guilt as two separate symptoms, while the DSM-5 combines them into a single symptom: worthlessness/ excessive or inappropriate guilt. Similar to the DSM-5, the ICD-10 requires the symptoms to be present for at least two weeks and result in an impairment of functioning. The ICD-10 classifies clinically significant depressive episodes as mild, moderate, or severe based on the number, severity, and type of symptoms present. A mild depressive episode requires four symptoms, moderate requires six symptoms, and severe requires eight or more symptoms. Additionally, for mild and moderate episodes, at least two of symptoms must be depressed mood, loss of interest in everyday activities, or reduction in energy, and all three of the symptoms must be present for severe episode. The ICD-10 also specifies the appropriate clinical sites for treatment depending on the

severity. Mild depressive episodes are suitable for primary care whereas moderate or severe depressive episodes are to be addressed in psychiatric settings.

The counterpart to PDD in the ICD-10 is dysthymic disorder. Dysthymic disorder is characterized by depressed mood lasting at least two years with no episodes of hypomania. In addition, three of the 10 listed symptoms must also be present. Unlike the DSM-5, the ICD-10 specifies that meeting the diagnostic criteria for mild depressive disorder is an exclusion criterion for dysthymic disorder.

Etiology. The exact cause of depression remains unknown, but both biological and environmental factors are thought to be involved. Biological influences associated with depression include genetics (Kendler et al., 2001), immune system abnormalities (Hoseinzadeh et al., 2016), and neurotransmitters (Werner & Covenas, 2010). Environmental factors focus on stressful life experiences such as adverse childhood events (Salokangas, From, Luutonen, & Hietala, 2018), current life circumstances (Cronkite et al., 1998), and other health problems (NICE, 2009). The diathesis-stress model suggests that depression is the result of existing vulnerability or diathesis combined with stressful life events (Monroe & Simons, 1991). Similarly, the biopsychosocial model also combines biological, social, and psychological factors in the cause of depression (Schotte, Bossche, Doncker, & Claes, 2006).

Impact. Depression is the fourth most disabling medical condition worldwide and is projected to be the second leading cause of disability by 2020, only behind ischemic heart disease (WHO, 2002). It affects between 10% to 25% of women and 5% to 12% of men (Nihalani et al., 2006). Apart from the negative subjective experience of depression, it also impacts physical health, social and occupational functioning, and life expectancy. Physical ailments coupled with depression adversely affect overall health outcome and increases the risk

of death compared to those without depression (Moussavi et al., 2007). Depression also exacerbates the experience of pain, distress, and disability associated with physical illnesses. It can reduce a person's ability to work effectively, and potentially lead to loss in income (Kessler et al., 2006). Aside from the financial cost of health services, the indirect cost of depression includes loss of employment (Thomas & Morris, 2003), reduced productivity (Lerner et al., 2004), and impact on quality of life (Lepine & Briley, 2011). Depression can also impair a person's social functioning through isolation, hindering his or her ability to communicate and sustain relationships (Judd et al., 2000). Another major risk associated with depression is suicide, with lifetime risk estimated to be around 15% (Möller, 2003).

Treatment. The most commonly used treatment options are antidepressants and psychotherapy (Friedman, Anderson, Arnone, & Denko, 2014). The development of antidepressants is based on the monoamine hypothesis of depression, which suggests that deficiency of monoamine neurotransmitters causes depression (Delgado, 2000). This hypothesized pathology is supported by effectiveness of antidepressants that target and elevate different classes of neurotransmitters, including serotonin, norepinephrine, and dopamine (Friedman et al., 2014). Selective serotonin reuptake inhibitors (SSRIs) are generally considered to be the first in line treatment, due to their tolerability. Other classes of antidepressants include the serotonin and norepinephrine reuptake inhibitor (SNRI), tricyclic antidepressants, and monoamine oxidase inhibitor (MAOI). Current evidence supports the use of antidepressants across the severity of depression, with little difference in efficacy between classes of medication (Friedman et al., 2014). The main considerations when choosing medication are side effect profile and tolerability (Friedman et al., 2014).

There are also a number of psychotherapy modalities aimed at reducing depressive symptoms. Two well-established and researched therapies include cognitive behavioural therapy (CBT) and interpersonal therapy (IPT) (Friedman et al., 2014). CBT targets negative thinking and behavioural styles and patterns that contribute to depression, whereas IPT aims to facilitate understanding and to work through difficult relationships that contribute to depression. Two meta-analyses of CBT and IPT in treatment of depression found effect sizes of 0.75 and 0.63, respectively, suggesting medium to large effect sizes (Cuijpers et al., 2016; Cuijpers et al., 2011).

It is difficult to predict individual response to a treatment option. Antidepressants and psychotherapies are considered generally equal in efficacy, though those with chronic severe depression may benefit more from a combination of the two (Friedman et al., 2014). Studies have tried to identify predictors of treatment outcomes, with mixed results (Ezquiaga et al., 1998; Meyers et al., 2002). Comorbidity with personality disorders, previous depressive episodes, and some social factors such as lack of social support have been associated with negative outcomes (Ezquiaga et al., 1988). Meyers and colleagues (2002) reported less depression severity, female sex, and being married as significant predictors of recovery.

Eisenberg and colleagues (2011) studied the treatment seeking behaviours of college students across the U.S. The results showed that approximately one third of students with mental health issues received treatment, which is comparable with the general adult population (Wang et al., 2005). The prevalence of psychotherapy and medication use was found to be approximately equal in college students in the U.S. (Eisenberg, Hunt, Speer, & Zivin, 2011), whereas the general population received more medication than psychotherapy for mood disorders (Olfson et al., 2002). The authors attributed this difference to the strong presence and availability of

counselling centers on college campuses. In another study on depression treatments in college, however, Apfel (2004) found a higher preference for psychotherapy compared to medication.

Overview of Depression in University Students

Epidemiology. Depression can occur throughout the life span, but the first episode often occurs in early childhood or adolescence. University students are at the peak period of depression onset, particularly for first episodes. The age group most likely to have a first MDE is between 15 and 24 (Blazer, Kessler, McGonagle, & Swartz, 1994), which includes the typical age of undergraduate students. In a survey of existing research on depression prevalence in university students, Ibrahim and colleagues (2013) found prevalence rates ranged from 10% to 85%, with a weighted mean prevalence of 30.6%. Results varied depending on methods of assessment, geographical location, and demographic factors, but prevalence is still considerably higher than the 9.0% found in the general adult population in the U.S. (CDC, 2010). Moreover, in a study of over 1400 students, Furr et al. (2001) reported that 53% of students labeled themselves as being depressed and 9% reported suicidal ideation. Many studies, but not all, reported a sex difference in university students, with statistically significantly higher prevalence found among female students compared to male students (Ibrahim et al., 2013). For instance, Roberts, Glod, Kim and Hounchell (2010) found prevalence rate of 25% and 17% in female and male students, respectively. In another study, female students were more likely to have moderate to severe depression (18%) compared to male students (9%) (Schwenk, Davis, & Wimsatt, 2010). Given the high rates of depression found in this particular population, Ibrahim et al. (2013) called for the validation of commonly used depression measures in the student population specifically.

Contributing factors in university. Young adult university students are in a phase

described as 'emerging adulthood', which is a transitional developmental stage between adolescence and adulthood (Arnett, 2004). This transitional developmental stage is considered stress arousing and anxiety provoking (Meadows, Brown, & Elder, 2006). It is also a period of identity development, which can be difficult and lead to lowered self-esteem and withdrawal from social support. Indeed, gender differences in depression sharply widen from childhood to adolescence, from no difference (Twenge & Nolen-Hoeksema, 2002) to a 2:1 of females to males (Nolen-Hoeksema & Girgus, 1994).

Depression has been consistently associated with increased perceived stress, and college is often considered a time of increased pressure with 76% of students reporting feeling overwhelmed, and 22% indicating difficulty with daily functioning due to stress and depression (American College Health Association, 2011). Higher depression has been specifically associated with higher levels of college stress (Dyson & Renk, 2006; MacGeorge et al., 2005), and female students also report higher levels of stress (Matud, 2004). Ross, Niebling and Heckert (1999) attempted to define college stress, which included adjusting to college life, interpersonal relationship challenges, academic pressure, changes in lifestyle, and living arrangements. Chronic sleep loss, often the result of high academic pressure, caffeine consumption, and social media use in university students also intensify the risk of depression (Owens, 2014). Zhang and colleagues (2018) reported that, in college students, the relationship between poor sleep quality and mood disorders, including depression, is mediated by perceived stress. In addition to college-related stress, family life stress is predictive of depressive symptoms as well. Loneliness and lack of familial support can interact and exacerbate the experience of stress, leading to depression (Wei, Russul, & Zakalik, 2005).

Symptomology in university students. Despite a higher prevalence rate of depression in

university students, depression has not been investigated extensively in this population (Ceyhan, Ceyhan, & Kurty, 2009). The DSM-5 does not distinguish between depression in university students and the general population. Previous studies, however, have indicated that there may be slight differences in symptom manifestation and severity in the depressive experience with age (Cox, Enns, Border, & Parker, 1999; Cox, Enns, & Larsen, 2001). Younger people tend to have more behavioural symptoms, while older adults have more complaints of somatic symptoms and fewer complaints of low moods. Students may experience more cognitive symptoms (Cox et al., 1999, Whisman, Perez, & Ramel, 2000), including perfectionist ideation, worthlessness, and low self-esteem (Vredenbrug et al., 1988), and experience difficulty in concentration, pessimism, and self-blame (Cox et al., 1999). Furthermore, somatic symptoms, such as change in sleep and appetite, can often be attributed to factors other than depression, such as social and academic schedule (Kitamura, Hirano, Chen, & Hirata, 2004; Smith Rosenstein, & Granaas, 2001). Vredenburg et al. (1988) found that students' symptoms tend to be less severe, but chronic.

Impact. Depression has been linked to relationship instability (Whitton & Whisman, 2010), lower self-esteem (Conti, Adams, & Kisler, 2014), and lower work performance (Harvey et al., 2011). Hysenbegasi and colleagues (2005) studied the impact of depression on academic outcome in university students in the U.S. and found that depression was associated with a 0.49-point decrease in grade point average (GPA). Depression has also been linked to other major concerning behaviours. Individuals with mental illness often resort to self-medication as a coping mechanism (Ford & Schroeder, 2009). In a large nationwide study, Zullig and Divin (2012) explored the associations among depression, suicidality, and substance use in U.S. college students. Their findings demonstrated that there is an increased likelihood of using non-prescription opioids, stimulants, sedatives, and anti-depressants in those who report being

depressed. Moreover, depression has repeatedly been identified as a major risk factor for suicidality in the college population (Furr et al., 2001; Garlow et al., 2008). Suicide is the 11th leading cause of death in the U.S. (CDC, 2010), but is the third leading cause of death among college students (Suicide Prevention Resource Center, 2004).

Depression in early adulthood can have an accumulation of negative consequences in adult life through its impact on educational attainment, financial and career outcomes, and social relationships. Thus, early detection and effective treatment in university can have a cascade of positive downstream effects. Universities are well positioned for prevention and treatment of mental disorders, including depression, because they already encompass many aspects of students' lives such as social networks, residence, health care, and academics (Mowbray et al., 2006).

Summary. Given the evidence of high rates of depressive symptoms among university students, screening for depression in university counselling centers has become a priority (Erdur-Baker, Aberson, Barrow, & Draper, 2006; Furr et al., 2001). Efficient screening that is sustainable is particularly important for institutions that are limited in resources. Additionally, depression screening is also important in university health centers. Shepardson and Funderburk (2014) reported that significant proportions of university students visit their university health center for non-mental health reasons but have undetected mental health concerns. They recommended standardized, self-report screening measures as a way to facilitate dialogues between care providers and students (Shepardon & Funderburk, 2014). Compared with interview-based methods, self-reports are associated with lower concerns of social desirability and increased willingness to disclose sensitive information (Bowling, 2005). Also, self-reports are efficient and can be completed without interrupting the normal pace within a health clinic

(Brown, 2011). The next section will provide an overview of depression screening tools that are commonly used with university students.

Self-Report Depression Inventories

A review of the literature has revealed several depression screening measures designed for the general population that are often used in the context of depression research with college or university students (Ibrahim et al., 2013). The Beck Depression Inventory (BDI), and its revision, the BDI-II, stand as the most frequently used depression screens, found in around half of the studies in this review (Wang & Gorenstein, 2013). Second is the CES-D, followed closely by the PHQ-9. This review also identified a relatively new depression measure designed for university students, the University Student Depression Inventory (USDI) (Khawaja & Bryden, 2006). A review of each measure, as well its strengths and weaknesses will be presented below.

Beck Depression Inventory. The BDI was designed to measure depression symptoms and severity in those aged 13 and above (Beck et al., 1996). The original BDI was developed in 1961 to assess depression severity in those who are clinically depressed. It has since gone through multiple revisions. The first revision, BDI-IA, was published in 1987 and is commonly referred to in the literature as the 'BDI'. This revision is similar to the original, with some items reworded to improve the ease of use. The BDI-IA was criticized for reflecting only six out of the nine DSM-II criteria for depression. The BDI and the BDI-IA are still freely available, but not the current, most up-to-date version, the BDI-II. The BDI-II was published in 1996 and contained substantial revisions. It contains 21 items, each with four statements of symptoms scored from 0 (not at all) to 3 (severely). There are no reverse scored items. The summed total score can range from 0 to 63, with higher scores indicating greater depression severity. The second revision corresponds better to the DSM-IV and DSM-5 criteria for depressive disorders. Revisions

included omission of items relating to weight loss, body image, hypochondria, and work difficulty, which were replaced by new items on agitation, worthlessness, concentration difficulty and loss of energy (Beck et al., 1996). Additionally, the time frame assessed in the measure was extended to two weeks from one week in the original BDI.

Since first published, the psychometric properties of the BDI-II have been studied in a range of clinical and non-clinical populations (Erford, Johnson, & Bardoshi, 2016). In this review, the BDI-II was found to be the most well studied for its use and interpretation with university students. It is noted that the majority of these studies relied on university students as convenience samples, predominantly using students in psychology courses.

In their original validity study, Beck and Steer (1996) administered the BDI-II to 120 undergraduate students and 500 psychiatric outpatients and reported a two-factor structure. The student sample of 120 is generally considered insufficient for factor analysis. However, further studies also supported the two-factor structure when administered with university students (Dozois, Bobson, and Ahnberg, 1998; Whisman et al., 2000; Contreras et al., 2004; Storch, 2004). Still, other studies have reported different factor structures. Osman and colleagues (1997) and Carmody (2005) both administered the BDI-II to students in university psychology courses and reported a 3-factor structure.

The BDI-II has also been frequently compared to measures of related constructs to provide convergent and discriminant validity evidence. When administered with university students, it demonstrated high correlations with other measures of depression, including the CES-D (0.71-0.86) (Lipps et al., 2007; Shean & Baldwin, 2008), the depression subscale of the Depression Anxiety Stress Scale (DASS) (0.77) (Osman et al., 1997), and the depression

subscale of the State Trait Anxiety Inventory-Trait version (STAI-T)[1] (0.76) (Storch, 2004). It also demonstrated comparatively lower correlations with measures of anxiety, including: Beck Anxiety Inventory (BAI) (0.56-0.62) (Osman et al., 2007; Contreas et al., 2007), and the anxiety subscale of the STAI (0.69) (Storch, 2004). Anxiety and depression are related constructs, given that they share some overlapping symptoms. So, it is expected that correlations with anxiety measures will be moderate to strong but relatively lower than those with other depression measures.

Most of the validity studies reported reliability estimates for the BDI-II. The majority of the studies reported a single Cronbach's alpha (0.83 - 0.91) (e.g. Dezois et al., 1998; Contreas et al., 2004), indicating a good level of reliability despite the two-factor structure. Other studies reported individual Cronbach's alphas for each factor. Osman et al. (1997) studied the BDI-II in a sample of 137 students from university psychology courses and reported 3 factors: negative attitude ($\alpha = 0.84$), performance difficulty ($\alpha = 0.77$), and somatic element ($\alpha = 0.68$). It is noted that 137 is generally not an adequate sample size to conduct a factor analysis. Storch (2004) also studied the BDI-II in a sample of 414 psychology students, but reported a two-factor structure: cognitive-affective ($\alpha = 0.87$) and somatic ($\alpha = 0.74$). The reliability of the BDI-II is also supported by high test-retest reliability (0.96), with an interval of administration ranging between 1-12 days and a mean of 3 days (Sprinkles et al., 2002).

Finally, BDI-II scores have also been compared with decisions from diagnostic interviews to provide test-criterion validity evidence in university students. This is a less frequently studied form of validity evidence with university students. Sprinkles et al., (2012) administered the BDI-II along with the major depressive episode portion of the Structured

[1] The depression subscale of the STAI-T is based factor analysis by Bieling et al. (1998), which identified STAI-T having two distinct factors assessing anxiety and depression.

Clinical Interview for DSM–IV Axis I Disorders (SCID-I), and reported a strong, positive correlation (0.83).

The BDI and the BDI-II are the most widely used depression screens in depression research with college students (Ibrahim et al., 2013). They are simple in terms of administration and scoring, and the BDI-II corresponds to current DSM criteria. A notable limitation of using the BDI-II is that a fee must be paid for each copy used. Another drawback of the BDI-II is its length. Each of the 21-items contains four statements scored from 0 to 3 and, as a consequence, answering the scale requires reading a total of 84 statements. Reading 84 statements can be burdensome to those experiencing depression, and is inefficient in research studies with multiple measures.

Center for Epidemiological Studies - Depression Scale. The CES-D was designed to measure the level of depression in the general population, by assessing mood and level of functioning during the past week (Radloff, 1977). The original version contained 20 items with a 4-point Likert-type response format ranging from 0 (rarely or none of the time) to 3 (most or almost all the time). There are four reverse-scored items, and the resulting total score can range from 0-60, with higher scores reflecting greater symptoms of depression. There is also a modified version available for children (CES-D for Children) as well as a 10-item version for older adults (Irwin, Artin, & Oxman, 1999).

In the initial validation of the measure, Radloff (1977) reported that the measure possessed high internal consistency, finding coefficient alphas of 0.85 and 0.90 for community and psychiatric samples, respectively. The validity of interpreting and using the CES-D has been studied in a wide range of populations, including college students. Gloria and colleagues (2012) studied the utility of depression screens, including the CES-D, with Latina/o undergraduate

students. They reported a high internal consistency reliability alpha for the CES-D (0.88), and evidence for convergent validity with the BDI-II (0.75) (Gloria, Castellanos, Kanagui-Munoz, & Rico, 2012). In a sample of 690 Jamaican university students, Lipps, Lowe, and Young (2007) also reported that the CES-D possessed good internal consistency (0.89) and a strong correlation with the BDI-II (0.71). Segal and colleagues (2008) administered the CES-D and the BDI-II to a sample of community dwelling older adults and university students and provided further convergent evidence (0.92), as well as high internal consistency for CES-D (0.92). In a study of 395 college students, the CES-D and BDI-II were found to have good and comparable criterion validity (Shean & Baldwin, 2008).

The CES-D is freely available and is commonly used to measure depression in research studies. In a review of the depression literature in college students, the CES-D was the second most commonly used scale (Ibrahim et al., 2013). An important consideration in choosing the CES-D in research is that it focuses on affective states, and the items are not based on diagnostic criteria. Increased appetite or sleep anhedonia, psychomotor agitation or retardation, guilt and suicidal thoughts are not assessed. Additionally, Orme, Reis and Herz (1986) found that the CES-D correlated highly with trait anxiety as measured by Spielberg's State-Trait Anxiety Inventory (0.71), and the authors raised the concern that the CES-D measured predisposition for anxiousness as well.

Patient Health Questionnaire-9. The PHQ-9 was originally designed to detect and measure depression and severity in primary care settings. The nine items on the scale correspond with the nine symptoms in the DSM-5 diagnostic criteria. Each item is scored from 0 to 3, based on the frequency of each symptom, from 0 (not at all) to 3 (nearly every day). The total score ranges from 0 to 27. The PHQ-9 can be used to identify cases of depressive disorders, as well as

measure severity. There are no reverse scored items, and higher scores indicate higher levels of depression. The original validation study for the PHQ-9 included a large sample of primary care and obstetrics/ gynecology patients (Spitzer et al., 1999). There has been an accumulation of validity evidence related to its use in a wide variety of medical populations as well as general populations including adolescents and older adults.

Since its publication, the PHQ-9 has been extensively studied for use in a large number of specialized populations, including patients in primary and secondary care, as well as community samples. Study of its use and interpretation in university students is limited. Adewuya, Ola, and Afolabi (2006) studied the utility of the PHQ-9 in a sample of 512 Nigerian university students from Obafemi Awolowo University in Nigeria. The authors reported good internal consistency (Cronbach's alpha = 0.85), good one-month test-retest reliability (r = 0.89), as well as test-criterion related evidence using the MINI as the 'gold standard"' diagnosis interview (Adewuyua, Ola, & Afola, 2006). This study finding supported the use of 10 as the cutoff score, with sensitivity of 0.85 and specificity of 0.99. However, the study also found a relatively lower correlation with the BDI, r = 0.67 (Adewuyua, et al., 2006). Zhang and colleagues (2013) conducted a similar validity study with 959 Chinese university students, using the SCID as the criterion measure. This study also reported good internal consistency (Cronbach's alpha = 0.85), good test-retest reliability (r = 0.87) when re-administered within four weeks, but a higher correlation between the PHQ-9 and the BDI (r = 0.79) (Zhang et al., 2013). In another validation study with Chinese university students, Du and colleagues (2017) reported comparable, but slightly lower, Cronbach's alpha (0.80) and two-week test-retest reliability (0.78). Using the MINI as the criterion, the study reported an optimal cut-off score of 10, with a sensitivity of 0.74 and specificity of 0.85. Sensitivity of 0.74 is considered quite low for a depression inventory.

However, using a cutoff score of 9 increased the sensitivity to 0.89, but lowered the specificity to 0.79 (Du et al., 2017). In a Japanese university sample, Umegaki and Todo (2016) used an item response theory model to compare three depression scales: the Zung Self-Rating Depression Scale (SDS), the CES-D, and the PHQ-9. The authors found the PHQ-9 performed better compared to the other depression measures, likely due to the absence of negatively worded items. This review identified one PHQ-9 validity study with university students that was administered in English (Keum, Miller, & Inkelas, 2018). The authors conducted a secondary data analyses based on data from a larger national survey of undergraduate students, and provided support for a one factor structure, as proposed by Kroenke et al (2001). Keum and colleagues (2018) also found evidence for good reliability (Cronbach's alpha = 0.89), and an adequate pattern of convergent and discriminant evidence with alcohol use and overall mental health.

The PHQ-9 is easy to administer and to interpret. It is less than half the length of other popular depression scales such as the BDI-II or the CES-D. So, it is more time efficient, which is critical in busy clinics. Additionally, the instrument is freely available, making it more cost effective and easier to access for clinical and research purposes. Despite becoming frequently used in depression research, including in university students, validity evidence to support the interpretation and use of scores in the North American higher education context remains scarce.

University Student Depression Inventory. The USDI is a 30-item scale designed specifically for assessing depression in university students (Khawaja & Bryden, 2006) and is the only self-rated depression measure specific to this population to date. The response format is a 5-point Likert-type format, ranging from 1 (not at all) to 5 (all the time), producing a total score range between 30 and 150, with higher scores indicating higher levels of depression.

Khawaja and Bryden (2006), in a study on the development and validation of the USDI with a sample of 322 Australian university students, aimed to design a depression scale that reflected symptoms commonly seen in student depression. The authors noted that most clinical depression scales contain items on somatic symptoms, such as changes in sleep and appetite, but these symptoms are not necessarily indicators of depression in university students. Instead, the USDI focuses on student experiences, with an emphasis on cognitive and affective symptoms. It is also the first depression scale to include items addressing academic motivation. As a result, this measure is not reflective of the DSM-5 diagnostic criteria.

The initial validation study for the USDI reported three factors: lethargy, cognitive/ emotional, and academic motivation (Khawaja & Bryden, 2006). The authors also reported a good reliability estimate using Cronbach's alpha for the total USDI (0.95), as well as for each of the three subscales: 0.89 for lethargy, 0.92 for cognitive/ emotional, and 0.84 for academic motivation (Khawaja & Bryden, 2006). The USDI correlated highest with the depression scale of Depression Anxiety Stress Scale (DASS) (r=0.76), followed by the Stress Scale (r=0.62), and the Anxiety Scale (r=0.56) (Khawaja & Bryden, 2006). Subsequent research with Australian and Iranian students supported the three-factor first-order and a one-factor second-order structure (Habibi, Khawaja, Moradi, Dehghani, & Fadaei, 2014; Romaniuk & Khawaja, 2013).

Romaniuk and Khawaja (2013) used z scores to determine cut scores for corresponding labels of ‘low’, ‘moderate’, ‘high’, and ‘very high’. However, these labels do not represent severity of depression. To date, there is no test-criterion-related validity evidence for the USDI, so its clinical utility is very limited. Additionally, at 30 items, the USDI is longer than most commonly used depression scales. There remains the need for a brief, efficient screen for depression in students in primary care settings and counselling centers.

Summary. This review identified three depression measures that are commonly used with university students: BDI-II, CES-D, and PHQ-9, as well as a depression measure designed for this population, the USDI. The PHQ-9 is the only measure that is only 9 items, less than half of the length of other measures. Unlike the CES-D and the USDI, the PHQ-9 is reflective of the diagnostic criteria, and, unlike the BDI-II, it is freely available. These qualities make the PHQ-9 an ideal candidate for measuring depression for research purposes and in fast-paced clinics with financial constraints. While the PHQ-9 is supported by validity evidence in medical settings, evidence for its use with university students in North American context is scarce.

Sources of Validity Evidence

This section will provide an overview of the literature on conducting validity studies. It will focus on the guidelines as described by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), which outlines five sources of validity evidences. This review will also assist in guiding the approaches and rationale for selecting the specific procedures used in my proposed study.

Validity is key in the interpretations and the decisions made based on test scores. When strong validity evidence exists to support the interpretations of scores based on measures, there is increased confidence our screening and research decisions. Alternatively, with low levels of validity, decisions may be misinformed or even harmful. The concept of validity has evolved over the past century. The traditional, trinitarian view of validity suggested that there were three types of validity: content, criterion, and construct. Since the late 1970s and 1980s, a more contemporary perspective, the unified view of validity, has gained traction. The unified perspective, proposed by Messick (e.g., Messick, 1989), emphasizes that construct validity is all of validity, and has been endorsed by the *Standards for Educational and Psychological Testing*

(AERA et al., 2014). According to the *Standards* (AERA et al., 2014), validity is "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (p. 11). The process of validation involves providing relevant, empirical evidence to support the proposed score interpretation and use (AERA et al., 2014).

Hence, validity is not an inherent characteristic of the measure. That is, the measure itself is neither valid nor invalid. Rather, validity concerns the interpretation and use of the scores, given the sample and the context. Hence, validity cannot be separated from the sample and context in which the test is administered (Zumbo, 2009). Additionally, validity is not an all or none phenomenon. Rather, it can be considered on a continuum ranging from strong to weak, taking into consideration the degree to which the current body of evidence supports the intended inference and use of the test scores. Finally, evaluation of validity is a continuous, ongoing process. Messick (1989, p. 13) emphasized, “[e]vidence is always incomplete.” Because values, social norms, language, and theories shift over time, it is important to continuously accumulate validity evidence for the intended inferences (Hubley & Zumbo, 2011). The 2014 *Standards* (AERA et al., 2014) outlines five sources of validity evidence: test content, internal structure, response processes, relations to other variables, and consequences of test use. They are described below.

Test content. Haynes, Richards, and Kubany (1995) defined content validity as "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose" (p. 238). That is, for a test to be interpreted and used as a measure of a construct, the test content, including test items, response format, and instructions, should reflect the key aspects of the intended construct. According to the *Standards* (AERA et al., 2014), test content includes themes, wording, and format of the

items in the test, as well as the administration and scoring. Content validation can be either quantitative or qualitative (Haynes et al., 1995). The quantitative approach relies on rating scales to evaluate aspects of the test such as the relevance of the test items, and qualitative methods can provide additional feedback from evaluators, such as recommended rewording (Haynes et al., 1995).

In providing a framework to assess test content, Sireci (1998) described four elements of content validity: domain definition, domain representation, domain relevance, and appropriateness of test construction procedures. First, domain definition specifies the concrete details regarding what is being measured. Domain representation then addresses the degree to which the domain is appropriately represented in the test. Third, domain relevance examines the degree of relevance of each item to the domain. Finally, appropriateness of the test development process refers to the process of test construction that minimizes construct irrelevant variance and construct underrepresentation. There are a number of methods used to evaluate validity evidence based on test content, most of which require people with expertise on the subject, called subject matter experts (SMEs) (Sireci & Faulkner-Bond, 2014). For a depression inventory, SMEs can include clinicians who work extensively with patients with depression and researchers who specialize in this topic. In these studies, SMEs can match test items to the intended content, rate the degree to which the item adequately represents the intended content and cognitive specification, or rate the degree to which items are relevant to the domain tested (Sireci & Faulkner-Bond, 2014). In addition to SMEs, experiential experts (EEs) can also provide judgements about the test content. EEs are members of the target population that the instrument is designed for, and can provide judgements, for example, on the clarity of the language of the test items (Tilden, Nelson, & May, 1990). In the context of a depression assessment, EEs may be

patients with depression. Finally, practical experts (PEs) may be research assistants who administer the questionnaire, and are well positioned to provide feedback based on their experience with the assessment takers. This feedback can be another source of information in content validity studies.

Internal structure. Another source of validity evidence involves examining the internal structure of the test. This process focuses on the relationships among the test items, impacts scoring, and assesses the degree to which these relationships reflect the theoretical structure of the intended construct (AERA et al., 2014). The *Standards* (AERA et al., 2014) describes two main sources of evidence in regard to internal structure: dimensionality and measurement invariance.

Dimensionality concerns the inter-relationships among the test item. As a source of validity evidence, it also involves examining if the internal structure findings support the intended use and inference of the test scores (Rios & Wells, 2014). Factor analysis is a commonly used statistical method used to assess dimensionality of responses to a measure (Kline, 2013). It identifies the number of latent variables, indicated by dimensions or factors, which are clusters of items that have high inter-correlations with each other relative to other items in the measure. In addition, factor analysis also identifies which items are linked to which factor. When the underlying dimensions are unknown, exploratory factor analysis (EFA) can be used (Rios & Wells, 2014). Alternatively, when the dimensions are theorized, a confirmatory factor analysis (CFA) is more appropriate. In CFA, the researcher explicitly specifies the number of factors using theory and previous research and which items will load on the different factors (if more than one factor) (Rios & Wells, 2014). CFA can verify the number of factors and pattern of factor loadings. Hence, if the hypothesized structure is not correct, the CFA model will result

in poor fit. If a factor analysis model produces one latent variable, as suggested by a unidimensional structure, this is evidence that support the use of a single composite (total or average) score. Alternatively, if a factor analysis model consists of multiple latent variables, suggested by a multidimensional model, then each factor can be treated as a subscale. This suggests that a score for each subscale should be calculated instead. Use of both subscale scores and a total score requires evidence of a more complex model (e.g., higher order factor structure).

Some validity studies have relied on Principal Component Analysis (PCA) instead of factor analysis methods to examine the internal structure of a test. It is important to note that PCA is not a form of factor analysis. Rather, it is a technique for variable reduction that is used when there are high correlations among variables (Suhr, 2005). Unlike factor analysis, PCA does not hypothesize latent constructs nor estimate the influence of factors on observed variables (Suhr, 2005). PCA can overestimate variance explained by components (Costello & Osborne, 2005), and is not recommended for scale development and validation (Cabrera-Nguyen, 2010; Costello & Osborne, 2005).

Another source of evidence to examine internal structure is measurement invariance. Measurement invariance assesses if members from different groups (e.g., different gender or cultural groups) attribute the same meaning to the construct being measured (Putnick & Bornstein, 2016). Measurement invariance also applies to a construct being measured at different time points, since there can be changes in the interpretation of the construct over time (Putnick & Bornstein, 2016). Measurement invariance is commonly assessed within the structural equation modeling framework (Putnick & Bornstein, 2016). There are four common measurement invariant tests, with increasing levels of stringency: configural invariance, metric invariance, scalar invariance, and residual invariance. Configural invariance is the first, least stringent step

that assesses if members of different groups have the same conceptualization of the construct. Once this is satisfied, metric invariance is tested to assess if strength of the relationships between test items and the latent construct is similar between different groups. The third step after metric invariance is supported is scalar invariance. Scalar invariance examines the relationship between the observed score and the latent construct across groups, such that changes in observed scores are reflective of changes in the latent construct, regardless of group membership. Finally, residual invariance tests if the measurement error of each test item is equal between groups. Residual invariance is reported less compared to the three other types of measurement invariant tests (Putnick & Bornstein, 2016).

Response processes. According to the *Standards* (AERA et al., 2014), response processes refers to the "cognitive processes engaged in by the test takers" (p.15). Hubley and Zumbo (2017) argued for a broader definition that includes thoughts, behaviours, motivations, and emotions that people engage in when responding to a test item and influence the observed score. This broader definition encourages evidence on response processes to take into account the impact of affect and motives. Additionally, it pushes researchers to consider contextual and situational influences, such as cultural differences, that can impact the test taker's interaction with test items. Hence, evidence based on response processes examines whether the mechanism the test takers engage in to respond to items matches the process expected theoretically.

There are a large variety of methods to assess response processes. A popular method is cognitive interviewing, including the use of the think aloud protocol (TAP) and verbal probing (Willis, 1999). In a TAP, participants are asked to complete the questionnaire and verbalize their thought process of arriving at an answer for an item. Verbal probing is often used in conjunction, where the experimenter probes for further details regarding the response (Willis, 1999).

Relations to other variables. A widely used source of validity evidence is relation to other variables. This source of evidence examines the relationships between the scores from the measure of interest with other variables, such as a criterion or the scores from other more or less related measures, and the degree to which these relationships conform to theoretical predictions based on the underlying construct (AERA et al., 2014). According to the *Standards* (AERA et al., 2014), there are three types of evidences that fit under this category: convergent and discriminant evidence, test-criterion relationships, and validity generalization.

Convergent and discriminant evidence. Convergent and discriminant evidence examines the relationship between the test score from the measure of interest and those of other measures (AERA et al., 2014). The current view on validity emphasizes the theoretical underpinning of the construct that the test is intended to measure. As such, it is important to consider the ways in which the construct is connected to other related constructs.

Convergent validity evidence focuses on the correlations between measures of the same construct (e.g., depression) or theoretically closely related ones (e.g., depression and anxiety). Discriminant validity assesses the association between scores on measures of theoretically distinct constructs. When examining validity evidence for depression measures, anxiety has been used as both a convergent and discriminant measure. As a convergent measure, researchers expect to see a relatively high correlation, because anxiety and depression have overlapping symptomology and are closely related. It can also act as a discriminant measure, since it is important to show that correlations between depression and anxiety measures are lower than that of two depression measures. Discriminant measures may also consistent of constructs that are theoretically less related or unrelated, such as depression and intelligence.

Rather than focusing on labeling a measure as convergent or discriminant, Hubley and Zumbo (2013) described convergent and discriminant validity evidence as existing on a continuum. So, it is more useful to consider the pattern of relationships between measures. Comparisons between convergent and discriminant validity coefficients from the same sample are be used to support test interpretation. Measures of the same construct should have the highest correlation, followed by measures of theoretically similar constructs, while correlation between theoretically dissimilar measures should be comparatively lower. For example, when providing convergent and discriminant validity evidence for a depression measure, two depression measures are expected to have strong and positive correlations, followed by a relatively lower correlation with a measure of a related construct (e.g., anxiety), and then by the correlation between depression and a measure of a less related construct (e.g., physical health).

Test-criterion evidence. Test-criterion evidence assesses the efficacy of the measure in predicting a criterion or outcome measure (AERA et al., 2014). Diagnostic interviews are often used as the gold standard to provide test-criterion related evidence for self-report depression inventories such as the PHQ-9. Unlike self-report questionnaires, diagnostic interviews are more time-intensive, require trained administers, and can produce diagnoses.

There are some distinctions between the diagnostic processes of depression in research compared to clinical settings (Targum, 2011). In clinical practice, the interview process can include an array of open- and close-ended questions, as well as empathic listening, paraphrasing, and reflection. Aside from gathering information, the initial interview is used to foster therapeutic alliance, and engender therapeutic benefits. In contrast, research interviews have a stronger emphasis on gathering information, while remaining neutral. While developing a rapport is still important, interviews in research is not intended to facilitate therapeutic

relationship or benefit. In order to provide validity evidence, standardized diagnostic interviews have become commonplace in research.

There are two main types of diagnostic interviews for depression: structured and semi-structured (Levis et al., 2018). Semi-structured interviews are more flexible and incorporate both standardized questions as well as additional queries based on clinical judgment (Levis et al., 2018). Examples of semi-structured interviews include Structured Clinical Interview for DSM (SCID) and Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Structured interviews are typically fully standardized and scripted, with questions read verbatim and no additional probing (Levis et al., 2018); examples include the Diagnostic Interview Schedule (DIS), the Composite International Diagnostic Interview (CIDI), and the Mini International Neuropsychiatric Interview (MINI).

Evaluation of criterion-related validity can be done with two designs: predictive and concurrent (Furr & Bacharach, 2013). In predictive studies, the criterion measure is obtained at a later time than the measure of interest. In contrast, in concurrent studies, the criterion measure and measure of interest are obtained at or around the same time. For depression, criterions are typically diagnostic interviews, such as the SCID, which is often used to confirm the presence of depression (Gilbody et al., 2007). Test-criterion evidence is often provided through receiver operating characteristic (ROC) curve analysis. In a ROC curve, true positives (sensitivity) are plotted on the vertical axis against false positives (1/specificity) on the horizontal axis, and the area under the curve (AUC) is an indicator for the accuracy and usefulness of the test. The greater the curve deviates from the diagonal straight line toward the upper-left corner of the group, the greater the AUC, and the more indicative of the scale's ability to discriminate between those who test positive (e.g., have depression) and negative (e.g., do not have depression).

Parameter sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are then calculated. Sensitivity is the proportion of true positives. For a depression measure, sensitivity is the percentage of depressed people in the sample identified by the criterion that is correctly identified as such by the measure. Specificity measures true negatives. For a depression measure, it is the percentage of non-depressed people as identified by the criterion that the depression measure correctly identified as such. Often, for depression screens in the clinical context, sensitivity can be weighed more than specificity, so that truly depressed are identified for further screening. Unlike sensitivity and specificity, predictive values vary with the prevalence rate, such that low prevalence will lead to lower PPV. PPV is the proportion of true positives, as identified by the criterion, of all of positives as identified by the measure. Finally, NPV is the proportion of true negatives that are identified as negative by the criterion, out of all the negatives as identified by the measure.

Validity Generalization. Validity generalization examines the extent to which validity evidence based on test-criterion relationships can be generalized to new contexts or groups (AERA et al., 2014). This type of evidence typically uses meta-analysis to evaluate the test-criterion validity coefficients across a large number of studies. The result summaries the evidence across studies and can provide information on the average level of validity coefficients, the degree of variability of these coefficients, and the source of variability (Furr & Bacharach, 2013).

Consequences of testing. Validity evidence regarding consequences of testing involves assessing the integrity of the proposed interpretation and intended use. Messick (1998) emphasized that test consequences refer to consequences of legitimate test interpretation and use, and not test misuse. To justify an interpretation and use of a test, value implication and social

consequences must be considered (Messick, 1989). Value implications involves critically examining the personal and social values of the construct, its naming, its underlying theory, as well as the social ideologies that influenced the development of the underlying theory (Messick, 1989). Social consequences include positive and negative consequences for society stemming from legitimate test use. Hubley and Zumbo (2011) highlighted that both intended and unintended social consequences need to be considered. It is important to note that not all social consequences are considered sources of validity and invalidity (Hubley & Zumbo, 2011). As a source of validity evidence, social consequences must be the result of construct underrepresentation or construct irrelevant variance (Messick, 1989). After identifying the intended or unintended consequences of legitimate test use, it is important to consider the impact of these consequence on the interpretation of score and its use.

Summary. This review highlights the current, unified view of validity. That is, validity is about the inferences and use of test scores, and not a property of the test itself. Also, validity cannot be separated from the context. Hence, the process of validation involves the accumulation of evidence that supports the intended inferences and uses of the test scores and is an ongoing process. The accumulation of evidence can come from five difference sources, including internal structure, relationships to other variables, test content, response processes, and consequences of testing (AERA et al., 2014). From this review, it is noted that relations to other variables and internal structure are two of the most frequently studied sources of validity evidence, especially in newer measures (e.g., Hubley, Zhu, Sasaki, & Gadermann, 2014). Content validity is recommended during the construction of the test to determine the appropriateness of test items, but can be studied later on as well. Response processes and consequences of testing as validity evidence are emerging fields and tend to be studied with well-established measures. Because the

PHQ-9 has very limited evidence for use with university students, this review suggests that relationships with other variable and internal structure are appropriate sources of evidence as starting point.

Reliability

This section will provide an overview of reliability, including common methods of calculating reliability. This will be used to guide the choice of reliability estimate for this study. Reliability is the degree to which a test measures a construct consistently, over repeated administrations under the same conditions with the same people. To estimate reliability, the reliability coefficient describes the degree to which the test scores are free from random measurement errors. Based on classical test theory (CTT), observed scores are the sum of true scores plus random error. Thus, reliability can also be described as the extent to which the variation in respondents' observed scores is attributed to the variation in true scores.

There are three main methods for estimating reliability: alternate forms, test-retest, and internal consistency. Each method depends on a unique set of assumptions about the participants and the testing procedures and may have different sources of error, so no one method will provide the single best estimate of reliability of true scores under all circumstances.

Alternate forms reliability. Alternate forms reliability can be obtained when two different forms of the test are administered. The correlation between the two scores can be interpreted as an estimate of the test score reliability. However, this interpretation rests on the assumption that the two forms are parallel. To be considered parallel, the true scores for the two forms must be equivalent and scores for the two forms must have the same error variance. In reality, it is impossible to be entirely confident that the two forms are truly parallel. In particular, it is infeasible to know for certain that the two forms are measuring the same construct and have the

same true scores. Another issue with the alternate forms reliability is that, due to repeated testing, one can see carryover or contamination effects. That is, the completion of the first form might have an effect on responses to the second form, due to memory of test content, desire for consistency, or mood. When that occurs, the error scores from the two forms may be correlated, which violates a foundational assumption of CTT. In addition to the issues with alternate forms reliability, because there is only one version of the PHQ-9, it is not feasible to administer another form.

Test-retest reliability. Test-retest reliability is particularly useful for measures of stable constructs. Test takers take the same test twice, and the correlation of the two sets of scores provides an estimate of the test score reliability. When calculating test-retest reliability, both assumptions of parallel tests need to be met as well. To meet the first assumption of equivalent true scores, the true score must be stable across the two testing sessions. There are three factors to consider. First some psychological attributes (e.g., personality characteristics) are more stable than others, whereas some constructs, such as mood, are more state-like. Test-retest reliability is more appropriate with trait-like attributes. Secondly, the length of time between the testing can influence results. It is important that a test-retest interval is selected during which stability of the true score is a reasonable expectation. Short intervals might risk carryover effects, but long intervals might permit change in the true score. Third, the second assumption of equal error variance can be reasonably satisfied. Error variance is strongly impacted by the testing situations such as noise or distractions, but these elements can be reasonably controlled. Additionally, test-retest reliability has the challenge of requiring two testing sessions, thus can require more time as well as additional concerns regarding participant attrition and missing data.

Internal consistency reliability. The third method of estimating reliability is through internal consistency and is the most widely used method for estimating reliability. The practical advantage is that it requires only one form of the test and the test takers only need to complete the test once. It includes several related, but different, procedures to estimate reliability. In this approach, different parts of a test, such as items or groups of items, are treated as different forms of the test. Then, a statistic can be computed to summarize the degree of consistency among the responses from the various parts. Internal consistency includes many estimators. The commonly utilized approaches for internal consistency are the split-half approach, Cronbach's alpha, and standardized coefficient alpha. More recently, approaches that take into account the ordinal nature of the response (i.e., less than five points on a Likert-type response format) have been promoted (e.g., ordinal alpha).

Split-half estimate of reliability. The split-half estimate of reliability was proposed by Spearman and by Brown, independently (Furr et al., 2001). The test is administered once and is divided equally (e.g., first/second half, odd/even items), and assumed to be parallel. The two halves are treated as alternate forms, and the correlation between the scores from the halves is calculated. However, because the test is split into two, this correlation is not representative of the reliability of the whole measure. To correct for the decreased estimate when the test is split in half, the Spearman-Brown formula is used to correct the reliability coefficient. A key disadvantage of the split-half estimate is that there is no single, most accurate way of splitting the test. So, depending on how the test is divided, there can be multiple estimates.

Cronbach's alpha. Cronbach's alpha is the mean of all possible split-half coefficients, corrected for the shortened length using the Spearman-Brown correction. It is the most widely used procedure for estimating reliability, but its limitations are well documented (Sijtsma, 2009;

Yang & Green, 2011). The use of the Cronbach's alpha as a reliability estimate rests on several assumptions, and some important ones include the uncorrelated-errors assumption, tau-equivalence and normality (Trizano-Hermosilla & Alvarado, 2016; Yang & Green, 2011). These assumptions dictate that the item error variations between any two items need to be uncorrelated, each item contributes to the total score equally, and that the test scores are normally distributed. Violation of these assumptions can lead to either underestimation or overestimation of reliability (Yang & Green, 2011).

Cronbach's alpha is based on the Pearson correlation matrix, which assumes that the item responses are continuous. Violation of this assumption may distort the Pearson correlation matrix (Rupp, Koh, & Zumbo, 2003). As a result, the Cronbach's alpha can underestimate the reliability when used with measures using a Likert-type response format (i.e., ordinal responses), especially ones with less than five scale points, such as the PHQ-9 (Zumbo, Gadermann, & Zeisser, 2007). Alternatively, an ordinal alpha can estimate reliability more accurately for ordinal scales (Zumbo et al., 2007; Gadermann, Guhn, & Zumbo, 2011). Conceptually equivalent, the ordinal alpha is based on the polychoric correlation matrix, which takes into consideration the ordinal nature of the response format (Zumbo et al., 2007).

Standardized coefficient alpha. Another reliability estimate is the standardized alpha estimate, or the generalized Spearman-Brown formula. This method is closely associated with the Cronbach's or raw alpha method and is used when the items have highly different variances from one other. Uncorrected, the score will be heavily influenced by the item with the largest variance. Thus, the reliability is computed using standardized item responses.

Summary. Test score reliability is intrinsically tied to the validity of its inferences. In fact, reliability may be viewed as necessary for validity or, by some, as a preliminary form of

validity evidence. Some theorists include reliability as a form of evidence for internal structure (Rios & Wells, 2014). High reliability increases the confidence that the test is measuring real individual difference or variability. Contrarily, low reliability is indicative of increased random errors, which increases the difficulty of replication of findings and can attenuate subsequent statistical findings (e.g., validity coefficients).

Review of PHQ-9 Validation Studies

This section will review existing research on the validity evidence for the PHQ-9 across a range of populations. The intent is to highlight and understand existing literature, which will guide and support the rationale for selecting the methods and measures used in this study. The inclusion criteria for this review are that studies: 1) use the 9-item PHQ-9, rather than a revised version, 2) use administration of the English version of the PHQ-9, and 3) were published in the last 10 years (i.e.. 2008-2018).

The majority of these studies were conducted in the U.S., with a few from the U.K., Australia, Canada, and New Zealand. Because the PHQ-9 was originally designed for the clinical setting, unsurprisingly, the vast majority of these studies were conducted with a medical population. Upon review, three general groupings emerge: primary care, secondary care, and general community. Validity studies with patients from the secondary care setting are, by far, the most frequently studied, which include a wide range of populations, such as dialysis patients and cancer patients. In terms of sources of validity evidence, aside from reliability estimates, test-criterion validity is the most frequently studied. The majority of these studies used a diagnostic interview as the gold standard criterion, and the SCID was the most frequently used. Other commonly used diagnostic interviews include the MINI, CIDI, CIS, and DIS. Many studies also examined convergent and discriminant validity evidence, where the PHQ-9 was often compared

to other measures of depression, anxiety, and well-being. Lastly, many studies also reported on the internal structure of the PHQ-9, mostly by conducting CFA, but others used EFA and PCA as well.

Primary Care Setting. The PHQ-9 was originally designed for the primary care setting as part of a more comprehensive self-administered measure, the Primary Health Questionnaire, a 3-page questionnaire that assessed eight disorders. The first validity study by Spitzer and colleagues (1999) evaluated the entire PHQ by comparing its results with independent diagnosis by mental health professionals using results from 3000 adult patients from eight primary care clinics. The authors found good agreement between the measure and diagnoses (overall accuracy = 85%, sensitivity = 75%, specificity = 90%). The PHQ-9 was examined by itself by Kroenke and colleagues (2001) based on 6000 patients in primary care and obstetrics-gynecology clinics. The PHQ-9 was found to have high correlation ($r = 0.84$) with interviews given by mental health professionals (Kroenke et al., 2001). Good sensitivity and specificity (88% for both) at a cut-score of 10 was reported. Additionally, an increase in severity as indicated by PHQ-9 scores was associated with a decrease in functional status on all six SF-20 quality of life subscales.

In the past 10 years, only three studies examined criterion-test validity evidence for the English version of the PHQ-9. All studies used diagnostic interviews to confirm the absence or presence of depression in the sample and to serve as the criterion measure. Two studies used the SCID (Fine et al., 2013; Phelan et al., 2010) and another used CIDI (Arroll et al., 2010). Phelan et al. (2010) focused on patients over 65 in primary care. To strengthen support for their results by controlling for potential confounds, the authors controlled for dementia, unstable medical conditions, and poor language fluency by having them as exclusion criteria (Phelan et al., 2010). Likely due to the age group sampled, this study had a smaller sample size compared to others,

with a sample of 71 participants. The authors reported that the AUC of 0.87 for the PHQ-9 was comparable to the 15-item GDS and similar for men and women. Results of ROC curve analysis also showed that a cut point of 9 offered the best combination of sensitivity and specificity in this study, and that the sensitivity of both the PHQ-9 and GDS decreased when a broader definition of depression (including minor and major depression) was used. With a sample of 498, Fine and colleagues also examined sensitivity and specificity of multiple scoring methods, including the diagnostic algorithm and various cutoff scores. The results demonstrated a cutoff of 10 had the most optimal statistics. Arroll et al. also supported the cut-off point of 10 in a large sample of 2,642 family practice patients. They used a computerized version of the CIDI, unlike the other two studies that conducted the diagnostic interview in person.

The only study that reported on convergent and discriminant validity related to the PHQ-9 in primary care in the past 10 years compared its results to that of the Hospital Anxiety and Depression Scale (HADS) (Cameron et al., 2008). The HADS contains two subscales, depression and anxiety, which provided convergent and discriminant evidence, respectively. This study was a part of a service audit, in which 1063 patients completed the measures at baseline and 544 returned to complete it again post treatment. Results found that the correlations were, overall, higher between the two depression measures (PHQ-9 and HADS-D), which provided convergent validity, and lower correlations were found with the anxiety subscale (HADS-A).

Cameron et al. (2008) also examined the factor structure of the English PHQ-9 using principal component analysis (PCA), which revealed a one-factor solution that accounted for 42% of variance. In this 10-year review of studies of the English version of the PHQ-9 in primary care patients, this study was the only one that reported a reliability estimate. The authors used Cronbach's alpha, which was 0.83 at baseline and 0.92 at the end of the treatment.

Secondary Care Setting. Validity studies of the PHQ-9 in the past 10 years are most frequently done in secondary care settings. Populations studied included patients with a wide range of health conditions, such as kidney issues requiring dialysis (Chilcot et al., 2018), substance use disorder (Hepner et al., 2009, and cancer (Randall et al., 2013). Test-criterion validity evidence was the most frequently studied, and the majority of the studies used diagnostic interviews as the criterion. The most commonly used diagnostic interview was the SCID, followed by the C-DIS, CIS-R, and MINI. In one study providing validity evidence for the PHQ-9 with a sample of 487 patients with depression, Schueller et al. (2015) used the HAM-D as the criterion measure. However, the HAM-D is a self-report screen and has less credibility than a diagnostic interview. Most of the studies supported using the cut-off score of 10 (Elderon et al., 2011; Rathore et al., 2014; Rooney et al., 2013; Thompson et al., 2010). However, there has been a wide range of proposed cut-off scores from 6 (Thombs et al., 2008) to 13 (Beard et al., 2016). Sensitivity reported generally ranged from 0.80 to 0.94 and specificity from 0.74 to 0.90. There are exceptions however. For instance, in a study of 214 patients with Parkinson's disease, Thompson et al. (2010) reported a low sensitivity of 0.50 for the cut-off score of 10.

The variability in reported cut-off scores and the range in sensitivity can be attributed to two major differences between studies. First, there is large heterogeneity and differences with the populations studied, from psychiatric populations to those with spinal cord injury. It is vital to gather validity evidence for each population. Furthermore, different diagnostic interviews and different administration methods were used, from a layman administered CIS-R to a computer administered DIS to a clinician administered SCID. In the review of this literature, it is noted that almost all of the studies focused on the diagnosis of MDD, with the exception of one study that also reported on results based on minor and major depression combined (Thompson et al., 2010).

In comparison to focusing on MDD alone, the authors reported lower sensitivity and specificity when accounting for both minor and major depression.

Compared to primary care and community populations, more studies have provided convergent and discriminant validity evidence in secondary care. In examining convergent validity evidence, the PHQ-9 has been compared to other commonly used depression screens. In particular, the CES-D and the BDI-II are the most commonly studied. In a sample of 129 patients with chronic hepatitis C, Dbouk and Arguedas (2008) provided convergent validity evidence for the PHQ-9 through large and significant correlations with both the CES-D and the BDI-II. Similar correlations have been reported by other studies with various patient populations. In general, the PHQ-9 correlated strongly with the CES-D ($r = 0.77 – 0.83$) (Beard et al., 2016; Dbouk & Arguedas, 2008; Milette et al., 2010; Pilkonis et al., 2014) and the BDI-II ($r = 0.72 – 0.84$) (Dbouk & Arguedas, 2008; Dum et al., 2008; Hepner et al., 2009; Titov et al., 2010; Turner et al., 2012). A few studies reported relatively lower correlations between the PHQ-9 and another depression measure. Haddad et al. (2013) reported a correlation of 0.71 between the PHQ-9 and the depression subscale of the HADS in a sample of 740 patients with coronary heart disease. Turner et al. (2012) also reported lower correlation ($r = 0.66$) between the HADS and the PHQ-9 in a group of 72 stroke patients.

The PHQ-9 is also frequently compared to measures of anxiety. Anxiety has overlapping symptomology but is a separate diagnosis. So, a comparably lower correlation with an anxiety measure relative to a depression measure can be used as evidence for discriminant validity. Beard et al. (2016) and Chilcot et al. (2018) compared the PHQ-9 to the GAD-7, an anxiety measure, in samples of psychiatric and dialysis patients, respectively. Both studies reported

similar correlations (r = 0.61 and 0.67) that are lower than the reported correlations with the depression screens, demonstrating discriminant validity.

Discriminant validity was also demonstrated by lower correlations with measures of health outcome (SF-36) (r = -0.68 with mental health, r = -0.43 with physical health) (Millette et al., 2010) and life satisfaction (SWLS) (r = -0.46 between an affective subscale of the PHQ-9 and SWLS, r = -0.35 between a somatic subscale of the PHQ-9 with the SWLS) (Richardson & Richards, 2008). Surprisingly, Turner et al. (2012) compared a number of depression screens, including the HADS, BDI-II, and PHQ-9 to distress measures (i.e., the distress thermometer (DT) and the Kessler Psychological Distress Scale-10 or K-10). While the correlations among the depression screens are similar to those of other studies, the correlation between the K-10 and the PHQ-9 as well as the BDI-II was 0.83. The large number of overlapping items can explain this high correlation. Despite named as a distress scale, the K-10 and the PHQ-9 share a number of very similar items, namely regarding feelings of hopelessness, tiredness, depression, restlessness, and lack of interest.

As a whole, several strengths were noted in the studies reviewed. Several studies provided convergent validity evidence by correlating with several measures of related constructs (Beard et al., 2016; Milette et al., 2010; Turner et al., 2012). Additionally, many of them further reinforced their convergent findings by confirming the depression diagnosis with diagnostic interviews (Beard et al., 2016; Haddad et al., 2013; Turner et al., 2012). While the patient population varied in terms of diagnosis and age, the PHQ-9 consistently demonstrated convergent validity with high correlations with the CES-D and the BDI-II, two of the most commonly used depression screens.

Factor analyses on PHQ-9 responses in the secondary care population have yielded conflicting results. Many studies reported the single factor structure as originally proposed (Dum et al., 2008; Hepner et al., 2009; Ryan et al. 2013), but others failed to find support for a unidimensional structure (Beard et al. 2016; Chilcot et al., 2013; Chilcot et al., 2018; Richardson & Richards, 2008; Titov et al., 2010). CFA was used to confirm a single factor model that provided a good fit for the PHQ-9 with a sample of 23,672 patients registered with the Improving Access to Psychological Therapies (IAPT) services for depression and anxiety disorders in U.K. When error variances for items 3 and 4, and for 7 and 8 were correlated, the CFI was found to be 0.97 for face-to-face administration and 0.95 for telephone administration. The authors also reported RMSEA to be 0.07 for both methods of administrations, which is close to the recommended 0.06 (Hu & Bentler, 1999). Hepner et al., (2009) also used CFA to confirm the one-factor structure and reported good fit in a sample of 240 patients with substance use disorder. Dum et al. (2008) used PCA with a sample 108 alcohol and substance users, which revealed a satisfactory one-factor structure that accounted for 59% of the variance. This study has a lower sample size relative to others. In considering the appropriateness of its sample size, Comrey and Lee (1992) considered a sample size of 100 to be poor, but Everitt (1975) argued that a ratio of at least 10 participants to 1 item is adequate. A sample of 108 exceeds the 10:1 ratio and thus may be found to be sufficient.

Other studies failed to support the one-factor structure (Beard et al. 2016; Chilcot et al., 2013; Chilcot et al., 2018; Richardson & Richards, 2008; Titov et al., 2010). Titov and colleague (2010) conducted a CFA in an attempt to confirm the one factor model with a sample of 172 patients with depression. However, the authors reported a significant chi-square statistic, CFI of 0.90, and high mean square error of approximation (RMSEA) of 0.10, which led the authors to

conclude poor model fit. The model fit did not improve when it allowed for correlated errors between related items. The authors attributed the disparity between their findings and previous literature to the homogeneity of the depressed sample, which reduced variability in the PHQ-9 responses. Chilcot et al. (2013) tested both one-factor and two-factor structures using CFA. While other studies used maximum likelihood (ML) estimation, the authors used weighted least-squares with mean and variance adjustment estimation (WLSMV) due to the ordinal response format and skewed distribution of the PHQ-9. The chi-square test was significant for both models, but the two-factor model demonstrated better model fit as evidenced by a higher comparative fit index (CFI), lower RMSEA, and lower weighted root mean square residual (WRMR) (Chilcot et al., 2013).

Other studies used EFA instead, arguing that it was the first investigation of the PHQ-9 factor structure in the population studied. Beard and colleagues (2016) investigated the factor structure in a study with 1023 psychiatric patients, appropriately conducting EFA on half of the sample, followed by CFA on the other half. The authors conducted EFA using ML estimation and oblique rotation. Based on the eigenvalue > 1 rule and the scree plot, a two-factor structure emerged, accounting for 60.2% of the variance. The two-factor structure was composed of one factor with cognitive and affective items, and the second factor captured somatic items. Subsequent CFA with modification indices by freeing up the covariance between item 7 and item 8 supported the two-factor structure with good model fit as indicted by a CFI of 0.98 and RMSEA of 0.06. Richardson and Richards (2008) also identified a two-factor structure using EFA in a group of 2570 patients with spinal cord injury. Chilcot et al. (2018) used CFA to confirm a bi-factor structure with a group of 182 dialysis patients. In the bi-factor structure, all nine items were loaded onto a general depression factor, and items were also loaded onto two

smaller groups, somatic and affective/cognitive. This model demonstrated excellent model fit, with a non-significant chi-square statistic, CFI of 1.0, and RMSEA < 0.01.

Lastly, most of the studies, although not all, reported reliability of the scores. Evidence of reliability can be provided a number of ways. The most frequently reported is an internal consistency estimate provided by Cronbach's coefficient alpha. The majority of the studies reported a single value for the Cronbach's alpha in the range of 0.80-0.90 (e.g., Beard et al., 2016; Dum et al., 2008). Titov et al. (2010) reported a lower alpha estimate of 0.74, which the authors believed to be due to the homogeneity of the sample (all of whom were patients with depression). Two studies that identified a two-factor structure also reported Cronbach's alpha for each subscale (Chilcot et al., 2013; Richardson & Richards, 2008). Both studies reported a higher reliability estimate for the affective component relative to the somatic component. Delgadillo et al. (2011) also investigated test-retest reliability by administering the PHQ-9 after 4-6 weeks and reported an intraclass correlation coefficient (ICC) of 0.78, indicative of good reliability (Koo & Li, 2016).

Community. Validity studies of the PHQ-9 in a community setting have examined several populations, including the general population (Kiely & Butterworth, 2015; Liu & Wang, 2015), low-income women (Kneipp et al., 2010), pregnant women (Gjerdingen et al., 2009; Sidebottom et al., 2012), and occupational health professionals (Volker et al., 2016). None of the studies reported estimates of reliability.

The majority of the validity studies with community samples focused on test-criterion validity as well. The SCID, CIDI and MINI have been used as the criterion. The majority of the studies supported the use of a cutoff score of 10, reporting good sensitivity and specificity above 0.80 (Gjerdingen et al., 2009; Kiely & Butterworth, 2015; Sidebotton et al., 2012). Liu and Wang

(2015), however, found a noticeably lower sensitivity of 0.51 but high specificity of 0.93 in their study that examined using the PHQ-9 in the general population. There are a number of differences to note. Many other studies reviewed were designed to be validity studies, but this study was a secondary analysis of a population-based longitudinal study. The criterion measure was the CIDI, a structured diagnostic interview based on the DSM, but, unlike the SCID, is layman administered. The CIDI is a commonly used in epidemiological studies, but has been criticized and is considered to be less optimal compared to other diagnostic interviews (Brugha et al., 2001). Furthermore, Gelaye and colleagues (2014) found the CIDI to bias the diagnostic accuracy estimate of the PHQ-9. Finally, the prevalence rate of depression in this study was 2.21%, much lower than that of other PHQ-9 validation studies.

Kiely and Butterworth (2015) also studied the utility of the PHQ-9 in the general community using the CIDI as the criterion. Unlike other studies, the authors reported an alternate cutoff score of 8. This is study was part of a longitudinal, multi-cohort study based in Australia. The authors speculated that the lower cutoff score might be due to the difference between the general population and the patient population, for which the PHQ-9 was originally designed. However, other community-based studies have also supported the cutoff score of 10 (Manea et al., 2012).

Kneipp et al. (2010) was the only study that evaluated convergent and discriminant evidence for validity. In a sample of 308 low-income women, the authors administered two depression screens, the PHQ-9 and the BDI-II, as well as measures of theoretically distinct constructs including perceived stress, using the Perceived Stress Scale (PSS), and anxiety, using the Beck Anxiety Inventory (BAI). Given the skewed distribution of scores from both depression screens, Spearman's rho was used to examine correlations. As hypothesized, the correlation

between the PHQ-9 and the BDI-II was the highest (0.80), demonstrating evidence for convergent validity. As evidence of discriminant validity, the PHQ-9 result had a correlation of 0.61 with both the PSS and the BAI.

University Population. The validity of inferences made from, and use of, the PHQ-9 has been studied extensively in primary and secondary care settings. However, evidence of its application in the university population is extremely limited. Studies of its use with Nigerian, Japanese, and Chinese university students have been conducted, but research with the English version in North America is limited to a single secondary data analysis in U.S. (Keum et al., 2018).

Keum et al. (2018) examined the reliability and internal structure of PHQ-9 scores as well as convergent and discriminant relationships with other constructs in a group of 857 racially diverse university students. The reliability estimate, using Cronbach's alpha, was 0.89, suggesting good reliability. CFA was used to verify a single factor model, as originally proposed by Kroenke et al. (2001). The goodness-of-fit indices for the model were reported by race and gender. The authors reported that the model fit was supported by appropriate CFI (0.99-1.0) and standardized root-mean residual (SRMR) (0.04-0.08) values. Both point towards a good model fit (Hu & Bentler, 1999). Additionally, the RMSEA ranged from 0.058 to 0.093, which indicates an adequate model fit. The authors also tested three variations of two-factor models, which all showed adequate to good fit across different racial and gender groups and suggested that both structures many be considered (Keum et al., 2018). However, they also found that all three two-factors models possessed high factor inter-correlation (0.85-0.97) and concluded that a single factor is best suited for clinical and measurement use. To examine relationships with other variables, PHQ-9 scores were correlated with Mental Health Continuum-Short Form (MHC-SF)

scores and alcohol use. MHC-SF is a 14-item measure of mental-health and well-being, and it has three subscales: emotional well-being, psychological well-being, social well-being. The authors hypothesized that PHQ-9 scores would be negatively correlated with the scores from the MHC-SF, but more specifically would be correlated most with the emotional well-being subscale, followed by psychological well-being and social well-being. Their findings supported this hypothesis, with the correlations of -0.60, -0.53 and -0.41 for emotional well-being, psychological well-being, and social well-being respectively. The PHQ-9 scores had the lowest correlation with alcohol use ($r = 0.10$).

Summary. This review of validity studies of the PHQ-9 has produced a substantial number of studies and it appears to be one of the most studied depression screens across many different populations. It is noted, however, that the vast majority of these studies focused on medical patients and only a small number that examined the general public. Around half of the validity studies of the PHQ-9 conducted in the last 10 years are secondary data analyses of an existing study or of a larger data set. In a secondary data analysis, researchers may not have purposefully chosen the measures used for comparison when studying convergent and discriminant validity. Validity evidence for inferences from the PHQ-9 with university students is particularly under-studied. Validity evidence for the English version of the PHQ-9 has only been examined in one study and was a secondary data analysis. University students have qualitatively different challenges in life than that of patients in specialist doctors' office. So, it is important to purposefully examine the validity of the inferences of we make from the PHQ-9 when being used with a university student population.

Review of reliability and validity evidence for other depression measures

This section will review the existing reliability and validity evidence for commonly used

depression screens, specifically the BDI-II and the CES-D, with an emphasis on university students. The purpose of this review is to inform and guide the methods used in the current study by providing information of common practices in examining convergent and discriminant validity, internal structure, and reliability in similar circumstances. Given that this study will examine convergent and discriminant validity evidence, internal structure, and reliability, the review will focus on studies that also included these sources of validity evidence. For convergent and discriminant validity, it will provide further understanding about commonly used constructs and measures, how they are chosen, and the number of comparisons made.

Both the BDI-II and the CES-D have been studied in a wide range of populations from community to medical settings, and there is an abundance of literature providing validity constructs and measures evidence for both screens. The literature review process is similar to the one conducted for the PHQ-9. Studies are included based on the following criteria: 1) the validity study was focused on either the BDI-II or the CES-D, 2) used the complete measure, rather than a revised version, 3) used the English version of the BDI-II or the CES-D, and 4) was published in the last 10 years (i.e., 2008-2018), unless the sample was of university students, in which studies published before 2008 will be included. This review is not limited to any single population, but greater emphasis will be put on studies that include a university student sample.

Internal Structure. Internal structure is another commonly studied source of validity evidence. In validity studies of the BDI-II and the CESD in university students, internal structure appears to be the most frequently used source of evidence. Several studies reported factor analysis as well as convergent and/ or discriminant validity evidence (Contreas et al., 2004; Storch et al., 2004; Osman et al., 1997), whereas others relied solely on internal structure evidence (Carmody, 2005; Whisman et al., 2000; Arbona et al., 2017).

The vast majority of the internal structure studies used either EFA or CFA, with a very small number using PCA. In a review of the literature, it appears that CFA has been used more frequently in recent years. EFA was often conducted first when the factor structure was relatively unknown. Alternatively, when the factor structures were well-understood or theoretically supported, CFA was used to confirm the predicted model. Because both the BDI-II and the CES-D are commonly used and studied measures, their factor structures have been well documented if not always agreed upon. When conducting CFA, some studies reported testing one model, typically the most predominant model, whereas other studies reported results from testing multiple models. In a sample of 502 undergraduate students, Carmody (2005) used CFA to confirm the factor structure of the BDI-II by assessing goodness of fit indices of three models, i.e., one factor, two factors and three factors.

In the reviewed studies, when EFA was used, the number of factors was typically determined using the eigenvalue greater than 1 rule, scree plot, or both. Eigenvalues and percentage of variance explained were typically reported, although not by all studies. When CFA was conducted, it is noted that at least two fit indexes were reported in addition to the chi-square statistic. The two most commonly reported indexes were CFI and RMSEA. The majority of the studies reported two or three fit indexes, and one reported five (Tran, Ngo & Conway, 2003). Other commonly reported fit indices included GFI, AGFI, and SRMR.

Reliability. Most of the studies in this literature review, although not all, reported at least one reliability estimate. In a meta-analysis of 118 validity studies on the BDI-II, the authors reported that 25% did not report reliability estimates (Wang & Gorenstein, 2013). Reliability can be estimated a number of ways. Based on the reviewed studies, the two most common methods are (1) internal consistency estimate, using Cronbach's coefficient alpha, and (2) test-retest

reliability. Carter et al. (2016) is the only study that used a reliability estimate based on a factor analytic model: omega (ω). To investigate the use of the CES-D in patients with diabetes, the authors obtained ω =0.95 from the bifactor model, indicating excellent reliability.

Cronbach's alpha is the most commonly used reliability estimate. When reporting Cronbach's alpha, it is noted that most of the studies reported a single estimate for the total score, even if the authors examined internal structure and reported multiple factors. For example, Palmer and Binks (2008) reported a two-factor structure when examining the use of the BDI-II in a sample of incarcerated males, and a single reliability coefficient of 0.90. The majority of studies reviewed reported alpha coefficients around 0.90 for both the CES-D (Herniman et al., 2017; Milette et al., 2010; Stafford et al., 2014) and the BDI-II (Dezois et al., 1998; Lipps et al., 2018; Whisman et al., 2000; Osman et al., 2008; Palmer et al., 2014), indicating high internal consistency. If a study included more than one distinct population, separate total score alpha is reported. For instance, Segal et al (2008) studied the use of the BDI-II in community older adults and in college students, and reported two reliability estimates: 0.85 for the older adults and 0.92 for the college students. If the study included convergent and discriminant validity evidence, alphas were calculated for those measures as well. A few studies reported reliability estimates for individual subscales, instead of the total score. Arbona and colleagues (2017) identified a four-factor structure for the CES-D in a study with college students, and reported a reliability coefficient for each factor: 0.75 for somatic concerns, 0.83 for negative affect, 0.72 for lack of positive affect, and 0.73 for interpersonal concerns. It is noted that reliability coefficients for individual subscales tended to be in the range of 0.75-0.85, lower than that for the total score (Osman et al., 1997; Storch, 2004; Manian et al., 2013).

Test-retest reliability estimates appear to have greater variation. Originally, Beck (1996) reported a one-week test-retest reliability of 0.93 for BDI-II scores in a sample of 120 undergraduate students. More recently, Sprinkle et al. (2002) also reported a high test-retest reliability of 0.96 in group of 46 undergraduate students with an interval ranging between 1-12 days (on average, 3.2 days) whereas Wiebe and Penley (2005) reported a less impressive one-week test-retest reliability of 0.73. The majority of the studies reviewed used intervals of one or two weeks between test and retest. A few studies used noticeably longer periods. For example, Cukrowicz and Joiner (2007) retested participants after approximately 2 months and obtained an unacceptably low correlation of 0.44 for BDI-II. This lower correlation may be due to the longer interval and actual changes in depression level.

Convergent and discriminant validity evidence. Approximately half of the studies reviewed reported either convergent or discriminant evidence, or both. Several things are noted based on this review. First, there appears to be no consensus on the appropriate number of constructs or measures to include in providing this type of evidence. Second, the most frequently used measures are of depression and anxiety. Finally, there is large variability in choosing another measure to provide evidence for discriminant validity. Identifying the measures used in providing convergent and discriminant validity evidence for depression measures will guide the procedure used in this study. Based on the literature review, the measures used to provide convergent and discriminant evidence are listed in Table 2.1 and 2.2 along with the obtained validity coefficients.

Few studies provided only convergent or discriminant evidence (Shean & Baldwin, 2008; Kung et al., 2013), while most reported both sources of evidence (e.g. Osman et al., 2008; Segal et al., 2008). Shean and Baldwin (2008) reported only convergent evidence by administering two

depression measures, BDI-II and CES-D, to a sample of 395 university students. Palmer and Blinks (2008) is the only study in this review that reported only discriminant evidence, by administering the BDI-II alongside the Beck Hopelessness Scale (BHS). It is noted that there is no consensus on what is considered to be an appropriate number of discriminant measures to use. Some studies do not use any at all, while a few used as many as four. Generally, it appears that most studies used one to two measures to support this evidence. For example, Milette and colleagues (2010) provided both convergent and discriminant evidence for the CES-D by administering, along with the PHQ-9, the Short-Form (36) Health Survey (SF-36), Health Assessment Questionnaire (HAQ), and the McGill Pain Questionnaire.

Table 2.1

Convergent and discriminant validity evidence for the CES-D

Construct/ Measures	**r**	**Study**
Depression measures		
BPRS-D	0.60	Herniman et al., 2017
PHQ-9	0.77	Milette et al., 2010
HADS-D	0.72	Stafford et al., 2014
Anxiety measures		
BAI	0.68	McQuaid et al., 2009
ANS	0.50	McQuaid et al., 2009
SPQ (social phobia) -anxiety	0.51	McQuaid et al., 2009
SPQ (social phobia) -avoidance	0.46	McQuaid et al., 2009
Other construct/ measures		
CCS	n.r.	Arbona et al., 2017
SF-36 Mental Component Summary	-0.76	Milette et al., 2010
SF-36 Physical Component Summary	-0.33	Milette et al., 2010
SANS	0.24	Herniman et al., 2017
HAQ-disability	0.42	Millette et al., 2010

CES-D = Center for Epidemiologic Studies Depression Scale; BPRS-D = The Brief Psychiatric Rating Scale-Depression; PHQ-9 = Patient Health Questionnaire - 9; HADS-D = Hospital Anxiety and Depression Scale-Depression; BAI = Beck Anxiety Inventory; ANS = Autonomic Nervous System Questionnaire; SPQ = Social Phobia Questionnaire; CCS = College Stress

Scale; SF-36 = Short-Form 36 health Survey Questionnaire; SANS = Scale for the Assessment of Negative Symptoms; HAQ-disability = Health Assessment Questionnaire-Disability.

Convergent validity evidence is often supported by a high correlation between instruments that are intended to measure the same constructs. In one meta-analysis, Erford et al. (2016) reported that the BDI-II has been compared to 43 other depression measures. Two frequently used ones include the CES-D and the PHQ-9 (Wang & Gorenstein, 2013). The correlations vary across studies. Generally, the range of correlation between depression screens is similar between validation studies for the BDI-II and the CES-D, generally falling between 0.70 and 0.85.

Table 2.2

Convergent and discriminant validity evidence for the BDI-II

Construct/ Measures	**r**	**Study**
Depression measures		
PHQ-9	0.77	Kung et al., 2013
CES-D	0.86	Shean & Baldwin, 2008
	0.68	Segal et al., 2008
STAI-D	0.74	Storch, 2004
HADS-total	0.85	Turner et al., 2012
Anxiety measures		
BAI	0.66	Cunningham et al., 2008
	0.56	Osman et al., 1997
	0.53-0.63	Osman et al., 2004
	0.62	Contreas et al., 2004
DASS-A	0.44	Osman et al., 1997
CATI-A	0.68	Segal et al., 2008
STAI-A	0.69	Storch et al., 2004
	0.69	Storch et al., 2004
STAI-S	0.37	Osman et al., 2008

Construct/ Measures	r	Study
Other construct/ measures		
BHS	0.42	Cunningham et al., 2008
	0.63	Osman et al., 2008
	0.55	Palmer & Binks, 2008
DASS-Stress	0.68	Osman et al., 1997
Rosenberg Self-Esteem	-0.60	Osman et al., 1997
SBQ-R	0.57	Osman et al., 2008
SPWB	-0.65	Segal et al., 2008
PSS	0.67	Segal et al., 2008
AUDIT	0.33	Dum et al., 2008
DAST	0.26	Dum et al., 2008
DT	0.59	Turner et al., 2012
K-10	0.83	Turner et al., 2012

Note. CES-D = Center for Epidemiologic Studies Depression Scale; STAI-D = State Trait Anxiety Inventory - Depression subscale; HADS = Hospital Anxiety and Depression Scale; BAI = Beck Anxiety Inventory; DASS-A = Depression Anxiety Stress Scale; CATI = Coolidge Axis II Inventory; BHS = Beck Hopelessness Scale; SBQ = Suicidal Behaviors Questionnaire; SPWB = Short Psychological Well-Being Scale; PSS = Perceived Stress Scale; AUDIT = Alcohol Use Disorder Identification Test; DAST = Drug Abuse Screening Test; DT = Distress Thermometer; K-10 = Kessler Psychological Distress Scale.

Anxiety is the second most frequently included construct. The majority of authors do not clarify whether the anxiety measure is intended to provide convergent or discriminant evidence. One exception is the validity study by Osman et al. (1997). The authors investigated the use of the BDI-II with a group of 230 students from a U.S. university, and suggested that all of the statistically significant correlations between the BDI-II with related measures of depression, anxiety, self-esteem, and stress all provided evidence for convergent validity (Osman et al., 1997). Given that anxiety and depression have overlapping symptomology, it is expected that measures of these constructs will produce high correlations that are lower than the correlations between two depression measures. Indeed, based on this review, correlations between the CES-D or BDI-II and measures of anxiety fall in the range of 0.50-0.70, lower than those observed between two measures of depression. This pattern of correlations, high and positive validity

coefficients between measures of the same construct (i.e., two depression measures) and relatively lower validity coefficient between measures of related construct (e.g., depression and anxiety) will illustrate a continuum of convergent and discriminant validity evidence.

Finally, there appears to be a wide variety of disparate constructs and measures used to provide further discriminant validity evidence. Other commonly used constructs and measures used include: (a) hopelessness, as assessed by the Beck Hopelessness Scale, (Osman et al., 2004; Palmer et al., 2008), (b) distress, as assessed by the Kessler's 10-item brief screening scale (Turner et al., 2012), (c) stress, as assessed by the DASS (Osman et al., 1997) or Perceived Stress Scale (Segal et al., 2008), and (d) health functioning, as assessed by the SF-36, (Milette et al., 2010).

Based on this review, convergent and discriminant validity evidence for a depression measure is largely based on (1) another depression measure, (2) an anxiety measures, and (3) a measure of a related construct such as hopeless. This pattern fits the idea of convergent and discriminant validity evidence as existing on a continuum, as suggested by Hubley and Zumbo (2013). Thus, this proposed study will follow a similar pattern of relationships by including three measures to provide evidence for convergent and discriminant validity.

Summary. It is noted that sample sizes varied by the population studied. Generally, community samples appeared to have larger sample sizes than specialized clinical samples such as patients with brain injury. It is noted that, similar to validity studies of the PHQ-9, around half of the validity studies of the BDI-II and the CES-D reviewed are secondary data analyses. Student sample sizes were typically in the range of 137 (Sprinkles et al., 2002) to 575 (Whisman et al., 2000). Studies with university students relied solely on students in psychology classes who completed studies for course credits, with three exceptions (Sprinkles et al., 2002; Lipps et al.,

2007; Makhubela, 2016). Sprinkles et al. (2002) recruited from the university counselling center. Lipps et al., (2007) and Makhubela (2016) recruited students from a wide range of disciplines.

Proposed Study

Although the PHQ-9 has been well studied in the context of medicalized settings, validity evidence for its use with university students is very limited. Based on this review, validity studies of the PHQ-9 in university students in English are limited to a single secondary data analysis (Keum et al., 2018). Thus, no study has been purposefully designed to study the validity of the inferences from, and use of, the PHQ-9 in university students in an English-speaking country. In order to provide evidence to support the use of, and inferences made from, the English PHQ-9 with university students in Canada, the proposed study will examine evidence for internal structure, reliability, and convergent and discriminant validity.

To examine the internal structure of the PHQ-9 in university students, multiple CFAs will be conducted to test a number of models for the PHQ-9 identified in the literature. CFA is appropriate because the factor structure of the PHQ-9 has been previously studied, including in university students (Keum et al., 2018). Specifically, the one-factor model originally proposed by Kroenke et al. (2001) and three two-factor models based on Krause et al., (2010) and Richardson and Richards (2008) will be tested. All of the two-factor models include a somatic factor and an affective/non-somatic factor, but have differences in the items loading on each factor.

To provide an estimate of reliability for the PHQ-9 scores, Cronbach's alpha and ordinal alpha will be used to estimate internal consistency. Cronbach's alpha is the most widely used reliability estimate in the reviewed studies and will allow a direct comparison with existing results. Ordinal alpha, however, will be better able to take into account the ordinal nature of the response format in the PHQ-9 and provide a more accurate reliability estimate.

Convergent and discriminant validity evidence has been well studied in the past. However, many of the results were based on secondary data analyses from existing studies. For instance, Keum et al. (2018) utilized data from a national study. They had to rely on existing data that were designed for a different research purpose rather than being able to select constructs and measures a priori to provide convergent and discriminant evidence. Thus, it is important that the proposed study will incorporate this source of evidence a priori.

The literature review of validation studies of the PHQ-9, BDI-II, and CES-D will guide the selection of constructs and measures. Convergent and discriminant validity evidence exists on a continuum. So, measures that assess the same construct and increasingly dissimilar constructs will be selected. Most of the studies reviewed included another measure of depression to assess convergent validity. In this proposed study, the PHQ-9 will be administered along with the CES-D. Because both are measures of depression, the correlation of scores between these two measures are expected to be highest compared to that of the PHQ-9 with measures of other constructs.

To determine the ability of the PHQ-9 to discriminate between related but different constructs, the correlation between the PHQ-9 and a measure of anxiety will be examined. Anxiety is the most frequently used construct besides depression in the reviewed studies. Because anxiety and depression have some overlapping symptoms, a lower correlation with an anxiety measure relative to a depression measure (i.e., the CES-D) provides evidence that the PHQ-9 is assessing depression rather than anxiety. In a systematic review of anxiety self-report screens, Herr and colleagues (2014) identified the GAD-7 as having the best performance in identifying GAD relative to other measures of anxiety. The GAD-7 is chosen over other anxiety

screeners, such as the BAI or the STAI, due to its brevity and satisfactory psychometric properties.

Health status, as assessed by the Short-Form 12 Health Survey Questionnaire (SF-12), will also provide evidence to support the convergent and discriminant validity continuum. Health status, assessed by the SF-12 or the longer, SF-36, has been used to provide discriminant validity evidence for the PHQ-9 in previous studies (Milette et al., 2010; Kneipp et al., 2010). The SF-12 evaluates both physical health and mental health. Depression is associated with lower physical health outcomes (Rawson, Bloomer, & Kendall, 1994) and lower mental health indicators (Keyes, Dhingra, & Simoes, 2010). Because depression is a mental health diagnosis, it is expected that the correlation between the PHQ-9 and the mental health component of the SF-12 will be higher than that with the physical health score.

Hypotheses

Given the types of validity evidence studied and the measures selected for this study, the following outcomes are hypothesized. It is anticipated that PHQ-9 scores will (a) best fit a one-factor model, (b) be reliable, with Cronbach's and ordinal alpha coefficients greater than 0.80, and (c) correlated highest with another depression screen, followed by next highest correlations with anxiety and mental health functioning, and show a lower correlation with physical health functioning.

Chapter 3: Manuscript[2]

Introduction

Depression is highly prevalent in university students. Ibrahim and colleagues (2013), in their survey of depression prevalence research, reported that prevalence in university students is higher than in the general population and ranges from 10% to 85%. University students experience multiple stressors, such as living independently for the first time, academic stress, relationship challenges, major life decisions, and reduced sleep. Combined, these stressors sharply increase students' vulnerability to depression. University students tend to also be at the peak period of depression onset. The first episode of depression often occurs between the ages of 15 and 24, which includes the age of a typical university student (Blazer, Kessler, McGonagle, & Swartz, 1994). Depression in university students has been linked to a plethora of negative consequences including relationship instability (Whitton & Whisman, 2010), lower self-esteem (Conti, Adams, & Kisler, 2014), lower work performance (Harvey et al., 2011), self-medication (Ford & Schroeder, 2009), and suicide (Furr et al., 2001). University students are in a critical period of development and depression can lead to a cascade of negative outcomes by disrupting social relationships (Whitton & Whisman, 2010), educational attainment (Fletcher, 2008), financial and career outcomes (Saunders et al., 2000), and increasing suicidal thoughts and attempts (Furr et al., 2001). Thus, early detection and effective treatment is vital. Universities are well positioned to provide prevention and treatment since they already encompass many aspects of students' lives, such as social networks, residence, health care, and academics (Mowbray et al., 2006).

[2] This chapter is written as a manuscript and contains some redundancy with the literature review and the conclusion chapters of the thesis.

Quickly administered and effective depression screening measures for university students can assist mental health clinicians and health care professionals to accurately identify depression, which can facilitate more immediate treatment, as well as to help monitor progress. Two of the most commonly used depression screening tools with the university population are the Beck Depression Inventory-II (BDI-II) (Beck et al., 1996) and the Center for Epidemiologic Studies – Depression scale (CES-D) (Radloff, 1977), which are moderately long at 21 items and 20 items, respectively. Lengthy measures add additional burden for depressed clients in clinics and can be daunting for research participants who have to complete a battery of questionnaires. Furthermore, the BDI-II can be costly for university counselling centers and clinics as it is not freely available for use. Thus, there is a need for a short, effective depression screen for both clinical and research purposes. The Patient Health Questionnaire-9 (PHQ-9) is an ideal candidate.

The PHQ-9 is the depression module of the more comprehensive self-administered measure, the Primary Health Questionnaire, a 3-page questionnaire that assesses eight disorders (Spitzer et al., 1999). The PHQ-9 was developed specifically for use as a depression screening tool in fast paced clinics. At just nine items, it is less than half of the length of the BDI-II or CES-D. It can be completed in a few minutes and scored rapidly, which reduces the burden on respondents and lessens time constraints for busy clinics. It is also freely available, which further eases the financial burden on institutions with limited resources. Finally, it is well established (Moriarty et al., 2015) and reflects the diagnostic criteria for depression based on the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5; American Psychiatric Association, 2013). Each item on the scale corresponds with one of the nine symptoms in the DSM-5 diagnostic criteria.

The PHQ-9 has become a popular measure for use with university students for both research and clinical purposes (e.g. Eisenberg et al., 2007; Garlow et al., 2008; Schwenk et al., 2010; Shepardson & Funderburk, 2014). Since its publication, reliability and validity evidence for use of the PHQ-9 has been studied with many patient populations in both primary and secondary care settings (e.g. Arroll et al., 2010; Beard et al., 2016; Dum et al., 2008). However, validity evidence for the inferences made from the PHQ-9 in the university population is scarce. To date, translated versions of the PHQ-9 have been studied with university students in China, Japan, and Nigeria. Adewuya, Ola, and Afolabi (2006) studied the utility of the PHQ-9 with Nigerian university students and reported a one-factor structure, good reliability (Cronbach's alpha = 0.85), and test-criterion related evidence. The PHQ-9 has also been found to have good reliability (Cronbach's alpha = 0.80 to 0.85), convergent validity evidence, and test-criterion related evidence with Chinese university students (Zhang et al., 2013; Du et al., 2017). In Japan, Umegaki and Todo (2016) used an item response theory model to compare three depression scales: the Zung Self-Rating Depression Scale, the CES-D, and the PHQ-9. The authors found the PHQ-9 to perform better compared to the other depression measures, likely due to the absence of negatively worded items. Validity evidence for the English version of the PHQ-9 for use with university students is limited to a single secondary analysis in the U.S. (Keum, Miller, & Inkelas, 2018). Keum et al. (2018) found support for a unidimensional factor structure, evidence for good reliability of scores (Cronbach's alpha = 0.89), and good convergent and discriminant evidence that showed higher correlations with mental health measures than with alcohol use. There has yet to be a validity study, however, that was purposefully designed to provide evidence to support the use of the English version of the PHQ-9 with university students.

The present study examined internal structure, internal consistency reliability, and convergent and discriminant evidence for the PHQ-9 to evaluate its use with university students in Canada. Multiple confirmatory factor analyses (CFA) were conducted to examine the internal structure of the PHQ-9 with university students based on models previously examined in the extant literature for the PHQ-9. Numerous studies have examined the internal structure of the PHQ-9 with mixed results. In the original validity study, Kroenke and colleagues (2001) suggested that the PHQ-9 had a one factor structure with a group of primary care and obstetrics/ gynecology patients. The one-factor structure has been supported with multiple other populations in later studies, including patients with substance abuse (Dum et al., 2008), depression and anxiety (Ryan et al., 2013; Titov et al., 2010), and university students (Keum et al., 2018). However, other studies have identified two-factor structures comprised of various somatic and non-somatic symptoms as the best fit for the PHQ-9 in psychiatric settings (Beard et al., 2016), palliative care (Chilcot et al., 2013), and spinal cord injury (Kause et al., 2010). A unidimensional model and three two-factor models were examined in the present sample.

To examine score reliability of the PHQ-9, Cronbach's alpha and ordinal alpha were used to estimate internal consistency. Cronbach's alpha is the most widely used reliability estimate and allows a direct comparison with existing studies. The PHQ-9 has been studied with various medical populations, and studies have generally reported good reliability with Cronbach's alpha in the range of 0.81 to 0.91 (e.g. Dum et al., 2008; Roony et al., 2013). With an American university student sample, Keum et al., (2018) also reported good reliability, with Cronbach's alpha of 0.89. Ordinal alpha, however, is better able to take into account the ordinal nature (i.e., less than 5 points) of the response format for the PHQ-9 and provide a more accurate reliability

estimate. It was anticipated that PHQ-9 scores will show adequate internal consistency, with Cronbach's alpha and ordinal alpha ≥ .80.

In determining constructs and measures to provide convergent and discriminant validity evidence, a review of literature of validity studies of the PHQ-9 from 2008-2018 was conducted. In the reviewed studies, the most frequently used constructs in providing convergent and discriminant validity evidence were depression and anxiety.

The majority of studies that examined convergent and discriminant validity evidence included another depression measure. High correlations between measures of the same construct provide strong support for convergent validity. The PHQ-9 has been most frequently compared to the BDI-II (e.g. Dbouk et al., 2008; Dum et al., 2008), the CES-D (e.g. Beard et al., 2016, Pilkonis et al., 2014), as well as the depression subscale of the Hospital Anxiety Depression Scale (HADS) (Cameron et al., 2008; Haddad et al., 2009), and the Zung Self-Depression Scale (SDS) (Turner et al., 2012; Titov et al., 2010). The correlations vary across studies, generally falling between 0.70 and 0.85.

Anxiety was the second most frequently included construct. The majority of authors did not clarify whether the anxiety measure was intended to provide convergent or discriminant evidence. The PHQ-9 has been compared with the GAD-7 (Beard et al., 2016; Chilcot et al., 2018; Ryan et al., 2013), Beck Anxiety Inventory (BAI) (Kneipp et al., 2010), and the anxiety subscale of HADS (Cameron et al., 2008). Given that anxiety and depression have overlapping symptomology, it is expected that measures of these constructs will produce high correlations that are lower than the correlations between two depression measures. Indeed, based on this review, correlations between the PHQ-9 and measures of anxiety generally fall in the range of 0.50-0.70, lower than those observed between two measures of depression.

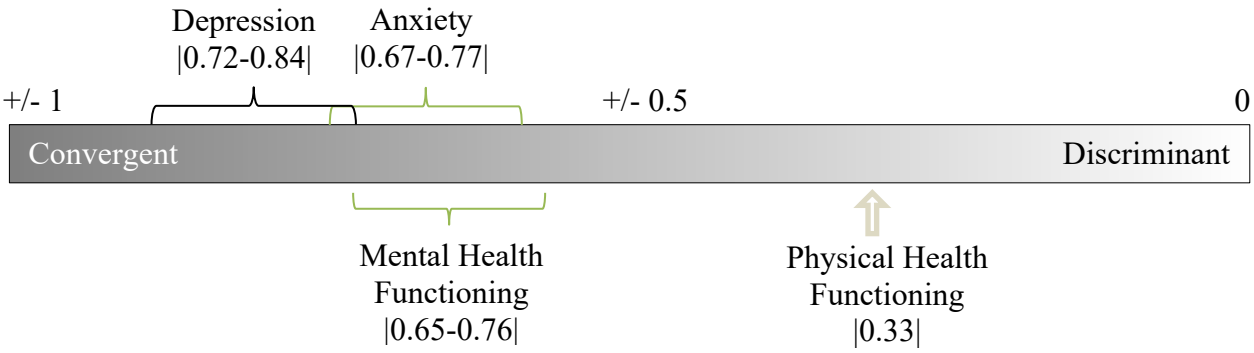
Finally, there was a wide variety of disparate constructs and measures used to provide further discriminant validity evidence. Examples of other commonly used constructs and measures used included: mental health functioning (0.68-0.76) and physical health functioning (0.33-0.43), as assessed by the SF-36 or the abbreviated SF-12, (Milette et al., 2010; Kneipp et al., 2010); stress (0.61), as assessed by the Perceived Stress Scale (Kenipp et al., 2010); and distress (0.83), as assessed by the Kessler's 10-item brief screening scale (Turner et al., 2012).

Based on this review, convergent and discriminant validity evidence for a depression measure was largely based on (1) another depression measure, (2) an anxiety measure, and (3) a measure of a more and less related constructs such as mental and physical health functioning, respectively. This pattern fits the idea of convergent and discriminant validity evidence as existing on a continuum, as suggested by Hubley and Zumbo (2013). The continuum of convergent and discriminant validity evidence focuses on the pattern of relationships between measures, rather than defining whether a measure is strictly convergent or discriminant. This study followed a similar pattern of relationships by including three measures to provide convergent and discriminant evidence. Specifically, scores from the PHQ-9 were compared to those of the CES-D, GAD-7, and SF-12. The correlation of the two depression measures, the PHQ-9 and CES-D, was expected to be strong and positive, and importantly, the highest. Next, the correlation between the PHQ-9 and the GAD-7 was also expected to be strong and positive, given that anxiety and depression have overlapping symptoms. It is also important that this correlation be relatively lower than that of the PHQ-9 and CES-D scores, to provide evidence that the PHQ-9 is assessing depression, rather than anxiety. Finally, the SF-12 evaluates both physical health and mental health functioning. Because depression is a mental health diagnosis, it was expected that the correlation between the PHQ-9 and the mental health component of the

SF-12 (MCS) would be higher than that with the physical health score (PCS). This convergent and discriminant evidence for validity is illustrated on a continuum in Figure 3.1.

Figure 3.1

Expected pattern of convergent and discriminant validity evidence for the PHQ-9



Methods

This study employed a web-based survey. This method was chosen because (a) university students are generally proficient in computer use and have good Internet access, and (b) this study uses measures related to depression and anxiety, which are potentially sensitive topics. Web-based surveys increase honest reporting of sensitive information compared to telephone interviews (Kreuter, Presser, & Tourangeau, 2008), and are associated with more socio-economically and ethnically diverse sampling, while remaining equally effective as in-person testing (Casler, Bickel, & Hacket, 2013). Particularly for mental health related self-report measures, the online survey method is associated with lower social desirability scores compared to face-to-face surveys (Henderson et al., 2012).

Participant Recruitment

University students ages 18 years or older and currently enrolled in an undergraduate or graduate program were recruited. Both undergraduate and graduate students were included because graduate students have been underrepresented in previous studies. Both full-time and

part-time students were also included. Importantly, all of these students access the same mental health services on campus. Moreover, university clinics and counselling centers do not differentiate among these groups in terms of treatment. It is important to show that inferences made from PHQ-9 scores are valid and appropriate for use with a range of students, and not just full-time undergraduate students alone. Exclusion criteria thus were age younger than 18 years, not enrolled in a university, and unable to speak, read, or write fluently in English. Data collection was conducted in the Greater Vancouver Area, and participation was limited to university students in this area.

Participant recruitment was completed through two main channels. First advertisements were posted across local university campuses in a variety of locations, including classrooms, student recreation centers, as well as university counselling centers. Online postings advertising the study were also made on social media sites such as Facebook and Twitter. Additionally, online postings appeared on social media platforms, such as Facebook. As an incentive, participants were entered into a draw at the end of the study, for one of 15 $20 CAD gift cards, by being redirected to another web page to enter their contact information.

Procedure

Participants were directed to an online survey hosted on https://ubc.qualtrics.com. Upon accessing the survey, the participant read a cover letter explaining the nature of the study, the procedure, time commitment, ability to withdraw, and the confidentiality and security of their data. By completing and submitting the questionnaires, participants consented to participate in the research. To minimize fatigue or practice effects due to order effects, the order of measures presented to the participants was randomized with the Randomizer option on Qualtrics.

Additionally, Evenly Present Elements on Qualtrics kept count to ensure that questionnaires were randomized evenly.

Ethical considerations

This study used an anonymous online survey. Given that the current study included questionnaires about negative experiences and moods, a concern was that participation might induce negative mood during or after participation. Participants were informed that they had the right to skip any questions they do not want to answer. Additionally, information on community-based counselling services and resources available in the Greater Vancouver area as well as online resources were provided to all participants at the end of the survey. Participants were informed that, due to the anonymity of the survey, the researchers could not identify participants with concerning levels of depression, anxiety, distress or other concerns and thus could not intervene in any way.

Measures

Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001). The PHQ-9 is a 9-item, self-report screen that assesses depressive symptoms over the past two weeks. Each of the nine items reflects a symptom for major depressive disorder (MDD) as outlined in the DSM-5. Respondents are asked to rate how often they have experienced each of the symptoms using a 4-point Likert-type response format: 0 (not at all), 1 (several days), 2 (more than half of the days), and 3 (nearly every day). Total scores range from 0 to 27. Higher scores indicate higher severity of depression, with cut scores of 5, 10, 15 and 20 representing mild, moderate, moderately severe, and severe levels of depression (Kroenke et al., 2001). A total score of greater than 10 is indicative of MDD (Moriarty et al., 2015).

Center for Epidemiological Studies-Depression (CES-D; Radloff, 1977). The CES-D is a 20-item, self-report measure that assesses the frequency of depressive symptoms in community samples. The CES-D covers a broad range of depressive symptoms, including cognitive, behavioural, somatic and interpersonal symptoms. Respondents are asked to report the frequency of each depressive symptom that they experienced during the past week using a 4-point Likert-type response format, ranging from 0 (rarely or none of the time) to 3 (most or all of the time). The total score ranges from 0-60 and higher scores indicate a greater level of depression. A standard cut-off score of 16 indicates a high probability of clinical depression (Andresen et al., 1994).

Generalized Anxiety Disorder 7-item scale (GAD-7; Spitzer et al., 2006). The GAD-7 is a 7-item, self-report anxiety screen based on the DSM-5 criteria. The GAD-7 assesses the frequency of anxiety symptoms experienced over the past two weeks using a 4-point Likert-type response format: 0 (not at all), 1 (several days), 2 (more than half of the days), and 3 (nearly every day). The total score can range between 0 and 21. Higher scores indicate greater generalized anxiety. Scores of 0-4, 5-9, 10-14, and 15-21 represent minimal, no anxiety, mild, moderate, and severe anxiety symptoms levels, respectively. The GAD-7 was originally developed as a screener for generalized anxiety disorder (GAD) with a large sample of patients in a primary care setting (Spitzer, Kroenke, Williams & Lowe, 2006).

The Short-Form 12 Health Survey Questionnaire (SF-12; Ware et al., 1996). The SF-12 is a widely used 12-item self-report measure that assess functional health status. It is abbreviated from the Short-Form 36 Health Survey Questionnaire (SF-36) as a shorter alternative for both general and specific populations to assess health outcomes. The SF-12 produces two scores: a physical component summary score (PCS) and a mental component summary score

(MCS) (Ware et al., 1995). Both subscales are scored using all 12 items and use a norm-based scoring method based on norms from the U.S. general population. The scores are expressed as T-scores with mean of 50 and standard deviation of 10 (Ware et al., 1995). Higher scores indicate better functional health.

Demographics. In order to describe the sample, participants were asked to complete a demographics questionnaire to provide information on age, racial background, gender identity, area of study, and type of degree (undergraduate, graduate, or other).

Data analysis

Confirmatory factor analyses (CFA) were used to examine the internal structure measurement models for the PHQ-9. Robust weighted least-squares (WLS) estimation, an approach recommended for ordinal data (Flora & Curran, 2004), was used. The fit of the one factor and three competing two-factor models presented in Figure 3.2 were tested. Fit indices were relied upon to evaluate the model fit. Absolute fit was assessed using the root-mean-square error of approximation (RMSEA) and the standardized root-mean-square residual (SRMR), while comparative fit was assessed with Bentler's comparative fit index (CFI). Suggested guidelines are: RMSEA < 0.08 for acceptable fit, SRMR < 0.08 for acceptable fit, and CFI > 0.95 for good fit (Hu & Bentler, 1999).

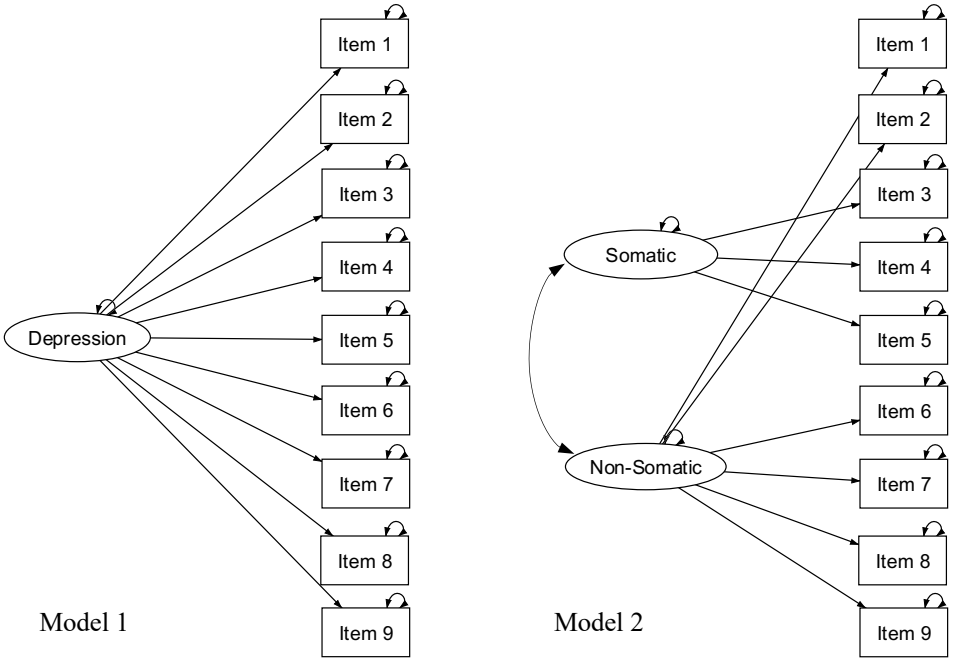
To examine internal consistency reliability of the PHQ-9 scores, both Cronbach's alpha coefficient and ordinal alpha were calculated for the PHQ-9. This is the first study, to authors' knowledge, that has reported ordinal alpha for the PHQ-9. Unlike the Cronbach's alpha, the ordinal alpha is based on the polychoric correlation matrix, which takes into consideration the ordinal nature of the response format (i.e., 4-point response options) and provide a more accurate estimate for ordinal scales (Zumbo et al., 2007; Gadermann, Guhn & Zumbo, 2011). Although

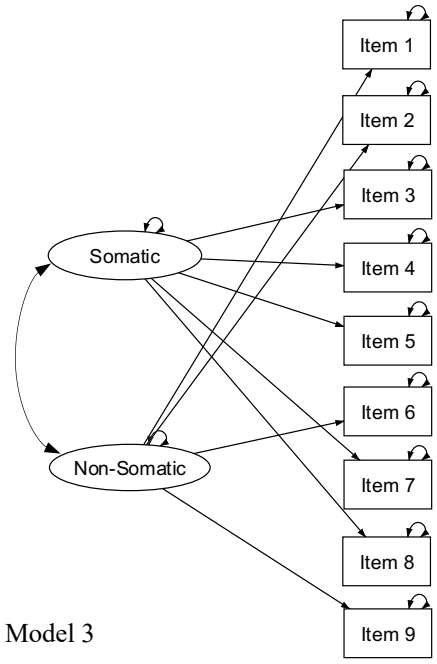
ordinal alpha is more appropriate for ordinal data, reporting Cronbach's alpha allows for comparison with previous studies. Internal consistency estimates of reliability were also examined for the convergent and discriminant measures in the study.

To examine convergent and discriminant validity evidence, Pearson's correlations (*r*) were calculated between scores from the PHQ-9 and those from the CES-D, GAD-7, and SF-12

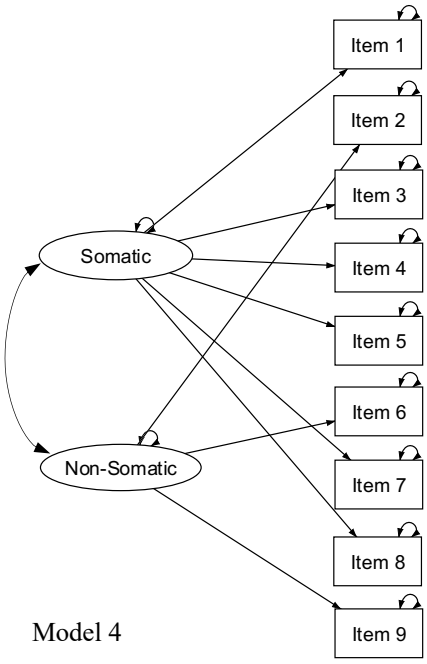
Figure 3.2

Internal structure models for the PHQ-9 to be tested with university students





Model 3



Model 4

Model 1 is based on Kroenke et al. (2010). Model 2 is based on Krause et al. (2010) baseline measurement and Chilcot et al. (2013). Model 3 is based on Krause et al. (2010) 17 months post-injury and Beard et al. (2016). Model 4 is based on Krause et al., (2010) 29 months post-injury.

Results

Sample Characteristics

The sample consisted of 204 students currently attending a post-secondary school in the lower mainland of British Columbia, Canada. Eight participants did not complete the demographics questionnaire but were included in the analyses because they completed other parts of the survey. Participants ranged in age from 18-45 years (M = 25.3, SD = 5.25). The majority of the participants were female (68.4%) of either Asian (46.6%) or White (39.9%) ethnicity, and undergraduate students (52.3%). Participants came from a range of academic disciplines. Most respondents (79.2%) did not report a previous diagnosis of depression. A more

detailed description of the sample who completed the demographics form (N = 196) is provided in Table 3.1.

Table 3.1

Sample Demographics

Demographic Variables	Frequency (Percentage)
Gender	
Male	59 (30.1%)
Female	134 (68.4%)
Other	3 (1.5%)
Racial background	
White	77 (39.9%)
Black	2 (1.0%)
Asian	90 (46.6%)
Other	24 (12.4%)
Degree	
Undergraduate	102 (52.3%)
Graduate	76 (39.0%)
Other	17 (8.7%)
Area of study	
Science	48 (27.0%)
Arts	42 (23.6%)
Education	36 (20.2%)
Business	16 (9.0%)
Medicine	14 (7.9%)
Other [a]	22 (12.4%)
Previous depression diagnosis	
Yes	41 (20.9%)
No	155 (79.1%)

Note. Not all cells add up to 196, as students could choose to skip questions that they did not want to answer.

[a] Other: contains a range of areas of study, including kinesiology, nursing, engineering, criminology, law, land & food system, and earth & ocean science.

Internal Structure: CFA

Table 3.2 provides a summary of the fit indices results for each of the four initially proposed models. Determining which model fits the data best is an important first step as it

determines the scores that should be used in subsequent reliability and validity analyses. Overall, all four models generally showed good fit based on CFI, TLI, and SRMR values. The critical difference among the models was based on the RMSEA values wherein only Model 2 showed an adequate fit based on the obtained RMSEA value of 0.075 (i.e., < 0.08). Notably, all three proposed two-factor models produced very high factor intercorrelations between 0.85 and 0.91, which suggested that there is little practical and statistical difference between any two factors.

Because Model 2 demonstrated the most adequate fit across all four fit indices, we examined Model 2 more closely. Previous literature suggested that the two-factor models represented somatic and non-somatic symptoms (Elhai et al., 2012; Keum et al., 2018). For Model 2, it has been suggested that items 3, 4, and 5 are somatic items, and the rest of the items represent non-somatic symptoms. However, item 8 (*Trouble concentrating on things, such as reading the newspaper, or watching television*), theoretically, should be generally considered a somatic symptom. Thus, using theory and based on examining the item content, we tested an additional two-factor structure, Model 5, with items 3, 4, 5, and 8 representing a somatic factor and items 1, 2, 6, 7, 9 representing a non-somatic factor. As seen in Table 3.2, Model 5 showed a pattern of fit similar to that of the one-factor model, with the RMSEA value being higher than the recommended cut-off. Because this more theoretically based model did not improve the model fit over Model 2, we re-examined the items in Model 2 and suggest that the names of the two factors are mislabelled. Upon closer examination, items 3 (sleep disturbance), 4 (change in energy level), and 5 (change in appetite) represent symptoms that are most commonly shared across a number of conditions and disorders. Thus, we suspect that these items converge as a factor because they are less specific to depression than the other items. As such, there may be

two clusters of symptoms that differ in terms of their specificity to depression but are not necessarily meaningful to separate in terms of scoring.

In conclusion, all of the models (i.e., unidimensional and two-dimensional) show similarly good fit except in terms of the RMSEA values. Given that all of the two factor models show high intercorrelations between the factors, that existing factor labels may misrepresent theoretical differences between symptoms as somatic and non-somatic, and that the best-fitting model overall (i.e., Model 2) may be described as having factors representing more vs. less specific depressive symptoms, use of a single total score seems the most efficient and reasonable for use in clinical and research settings.

Table 3.2

Confirmatory factor analysis results for each model

Model	CFI	TLI	RMSEA	SRMR	Inter-factor correlation
1[a]	0.970	0.960	0.100	0.069	N/A
2[b]	0.984	0.977	0.075	0.057	0.849
3[c]	0.976	0.967	0.090	0.066	0.907
4[d]	0.976	0.967	0.091	0.064	0.904
5[e]	0.974	0.963	0.096	0.069	0.920

Note. CFI = comparative fit index; TLI = Tucker Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual.

[a] Model 1: Unidimensional
[b] Model 2: Factor 1 = items 3, 4, 5; Factor 2 = 1, 2, 6, 7, 8, 9.
[c] Model 3: Factor 1 = items 3, 4, 5, 7, 8; Factor 2 = 1, 2, 6, 9.
[d] Model 4: Factor 1 = items 1, 3, 4, 5, 7, 8; Factor 2 = 1, 6, 9.
[e] Model 5: Factor 1 = items 3, 4, 5, 8; Factor 2 = 1, 2, 6, 7, 9

Reliability Analyses and Observed Score Descriptives

Reliability estimates for, and mean performance on, the PHQ-9 and the other measures used in this study are reported in Table 3.3. In particular, the ordinal alpha for the PHQ-9 total

score was excellent at 0.91. Based on Model 2, reliability estimates for the two PHQ-9 subscale scores (depression-specific vs. not) for the PHQ-9 were calculated. The lower ordinal alpha values of 0.88 and 0.82 for these two factors, respectively, are not surprising given (a) the fewer number of items in each subscale as well as (b) the more generic symptoms in the second (non-specific) subscale. Furthermore, the higher reliability estimate for the total score provides more confidence in the consistency of this score. Thus, the total score will be used when examining convergent and discriminant evidence.

Table 3.3

Observed score descriptive and internal consistency reliability for all study measures

	Possible Score Range	Obtained Score Range	***M***	***SD***	% Over Cut-score	Cronbach's α	Ordinal α
PHQ-9 Total Score	0-27	0-27	7.04	5.59	18.32%	0.87	0.91
PHQ-9 Subscale 1 (depression specific)	0-18	0-18	3.77	3.64	N/A	0.83	0.88
PHQ-9 Subscale 2 (non-specific)	0-9	0-9	3.27	2.46	N/A	0.78	0.82
CES-D	0-60	0-57	17.64	11.70	49.25%	0.92	0.94
GAD-7	0-21	0-21	6.73	5.44	27.50%	0.91	0.93
SF-12: MCS	0-100	7.37-58.45	39.21	10.98	N/A	N/A	N/A
SF-12: PCS	0-100	34.44-66.74	54.85	5.90	N/A	N/A	N/A

Note: PHQ-9 = Patient-Health Questionnaire-9; CES-D = Center for Epidemiological Studies Depression; GAD-7 = Generalized Anxiety Disorder-7; MCS = Mental health Composite Score; PCS = Physical Health Composite Score; N/A = not applicable.

The mean PHQ-9 score was 7.04, and under 20% of the sample scored over 10 on the PHQ-9 (n=37, 18.3%), a standard cut-off score for moderate severity of depressive symptoms.

The mean CES-D score was 17.67, and over two-thirds of the sample had a score over 16 (n=98, 49.3%), a standard cut-off score for possible depression, and around one-third had a score over 23 (n=66, 33.2%), a standard cut-off score for probable depression. The mean GAD-7 score was 6.73, and over one-quarter of the sample scored over 10, a standard cut-off score for moderate anxiety (n=55, 27.5%).

Convergent and Discriminant Evidence for Validity

A summary showing the obtained pattern of convergent and discriminant validity coefficients is presented in Table 3.4.

Table 3.4

Correlations between the PHQ-9 total scores and scores on other measures

Measure	CES-D	GAD-7	SF-12: MCS	SF-12: PCS
PHQ-9	0.86**	0.74**	|0.72|**	|0.12|

*Note. *p < 0.05, **p < 0.01*

Discussion

This the first study that was designed to examine the validity evidence for the English version of PHQ-9 for use with university students. Findings from this study provide evidence in terms of internal structure, reliability, and convergent and discriminant evidence of validity to support the interpretation and use of the English version of PHQ-9 with Canadian university students.

Consistent with some previous studies (e.g. Keum et al., 2018), our findings provided some additional support for the original unidimensional structure (Kroenke et al., 2001). We tested four models proposed in the literature using CFA. Many previous factor analyses used maximum likelihood estimation (e.g. Krause et al., 2008; Elhai et al., 2012), which does not take into account the ordinal nature of the response scale. Instead, we used the robust WLS

estimation, which is recommended for ordinal data (Flora & Curran, 2004). Even though the two-factor models attempted to provide a useful theoretical distinction between somatic and non-somatic symptoms, they were not consistent in terms of what qualified as a somatic vs non-somatic symptom. Furthermore, consideration of the items suggested that the best fitting two-factor model (Model 2) is better described as depression-specific symptoms vs more generic symptoms. Even still, the very high inter-factor correlation as well as the lower internal consistency reliability estimates do not provide compelling support for a two-factor structure over a unidimensional structure. Altogether, the unidimensional structure provides the most measurement and clinical utility. This is consistent with findings from Keum et al., (2018), which studied the use of the English PHQ-9 in U.S. university students based on a secondary data analysis.

Further evidence to support the use of the PHQ-9 total score with Canadian university students is provided by the high reliability estimate. The Cronbach's alpha of 0.87 and ordinal alpha of 0.91 for the PHQ-9 total score indicates good reliability and corresponds to that found in previous research with university students (Keum et al., 2018).

Many previous studies examining convergent and discriminant evidence for validity for the PHQ-9 were based on secondary data analysis from existing studies or data sets. Existing data are often designed for a different research purpose and thus did not select constructs and measures a priori to provide convergent and discriminant evidence. A strength of the present study was that constructs and measures were selected a priori using both a theoretically and empirically guided approach. We found the expected pattern of convergent and discriminant evidence for the PHQ-9 by demonstrating a higher correlation between the total scores on the PHQ-9 with the CES-D (another measure of depression), relative to measures of similar

constructs (GAD-7 for anxiety and SF-12 MCS for mental health functioning), and a much lower correlation with the SF-12 PCS (measuring the far more disparate construct of physical health functioning). It is important that the PHQ-9 demonstrated a stronger correlation with another depression measure relative to the anxiety measure. While anxiety and depression have some overlapping symptoms and are expected to correlate to a fairly high degree, a lower correlation with an anxiety measure relative to a depression measure provides evidence that the PHQ-9 is assessing depression rather than anxiety. As expected, the PHQ-9 also had a much stronger correlation with the SF-12 MCS than the SF-12 PCS. It is also important to point out that all of the measures used in this study for which internal consistency reliabilities could be computed showed satisfactory reliability estimates and thus would not cause notably attenuated validity coefficients due to measurement error.

Strengths and Limitations

The PHQ-9 was originally designed for use in primary care but is used by some university counselling centres. Validity evidence to support its use with this population, however, remains extremely limited. As only the second study to examine evidence of validity for the use of the English version of the PHQ-9 in university students in North America and the first to design the study a priori for this purpose rather than conduct a secondary data analysis, our study findings address an important gap in the literature. Our student sample is based on a general student population, inclusive of both undergraduate and graduate students, as well as students from a wide range of disciplines.

Because this study was designed from the start to be a validity study, constructs and measures to support convergent and discriminant evidence of validity were purposefully selected

and based on a review of validity research with the PHQ-9 and other common depression measures. Several measures were included to provide multiple points of comparison for illustrating the convergent and discriminant continuum, which allowed us to examine critically the inference that the PHQ-9 is a measure of depression, rather than a measure of a related construct, such as anxiety or general mental health functioning. Finally, use of an anonymous web-based survey helped to reduce social desirability bias compared to in-person or telephone administered formats (Henderson et al., 2012).

There were several limitations to this study. First, the volunteer nature of the study increased self-selection bias. Students with a stronger interest in the topic, whether due to personal experience with depression or an interest in learning more about such symptoms, were probably more likely to participate. There was a larger proportion of female students, which is likely due to self-selection bias given that women are more likely to volunteer for research and to disclose and discuss mental health related issues (Kent, Philip, Zimbardo, & Boyd, 2010; Smith, 2008). Finally, while the current sample size is sufficient for the analyses conducted in this study (e.g., CFA), it was not sufficient to enable subgroup analyses, such as internal structure or measurement invariance by educational level or gender.

Future Directions

The evaluation of reliability and validity evidence for a measure is an ongoing process. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) describes five sources of validity evidence (i.e., internal structure, test content, relations to other variables, response processes, and test consequences) in addition to several types of reliability evidence. The most critical forms of evidence needed next to support the use of the English

version of the PHQ-9 with university students include test-criterion evidence and further internal structure evidence.

Future studies should investigate test-criterion validity evidence in a sample of university students to determine the ideal cut-off score in this population. Test-criterion evidence assesses the ability of the PHQ-9 to accurately identify depression compared to a criterion and confirm the ideal cut score for university students. Diagnostic interviews, such as the Structured Clinical Interview for DSM (SCID), have been used as a criterion to examine test-criterion evidence for the PHQ-9 in multiple medical populations (e.g., Fann et al., 2005; Rooney et al., 2013; Turner et al., 2012). Thus, future studies should incorporate a diagnostic interview as a criterion in a study with university students to empirically evaluate the appropriateness of the typically recommended cut-off score of 10 in the student population.

The factor analysis results from the present study replicated the unidimensional structure reported in many medical populations (Dum et al., 2008; Hepner et al., 2009; Ryan et al. 2013) as well as in the previous secondary analysis with university students (Keum et al. 2018). Our results suggested that both the unidimensional and the two-factor structure provided an adequate fit to our data. In the present case, the unidimensional structure is preferred because it provides more clinical utility. However, the two-factor structure has provided the best fit to the data in our study, as well as in previous studies in a variety of populations, including university students (Keum et al., 2018), palliative care patients (Chilcot et al., 2013), psychiatric patients (Beard et al., 2016), and spinal cord patients (Richards & Richards, 2008). Given the mixed results, it is unsurprising that the unidimensional and the two-factor structure both provided adequate model fit in our case, and future studies should continue to evaluate the internal structure of the PHQ-9

in university students. It is recommended that future factor analyses to take into account the ordinal nature of the response format and use an appropriate estimation method, such as weighted least squares estimation. In addition to testing the current one-factor and two-factor models, future work might also examine a hierarchical structure with an overarching factor (depression) and two subordinate factors in a study with a larger sample size.

The ordinal nature of the response format is also an important consideration when examining the reliability of the measure. Future studies should continue to report ordinal alpha when investigating the internal structure of this measure. Another avenue for future assessment of reliability in the PHQ-9 is test-retest reliability, which has yet to be explored with university students. Given the potentially transitory nature of depression, it is recommended that duration between test and retest be kept to a short period of time (e.g., 1 week).

Future studies that examine convergent and discriminant validity evidence should continue to include another measure of depression and of anxiety but also explore relationships with other related measures. One relevant construct for university students would be stress, which has been used as a discriminant measure in previous validity studies on depression measures (Arbona et al., 2017; Segal et al., 2008). Given that higher education is associated with multiple sources of stress (e.g., academic stress, relationship challenges), it would be worthwhile include a measure of stress as a discriminant measure to ensure that the PHQ-9 is assessing depression rather than stress.

Existing validity evidence has focused on internal structure and relations to other variables, which includes test-criterion and convergent and discriminant evidence. Future studies should also explore new areas of validity evidence, in particular, response processes. Response

processes evidence evaluates the cognition, behaviours, and emotions that respondents engage in when completing a measure (AERA et al., 2014; Hubley & Zumbo, 2017). When the observed processes match expected ones for a construct, it is considered to provide evidence to support the use of the measure. Additionally, response processes evidence can provide insight on how students approach items in the PHQ-9. Identifying similarities and differences in response processes between items can provide additional insight about factor structure.

Concluding Statement

The results from the present study supported the reliability of scores and validity of inference made from the PHQ-9 as a depression measure for use with university students. Given its brevity, cost effectiveness, and DSM-5 based approach, we recommend the use of PHQ-9 with university students in clinical and research applications.

Chapter 4: Conclusion

Summary of Purpose and Findings

University students are in a unique and challenging period in their development. They experience stress from a number of sources, such as navigating independent living, establishing relationships, and academic and career pressures. Unsurprisingly, rates of depression in this population are higher compared to the general public. Having a brief depression screen with good psychometric properties is a key first step to better identify students experiencing depressive symptoms and depression disorders and support them through a challenging time, as well as being key in depression research. Compared to other commonly used depression measures, such as the BDI-II or the CES-D, the PHQ-9 provides several important advantages, as it is brief, cost-effective, and based on the current DSM-5 diagnostic criteria.

The present study was designed to examine the reliability of scores and validity of inferences made from the PHQ-9 as a depression measure for use with undergraduate and graduate university students. This research addressed a gap in the extant literature as only one other study has examined the psychometric properties of the English version of the PHQ-9 with university students and this was a secondary analysis. Findings from the present study provided support for the unidimensional structure of the PHQ-9. The alternative two-factor structures produced very high inter-factor correlations (suggesting redundancy in including two factors), lower reliability estimates (suggesting greater measurement error), and were of limited theoretical and practical value. Taken together, the unidimensional structure provided the most measurement and clinical utility. In this study, the PHQ-9 produced reliable total scores, as evidenced by high internal consistencies based on both Cronbach's α (0.87) and ordinal α (0.91).

The PHQ-9 also showed the expected pattern of convergent and discriminant validity coefficients wherein the PHQ-9 showed the highest correlation with another depression measure, followed by a similar but lower magnitude of correlation with measures of anxiety and mental health functioning, and the lowest correlation with the physical health functioning measure. These results suggest that the PHQ-9 is measuring depression rather than other closely related constructs, such as anxiety or general mental health functioning. Based on the psychometric results from this study, the PHQ-9 is tentatively recommended for use with university students.

Areas for Future Research

The evaluation of validity evidence is a continuous, ongoing process. The present study provided evidence in the terms of internal structure, reliability, and convergent and discriminant validity. With regard to future validity evidence for interpretation and use of PHQ-9 total scores with university students, the two most critical sources would be the test-criterion evidence and further internal structure evidence.

A test-criterion study would be valuable in examining the efficacy of the PHQ-9 in predicting a criterion or outcome measure. Importantly, test-criterion evidence would examine the appropriateness of 10, the most common recommendation, as a cut-off score for the PHQ-9 in university students. Cut-off scores indicate the probability of a depressive disorder. Having evidence for the use of an appropriate cut-off score for university students will help counselling centers, doctors' clinics, or any other assessment and monitoring on depression with university students to better and more confidently identify students in distress.

In the present study, the PHQ-9 and the CES-D identified different proportions of students as having probable depression based on respective standard cut-off scores (18.3% from

PHQ-9 versus 49.3% from the CES-D). It would be advantageous to conduct a future validity study of the PHQ-9 comparing results to those from the CES-D, as well as the BDI-II, another commonly used depression measure with university students, concurrent with a diagnostic interview. By directly examining and comparing the diagnostic performance of the commonly used depression measures, future studies can clarify the appropriateness of different cut-off scores for each measure as well as examine their ability to differentiate depressed students in terms of sensitivity and specificity. Such evidence will allow clinicians and researchers to be more informed of any overestimation or underestimation based on standard cut-off scores and assist them in their decision in choosing an appropriate depression measure.

In regard to the factor structure of the PHQ-9 with university students, future research with a larger sample size would be valuable to confirm the appropriateness of the use of a total score. With a larger sample size, future research can re-examine currently tested models, including the unidimensional structure as well as the four 2-factor models to provide further confirmation for the unidimensional structure. Additionally, given that a two-factor structure demonstrated adequate model fit, future studies might examine a hierarchical structure with an overarching factor (depression) and two subordinate factors in a study with a larger sample size. A hierarchical structure will still support the use of a total score but will provide evidence for the use of subscale scores as well. The use of subscale scores in assessing and monitoring depressive symptoms in university students might provide additional details about symptom types to address for clinicians working with depressed students as well as provide a more detailed information for university policy makers wishing to survey and address issues around depression for university students.

Response processes is another source of validity evidence worth exploring, particularly with university students. Based on the literature review for this study, there is yet to be a study that examines response processes as a source of validity evidence for the PHQ-9. For university students in particular, evidence based on response processes, which can examine thoughts, behaviours, motivations and emotions that people engage in when responding to a test item, can provide information on how students approach items in the PHQ-9. Understanding if students are engaging in different processes for different items can provide additional evidence to further provide understand any obtained factor structure.

This is the first study to report reliability in terms of both the Cronbach's alpha and the ordinal alpha. Future studies that investigate the use of the PHQ-9 should also take into account the ordinal nature of the response format and use ordinal alpha (or alternative estimates such as ordinal omega). Additionally, other ways of assessing reliability, such as test-retest reliability, have yet to be explored with university students, and can be an avenue for future psychometric work.

Counselling Implications

As a brief assessment, the PHQ-9 can be effective as an initial clinical assessment as well as in progress monitoring with university students in counselling centers and doctors' clinics. Other frequently used measures are at least double the length, which adds an additional burden for depressed clients. In clinical practice, the brief nature of the PHQ-9 can not only minimize burden on depressed students, it can also potentially reduce perceived barriers in accessing services and increase willingness to seek mental health support. Brief questionnaires can provide clinicians with valuable information on student distress without making the assessment taxing on

the students. In research, lengthy measures are also daunting for participants who have to complete a battery of questionnaires. Thus, brief assessments can facilitate higher survey completion rates.

Additionally, an efficient and accessible screen such as the PHQ-9 in counselling practice can act as an avenue for clinicians to explore symptoms of depression with clients with suspected depression and facilitate discussions around related experiences. As the PHQ-9 is based on DSM-5 diagnostic criteria, it covers agreed-upon depression symptoms. These conversations can be particularly beneficial for clients new to counselling or who have difficulty navigating how to discuss their experiences.

Counselling of university students is becoming increasingly important with increased awareness that this population is disproportionally impacted by mental health challenges. Counsellors are a key aspect of depression treatment in university students through offering psychoeducation on prevention and treatment, as well as being the front-line workers in identifying and treating depressive symptoms experienced by university students. Using screening tools with adequate reliability and validity evidence with this population is critical in properly identifying those in need and facilitating appropriate treatment and support.

References

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders*, *96*(1-2), 89-93.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Washington, DC: Publisher.

Apfel, J. L. (2004). Depression and its treatments: A college sample. *Journal of College Student Psychotherapy*, *18*(2), 67-81.

Arbona, C., Burridge, A., & Olvera, N. (2017). The Center for Epidemiological Studies Depression Scale (CES-D): Measurement equivalence across gender groups in Hispanic college students. *Journal of Affective Disorders, 219*, 112-118.

Arnett, J. (2004). *Emerging adulthood: The winding road from the late teens through the twenties*: New York, NY: Oxford University Press.

Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., ... & Hatcher, S. (2010). Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, *8*(4), 348-353.

Barbic, S. P., Leon, A., MacEwan, W. G., & Barbic, D. (2017). Validation of the PHQ-9 as a screen for depression in the emergency department. *Canadian Journal of Emergency Medicine, 19,* S48-S48.

Beard, C., Hsu, K. J., Rifkin, L. S., Busch, A. B., & Björgvinsson, T. (2016). Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, *193*, 267-273.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Benton, S. A., Robertson, J. M., Tseng, W., Newton, F. B., & Benton, S. L. (2003). Changes in counseling center client problems across 13 years. *Professional Psychology: Research and Practice, 34*(1), 66-72.

Blazer, D. G., Kessler, R. C., McGonagle, K. A., & Swartz, M. S. (1994). The prevalence and distribution of major depression in a national community sample: The national comorbidity survey. *American Journal of Psychiatry, 151*(7), 979-986.

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.

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *British Journal of General Practice, 58*(546), 32-36.

Carey, M., Boyes, A., Noble, N., Waller, A., & Inder, K. (2016). Validation of the PHQ-2 against the PHQ-9 for detecting depression in a large sample of Australian general practice patients. *Australian Journal of Primary Health*, *22*(3), 262-266.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160.

Carmody, D. P. (2005). Psychometric characteristics of the beck depression inventory-II with college students of diverse ethnicity. *International Journal of Psychiatry in Clinical Practice, 9*(1), 22-28.

Centers for Disease Control and Prevention, (CDC). (2010). Current depression among adults - United States, 2006 and 2008. *Morbidity and Mortality Weekly Report, 59*(38), 1229-1235.

Ceyhan, A. A., Kurtyilmaz, Y., & Ceyhan, E. (2009). Investigation of university student's depression. *Egitim Arastirmalari, (36),* 75-90.

Chilcot, J., Hudson, J. L., Moss-Morris, R., Carroll, A., Game, D., Simpson, A., & Hotopf, M. (2018). Screening for psychological distress using the Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS): Initial validation of structural validity in dialysis patients. *General Hospital Psychiatry, 50*, 15-19.

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., ... & Hotopf, M. (2013). The factor structure of the PHQ-9 in palliative care. *Journal of Psychosomatic Research*, *75*(1), 60-64.

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Conti, J. R., Adams, S. K., & Kisler, T. S. (2014). A pilot examination of self-esteem, depression, and sleep in college women. *NASPA Journal About Women in Higher Education*, *7*(1), 47-72.

Contreras, S., Fernandez, S., Malcarne, V. L., Ingram, R. E., & Vaccarino, V. R. (2004). Reliability and validity of the Beck Depression and Anxiety Inventories in Caucasian Americans and Latinos. *Hispanic Journal of Behavioral Sciences, 26*(4), 446-462.

Cox, B. J., Enns, M. W., Borger, S. C., & Parker, J. D. A. (1999). The nature of the depressive experience in analogue and clinically depressed samples. *Behaviour Research and Therapy, 37*(1), 15-24.

Cox, B. J., Enns, M. W., & Larsen, D. K. (2001). The continuity of depression symptoms: Use of cluster analysis for profile identification in patient and student sample. *Journal of Affective Disorders, 65*(1), 67-73.

Cronkite, R. C., Moos, R. H., Twohey, J., Cohen, C., & Swindle Jr, R. (1998). Life circumstances and personal resources as predictors of the ten-year course of depression. *American Journal of Community Psychology*, *26*(2), 255-280.

Cuijpers, P., Geraedts, A. S., van Oppen, P., Andersson, G., Markowitz, J. C., & van Straten, A. (2011). Interpersonal psychotherapy for depression: A meta-analysis. *American Journal of Psychiatry*, *168*(6), 581-592.

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*, *15*(3), 245-258.

Cukrowicz, K. C., & Joiner, T. E. (2007). Computer-based intervention for anxious and depressive symptoms in a non-clinical population. *Cognitive Therapy and Research, 31*(5), 677-693.

Dbouk, N., Arguedas, M., & Sheikh, A. (2008). Assessment of the PHQ-9 as a screening tool for depression in patients with chronic hepatitis C. *Digestive Diseases and Sciences, 53*(4), 1100-1106.

Delgadillo, J., Payne, S., Gilbody, S., Godfrey, C., Gore, S., Jessop, D., & Dale, V. (2011). How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. *Journal of Affective Disorders*, *134*(1-3), 266-271.

Delgado, P. L. (2000). Depression: The case for a monoamine deficiency. *The Journal of Clinical Psychiatry, 61* (Suppl. 6), 7-11.

Denovan, A., Dagnall, N., Dhingra, K., & Grogan, S. (2019). Evaluating the perceived stress scale among UK university students: Implications for stress measurement and management. *Studies in Higher Education, 44*(1), 120-133.

Doris, A., Ebmeier, K., & Shajahan, P. (1999). Depressive illness. *The Lancet*, *354* (9187), 1369-1375.

Dozois, D. J., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, *10*(2), 83-89.

Dum, M., Pickren, J., Sobell, L. C., & Sobell, M. B. (2008). Comparing the BDI-II and the PHQ-9 with outpatient substance abusers. *Addictive Behaviors*, *33*(2), 381-387.

Dyson, R., & Renk, K. (2006). Freshmen adaptation to university life: Depressive symptoms, stress, and coping. *Journal of Clinical Psychology*, *62*(10), 1231-1244.

Eisenberg, D., Gollust, S. E., Golberstein, E., & Hefner, J. L. (2007). Prevalence and correlates of depression, anxiety, and suicidality among university students. *American Journal of Orthopsychiatry, 77*(4), 534-542.

Eisenberg, D., Hunt, J., Speer, N., & Zivin, K. (2011). Mental health service utilization among college students in the United States. *The Journal of Nervous and Mental Disease*, *199*(5), 301-308.

Elderon, L., Smolderen, K. G. E., Na, B., & Whooley, M. A. (2011). Accuracy and prognostic value of American Heart Association recommended depression screening in patients with coronary heart disease: Data from the heart and soul study. *Circulation-Cardiovascular Quality and Outcomes, 4*(5), 533-540.

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., ... & Calabrese, J. R. (2012). The factor structure of major depression symptoms: A test of

four competing models using the Patient Health Questionnaire-9. *Psychiatry Research*, *199*(3), 169-173.

Erford, B. T., Johnson, E., & Bardoshi, G. (2016). Meta-analysis of the English version of the Beck Depression Inventory–Second Edition. *Measurement and Evaluation in Counseling and Development*, *49*(1), 3-33.

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry, 126*(3), 237-240.

Ezquiaga, E., Garcia, A., Bravo, F., & Pallares, T. (1998). Factors associated with outcome in major depression: A 6-month prospective study. *Social Psychiatry and Psychiatric Epidemiology*, *33*(11), 552-557.

Fann, J. R., Bombardier, C. H., Dikmen, S., Esselman, P., Warms, C. A., Pelzer, E., . . . Temkin, N. (2005). Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *The Journal of Head Trauma Rehabilitation, 20*(6), 501-511.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., ... & Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010. *PLoS Medicine*, *10*(11), e1001547.

Fine, T. H., Contractor, A. A., Tamburrino, M., Elhai, J. D., Prescott, M. R., Cohen, G. H., ... & Liberzon, I. (2013). Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I major depression module. *Journal of Affective Disorders*, *150*(3), 1001-1007.

Fletcher, J. M. (2008). Adolescent depression: diagnosis, treatment, and educational attainment. *Health Economics*, *17*(11), 1215-1235.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491.

Ford, J. A., & Schroeder, R. D. (2008). Academic strain and non-medical use of prescription stimulants among college students. *Deviant Behavior, 30*(1), 26-53.

Friedman, E. S., Anderson, I. M., Arnone, D., & Denko, T. (2014). *Handbook of depression*. Tarporley: Springer Healthcare Limited.

Furr, S. R., Westefeld, J. S., McConnell, G. N., & Jenkins, J. M. (2001). Suicide and depression among college students: A decade later. *Professional Psychology: Research and Practice*, *32*(1), 97-100.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*(1), 37-60.

Garlow, S. J., Rosenberg, J., Moore, J. D., Haas, A. P., Koestner, B., Hendin, H., & Nemeroff, C. B. (2008). Depression, desperation, and suicidal ideation in college students: Results from the American Foundation for Suicide Prevention College Screening Project at Emory University. *Depression and Anxiety*, *25*(6), 482-488.

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, *22*(11), 1596-1602.

Gjerdingen, D., Crow, S., McGovern, P., Miner, M., & Center, B. (2009). Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *The Annals of Family Medicine*, *7*(1), 63-70.

Gloria, A. M., Castellanos, J., Kanagui-Muñoz, M., & Rico, M. A. (2012). Assessing Latina/o undergraduates' depressive symptomatology: Comparisons of the Beck Depression Inventory-II, the Center for Epidemiological Studies-Depression Scale, and the Self-Report Depression Scale. *Hispanic Journal of Behavioral Sciences*, *34*(1), 160-181.

Haddad, M., Walters, P., Phillips, R., Tsakok, J., Williams, P., Mann, A., & Tylee, A. (2013). Detecting depression in patients with coronary heart disease: A diagnostic evaluation of the PHQ-9 and HADS-D in primary care, findings from the UPBEAT-UK study. *PloS One*, *8*(10), e78493.

Habibi, M., Khawaja, N. G., Moradi, S., Dehghani, M., & Fadaei, Z. (2014). University Student Depression Inventory: Measurement model and psychometric properties. *Australian Journal of Psychology*, *66*(3), 149-157.

Harber, K. D., Zimbardo, P. G., & Boyd, J. N. (2003). Participant self-selection biases as a function of individual differences in time perspective. *Basic and Applied Social Psychology*, *25*(3), 255-264.

Harvey, S. B., Glozier, N., Henderson, M., Allaway, S., Litchfield, P., Holland-Elliott, K., & Hotopf, M. (2011). Depression and work performance: An ecological study using web-based screening. *Occupational Medicine*, *61*(3), 209-211.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247.

Henderson, C., Evans-Lacko, S., Flach, C., & Thornicroft, G. (2012). Responses to mental health stigma questions: the importance of social desirability and data collection method. *The Canadian Journal of Psychiatry*, *57*(3), 152-160.

Hepner, K. A., Hunter, S. B., Edelen, M. O., Zhou, A. J., & Watkins, K. (2009). A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. *Journal of Substance Abuse Treatment*, *37*(3), 318-325.

Herniman, S. E., Allott, K. A., Killackey, E., Hester, R., & Cotton, S. M. (2017). The psychometric validity of the Center for Epidemiological Studies–Depression scale (CES-D) in first episode schizophrenia spectrum. *Psychiatry Research*, *252*, 16-22.

Herr, N. R., Williams, J. W., Benjamin, S., & McDuffie, J. (2014). Does this patient have generalized anxiety or panic disorder?: The Rational Clinical Examination systematic review. *JAMA*, *312*(1), 78-84.

Hoseinzadeh, F., Abadi, P. H., Agheltar, M., Aghayinejad, A., Torabian, F., Rezayat, A. A., ...& Rahimi, H. R. (2016). The role of immune system in depression disorder. *Health, 8*(15), 1726-1743.

Hu, L., & 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.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219-230.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. *APA Handbook of Testing and Assessment in Psychology*, *1*, 3-19.

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo, and A. M. Hubley, (Eds.), *Understanding and investigating response processes in validation research* (pp. 1-12). Cham, Switzerland: Springer.

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 and E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193-213). New York, NY: Springer.

Hysenbegasi, A., Hass, S. L., & Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *Journal of Mental Health Policy and Economics*, *8*(3), 145.

Ibrahim, A. K., Kelly, S. J., Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of Psychiatric Research*, *47*(3), 391-400.

Irwin, M., Artin, K. H., & Oxman, M. N. (1999). Screening for depression in the older adult: Criterion validity of the 10-item Center for Epidemiological Studies Depression Scale (CES-D). *Archives of Internal Medicine*, *159*(15), 1701-1704.

Kalpakjian, C. Z., Toussaint, L. L., Albright, K. J., Bombardier, C. H., Krause, J. K., & Tate, D. G. (2009). Patient Health Questionnaire-9 in spinal cord injury: An examination of factor structure as related to gender. *The Journal of Spinal Cord Medicine*, *32*(2), 147-156.

Kendler, K. S., Thornton, L. M., & Gardner, C. O. (2001). Genetic risk, number of previous depressive episodes, and stressful life events in predicting onset of major depression. *American Journal of Psychiatry*, *158*(4), 582-586.

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychological Assessment, 30*(8), 1096-1106.

Keyes, C. L., Dhingra, S. S., & Simoes, E. J. (2010). Change in level of positive mental health as a predictor of future risk of mental illness. *American Journal of Public Health*, *100*(12), 2366-2371.

Khawaja, N. G., & Bryden, K. J. (2006). The development and psychometric investigation of the University Student Depression Inventory. *Journal of Affective Disorders*, *96*(1-2), 21-29.

Kiely, K. M., & Butterworth, P. (2015). Validation of four measures of mental health against depression and generalized anxiety in a community-based sample. *Psychiatry Research*, *225*(3), 291-298.

Kitamura, T., Hirano, H., Chen, Z., & Hirata, M. (2004). Factor structure of the Zung Self-rating Depression Scale in first-year university students in Japan. *Psychiatry Research*, *128*(3), 281-287.

Kline, R. (2013). Exploratory and confirmatory factor analysis. InY Petscher, C Schatschneider, and D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 183-217). New York, NY: Routledge.

Kneipp, S., Kairalla, J., Stacciarini, J. M., Pereira, D., & Miller, M. D. (2010). Comparison of depressive symptom severity scores in low-income women. *Nursing Research, 59*(6), 380-388.

Krause, J. S., Reed, K. S., & McArdle, J. J. (2010). Factor structure and predictive validity of somatic and nonsomatic symptoms from the Patient Health Questionnaire-9: A

longitudinal study after spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, *91*(8), 1218-1224.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

Kung, S., Alarcon, R. D., Williams, M. D., Poppe, K. A., Moore, M. J., & Frye, M. A. (2013). Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an integrated mood disorders practice. *Journal of Affective Disorders*, *145*(3), 341-343.

Leigh, I. W., & Anthony-Tolbert, S. (2001). Reliability of the BDI-II with deaf persons. *Rehabilitation Psychology, 46*(2), 195-202.

Lerner, D., Adler, D. A., Chang, H., Lapitsky, L., Hood, M. Y., Perissinotto, C., . . . Rogers, W. H. (2004). Unemployment, job retention, and productivity loss among employees with depression. *Psychiatric Services, 55*(12), 1371-1378.

Levis, B., Benedetti, A., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., ... & Gilbody, S. (2018). Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *The British Journal of Psychiatry*, *212*(6), 377-385.

Lewinsohn, P. M., Solomon, A., Seeley, J. R., & Zeiss, A. (2000). Clinical implications of “subthreshold" depressive symptoms. *Journal of Abnormal Psychology*, *109*(2), 345-351.

Lipps, G. E., Lowe, G. A., & Young, R. (2007). Validation of the Beck Depression Inventory-II in a Jamaican university student cohort. *West Indian Medical Journal*, *56*(5), 404-408.

Liu, Y., & Wang, J. (2015). Validity of the Patient Health Questionnaire-9 for DSM-IV major depressive disorder in a sample of Canadian working population. *Journal of Affective Disorder*, *187*, 122-126.

Lopez Ibor, J. J., Frances, A., & Jones, C. (1994). Dysthymic disorder: A comparison of DSM-IV and ICD-10 and issues in differential diagnosis. *Acta Psychiatrica Scandinavica*, *89*, 12-18.

MacGeorge, E. L., Samter, W., & Gillihan, S. J. (2005). Academic stress, supportive communication, and health. *Communication Education*, *54*(4), 365-372.

Makhubela, M. S. (2016). Measurement invariance of the Beck Depression Inventory across race with South African university students. *South African Journal of Psychology*, *46*(4), 449-461.

Manea, L., Gilbody, S., & McMillan, D. (2015). A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *General Hospital Psychiatry*, *37*(1), 67-75.

Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the brief Patient Health Questionnaire mood scale (PHQ-9) in the general population. *General Hospital Psychiatry*, *28*(1), 71-77.

Matud, M. P. (2004). Gender differences in stress and coping styles. *Personality and Individual Differences*, *37*(7), 1401-1415.

Meadows, S. O., Brown, J. S., & Elder, G. H. (2006). Depressive symptoms, stress, and support: Gendered trajectories from adolescence to young adulthood. *Journal of Youth and Adolescence*, *35*(1), 89-99.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Meyers, B. S., Sirey, J. A., Bruce, M., Hamilton, M., Raue, P., Friedman, S. J., ... & Alexopoulos, G. (2002). Predictors of early recovery from major depression among persons admitted to community-based clinics: An observational study. *Archives of General Psychiatry*, *59*(8), 729-735.

Milette, K., Hudson, M., Baron, M., Thombs, B. D., & Canadian Scleroderma Research Group*. (2010). Comparison of the PHQ-9 and CES-D depression scales in systemic sclerosis: Internal consistency reliability, convergent validity and clinical correlates. *Rheumatology*, *49*(4), 789-796.

Monroe, S. M., & Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders. *Psychological Bulletin*, *110*(3), 406-425.

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B. (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *The Lancet*, *370*(9590), 851-858.

Mowbray, C. T., Megivern, D., Mandiberg, J. M., Strauss, S., Stein, C. H., Collins, K., ... & Lett, R. (2006). Campus mental health services: recommendations for change. *American Journal of Orthopsychiatry*, *76*(2), 226-237.

Meyers, B. S., Sirey, J. A., Bruce, M., Hamilton, M., Raue, P., Friedman, S. J., ... & Alexopoulos, G. (2002). Predictors of early recovery from major depression among persons admitted to community-based clinics: an observational study. *Archives of General Psychiatry*, *59*(8), 729-735.

Moriarty, A. S., Gilbody, S., McMillan, D., & Manea, L. (2015). Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): A meta-analysis. *General Hospital Psychiatry*, *37*(6), 567-576.

National Institute for Clinical Excellence (NICE). (2009). Depression: the treatment and management of depression in adults. *London: National Institute for Health and Clinical Excellence*.

Nihalani, N., Simionescu, M., & Dunlop, B. W. (2006). Depression: Phenomenology, epidemiology, and pathophysiology. In T. L. Schwartz and T. J. Peterson (Eds.), *Depression: Treatment strategies and management* (pp. 13-33). New York: Taylor & Francis.

Nolen-Hoeksema, S., & Girgus, J. S. (1994). The emergence of gender differences in depression during adolescence. *Psychological Bulletin*, *115*(3), 424-443.

Olfson, M., Marcus, S. C., Druss, B., Elinson, L., Tanielian, T., & Pincus, H. A. (2002). National trends in the outpatient treatment of depression. *JAMA*, *287*(2), 203-209.

Orme, J. G., Reis, J., & Herz, E. J. (1986). Factorial and discriminant validity of the Center for Epidemiological Studies Depression (CES-D) scale. *Journal of Clinical Psychology*, *42*(1), 28-33.

Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review, 12*(2), 131-146.

Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008). Psychometric properties of the Beck Depression Inventory-II in nonclinical adolescent samples. *Journal of Clinical Psychology, 64*(1), 83-102.

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Gutierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory-II. *Journal of Psychopathology and Behavioral Assessment, 19*(4), 359-376.

Osman, A., Kopper, B. A., Barrios, F., Gutierrez, P. M., & Bagge, C. L. (2004). Reliability and validity of the Beck Depression Inventory-II with adolescent psychiatric inpatients. *Psychological Assessment, 16*(2), 120-132.

Owens, J., & Adolescent Sleep Working Group. (2014). Insufficient sleep in adolescents and young adults: An update on causes and consequences. *Pediatrics*, *134*(3), e921.

Palmer, E. J., & Binks, C. (2008). Psychometric properties of the Beck Depression Inventory-II with incarcerated male offenders aged 18–21 years. *Criminal Behaviour and Mental Health, 18*(4), 232-242.

Phelan, E., Williams, B., Meeker, K., Bonn, K., Frederick, J., LoGerfo, J., & Snowden, M. (2010). A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Family Practice*, *11*(1), 63.

Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *Journal of Psychiatric Research*, *56*, 112-119.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90.

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.

Randall, J., Voth, R., Burnett, E., Bazhenova, L., & Bardwell, W. (2013). Clinic-based depression screening in lung cancer patients using the PHQ-2 and PHQ-9 depression questionnaires: A pilot study. *Supportive Care in Cancer, 21*(5), 1503-1507.

Rathore, J. S., Jehi, L. E., Fan, Y., Patel, S. I., Foldvary-Schaefer, N., Ramirez, M. J., ... & Tesar, G. E. (2014). Validation of the Patient Health Questionnaire-9 (PHQ-9) for depression screening in adults with epilepsy. *Epilepsy & Behavior*, *37*, 215-220.

Rawson, H. E., Bloomer, K., & Kendall, A. (1994). Stress, anxiety, depression, and physical illness in college students. *The Journal of Genetic Psychology*, *155*(3), 321-330.

Richardson, E. J., & Richards, J. S. (2008). Factor structure of the PHQ-9 screen for depression across time since injury among persons with spinal cord injury. *Rehabilitation Psychology, 53*(2), 243-249. doi:10.1037/0090-5550.53.2.243

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

Roberts, S. J., Glod, C. A., Kim, R., & Hounchell, J. (2010). Relationships between aggression, depression, and alcohol, tobacco: Implications for healthcare providers in student health. *Journal of the American Academy of Nurse Practitioners*, *22*(7), 369-375.

Romaniuk, M., & Khawaja, N. G. (2013). University Student Depression Inventory (USDI): Confirmatory factor analysis and review of psychometric properties. *Journal of Affective Disorders*, *150*(3), 766-775.

Rooney, A. G., McNamara, S., Mackinnon, M., Fraser, M., Rampling, R., Carson, A., & Grant, R. (2012). Screening for major depressive disorder in adults with cerebral glioma: An initial validation of 3 self-report instruments. *Neuro-Oncology*, *15*(1), 122-129.

Ross, S. E., Niebling, B. C., & Heckert, T. M. (1999). Sources of stress among college students. *Social Psychology*, *61*(5), 841-846.

Rupp, A., Koh, K. & Zumbo, B.D. (2003, April). *What is the impact on exploratory factor analysis results of a polychoric correlation matrix from LISREL/PRELIS and EQS when some respondents are not able to follow the rating scale*. Paper presented at the annual meeting of the American Educational Research Association (AERA) in Chicago, Illinois.

Ruiz, M. A., Zamorano, E., García-Campayo, J., Pardo, A., Freire, O., & Rejas, J. (2011). Validity of the GAD-7 scale as an outcome measure of disability in patients with generalized anxiety disorders in primary care. *Journal of Affective Disorders*, *128*(3), 277-286.

Ryan, T. A., Bailey, A., Fearon, P., & King, J. (2013). Factorial invariance of the patient health questionnaire and generalized anxiety disorder questionnaire. *British Journal of Clinical Psychology, 52*(4), 438-449.

Salokangas, R. K., From, T., Luutonen, S., & Hietala, J. (2018). Adverse childhood experiences lead to perceived negative attitude of others and the effect of adverse childhood experiences on depression in adulthood is mediated via negative attitude of others. *European Psychiatry*, *54*, 27-34.

Saunders, D. E., Peterson, G. W., Sampson Jr, J. P., & Reardon, R. C. (2000). Relation of depression and dysfunctional career thinking to career indecision. *Journal of Vocational Behavior*, *56*(2), 288-298.

Schotte, C. K. W., Van Den Bossche, B., De Doncker, D., Claes, S., & Cosyns, P. (2006). A biopsychosocial model as a guide for psychoeducation and treatment of depression. *Depression and Anxiety, 23*(5), 312-324.

Schueller, S. M., Kwasny, M. J., Dear, B. F., Titov, N., & Mohr, D. C. (2015). Cut points on the Patient Health Questionnaire (PHQ-9) that predict response to cognitive–behavioral treatments for depression. *General Hospital Psychiatry*, *37*(5), 470-475.

Schwenk, T. L., Davis, L., & Wimsatt, L. A. (2010). Depression, stigma, and suicidal ideation in medical students. *JAMA*, *304*(11), 1181-1190.

Segal, D. L., Coolidge, F. L., Cahill, B. S., & O'riley, A. A. (2008). Psychometric properties of the Beck Depression Inventory-II (BDI-II) among community-dwelling older adults. *Behavior Modification, 32*(1), 3-20.

Shean, G., & Baldwin, G. (2008). Sensitivity and specificity of depression questionnaires in a college-age sample. *The Journal of Genetic Psychology*, *169*(3), 281-292.

Shepardson, R. L., & Funderburk, J. S. (2014). Implementation of universal behavioral health screening in a university health setting. *Journal of Clinical Psychology in Medical Settings*, *21*(3), 253-266.

Sidebottom, A., Harrison, P., Godecker, A., & Kim, H. (2012). Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Archives of Women's Mental Health, 15*(5), 367-374.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1-3), 83-117.

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

Sjonnesen, K., Berzins, S., Fiest, K. M., M Bulloch, A. G., Metz, L. M., Thombs, B. D., & Patten, S. B. (2012). Evaluation of the 9-item Patient Health Questionnaire (PHQ-9) as an assessment instrument for symptoms of depression in patients with multiple sclerosis. *Postgraduate Medicine, 124*(5), 69-77.

Smith, G. (2008). Does gender influence online survey participation?: A record-linkage analysis of university faculty online survey response behavior. *ERIC Document Reproduction Service No. ED 501717*.

Smith, T. B., Rosenstein, I., & Granaas, M. M. (2001). Intake screening with the Self-Rating Depression Scale in a university counseling center. *Journal of College Counseling*, *4*(2), 133-141.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA*, *282*(18), 1737-1744.

Sprinkle, S. D., Lurie, D., Insko, S. L., Atkinson, G., Jones, G. L., Logan, A. R., & Bissada, N. N. (2002). Criterion validity, severity cut scores, and test-retest reliability of the Beck Depression Inventory-II in a university counseling center sample. *Journal of Counseling Psychology*, *49*(3), 381.

Storch, E. A., Roberti, J. W., & Roth, D. A. (2004). Factor structure, concurrent validity, and internal consistency of the Beck Depression Inventory-Second Edition in a sample of college students. *Depression and Anxiety*, *19*(3), 187-189.

Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis (paper 203-30). In *Proceedings of the Thirtieth Annual SAS® Users Group International Conference* (Vol. 203, p. 30).

Targum, S. D. (2011). The distinction between clinical and research interviews in psychiatry. *Innovations in Clinical Neuroscience*, *8*(3), 40.

Thekkumpurath, P., Walker, J., Butcher, I., Hodges, L., Kleiboer, A. M., O'Connor, M., . . . Sharpe, M. (2011). Screening for major depression in cancer outpatients: The diagnostic accuracy of the 9-item Patient Health Questionnaire. *Cancer, 117*(1), 218-227.

Thomas, C. M., & Morris, S. (2003). Cost of depression among adults in England in 2000. *The British Journal of Psychiatry*, *183*(6), 514-519.

Thombs, B., Ziegelstein, R., & Whooley, M. (2008). Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: Data from the heart and soul study. *Journal of General Internal Medicine, 23*(12), 2014-2017.

Thompson, A. W., Liu, H., Hays, R. D., Katon, W. J., Rausch, R., Diaz, N., ... & Vickrey, B. G. (2011). Diagnostic accuracy and agreement across three depression assessment measures for Parkinson's disease. *Parkinsonism & Related Disorders*, *17*(1), 40-45.

Tilden, V. P., Nelson, C. A., & May, B. A. (1990). Use of qualitative methods to enhance content validity. *Nursing Research*, *39*(2), 172-175.

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy, 40*(2), 126-136.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*, 769.

Turner, A., Hambridge, J., White, J., Carter, G., Clover, K., Nelson, L., & Hackett, M. (2012). Depression screening in stroke: A comparison of alternative measures with the structured

diagnostic interview for the Diagnostic and Statistical Manual of Mental Disorders , Fourth Edition (major depressive episode) as criterion standard. *Stroke, 43*(4), 1000-1005.

Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort difference on the children's depression inventory: A meta-analysis. *Journal of Abnormal Psychology*, *111*(4), 578-588.

Tyrer, P. (2014). A comparison of DSM and ICD classifications of mental disorder. *Advances in Psychiatric Treatment*, *20*(4), 280-285.

Uher, R., Payne, J. L., Pavlova, B., & Perlis, R. H. (2014). Major depressive disorder in DSM-5: Implications for clinical practice and research of changes from DSM-IV. *Depression and anxiety*, *31*(6), 459-471.

Umegaki, Y., & Todo, N. (2016). Psychometric properties of the Japanese BDI, CES-D, PHQ-9, and SDS depression scales in university students: Analysis based on classical test theory. *International Journal of Psychology*, *51*, 947-948.

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): A systematic review with meta-analysis. *Plos One, 11*(5), e0155431.

Volker, D., Zijlstra-Vlasveld, M. C., Brouwers, E. P. M., Homans, W. A., Emons, W. H. M., & van der Feltz-Cornelis, C M. (2016). Validation of the Patient Health Questionnaire-9 for major depressive disorder in the occupational health setting. *Journal of Occupational Rehabilitation, 26*(2), 237-244.

Vredenburg, K., O'brien, E., & Krames, L. (1988). Depression in college students: Personality and experiential factors. *Journal of Counseling Psychology*, *35*(4), 419-425.

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-month use of mental health services in the United States: Results from the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 629-640.

Wang, Y. P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Revista Brasileira de Psiquiatria*, *35*(4), 416-431.

Ware, John & A. Kosinski, M & D. Keller, S. (1995). *SF-12: How to score the SF-12 Physical and Mental Health Summary Scales*. Boston, MA: The Health Institute, New England Medical Center.

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.

Wei, M., Russell, D. W., & Zakalik, R. A. (2005). Adult attachment, social self-efficacy, self-disclosure, loneliness, and subsequent depression for freshman college students: A longitudinal study. *Journal of Counseling Psychology, 52*(4), 602-614.

Werner, F. M., & Covenas, R. (2010). Classical neurotransmitters and neuropeptides involved in major depression: A review. *International Journal of Neuroscience*, *120*(7), 455-470.

Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory-Second Edition (BDI-II) in a student sample. *Journal of Clinical Psychology, 56*(4), 545-551.

Whitton, S. W., & Whisman, M. A. (2010). Relationship satisfaction instability and depression. *Journal of Family Psychology*, *24*(6), 791-794.

Wiebe, J. S., & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory-II in English and Spanish. *Psychological Assessment, 17*(4), 481-485.

Willis, G. B. (1999). *Cognitive interviewing: A "how to" guide.* Research Triangle Park, NC: Research Triangle Institute.

Wittkampf, K. A., Naeije, L., Schene, A. H., Huyser, J., & van Weert, H. C. (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. *General Hospital Psychiatry*, *29*(5), 388-395.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392.

Zhang, Y. L., Liang, W., Chen, Z. M., Zhang, H. M., Zhang, J. H., Weng, X. Q., ... & Zhang, Y. L. (2013). Validity and reliability of Patient Health Q uestionnaire-9 and Patient Health Questionnaire-2 to screen for depression among college students in China. *Asia-Pacific Psychiatry*, *5*(4), 268-275.

Zhang, Y., Peters, A., & Chen, G. (2018). Perceived stress mediates the associations between sleep quality and symptoms of anxiety and depression among college nursing students. *International Journal of Nursing Education Scholarship*, *15*(1), 1-9.

Zullig, K. J., & Divin, A. L. (2012). The association between non-medical prescription drug use, depressive symptoms, and suicidality among college students. *Addictive Behaviors*, *37*(8), 890-899.

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.