

Overview Review

Cite this article: Yang M, Seon Q, Gomez Cardona L, Karia M, Velupillai G, Noel V and Linnaranta O (2023). Safe and valid? A systematic review of the psychometric properties of culturally adapted depression scales for use among Indigenous populations. *Cambridge Prisms: Global Mental Health*, **10**, e60, 1–14 <https://doi.org/10.1017/gmh.2023.52>

Received: 13 March 2023

Revised: 26 July 2023

Accepted: 02 September 2023

Keywords:

psychometrics; Indigenous; validity; reliability; clinical utility

Corresponding author:

Outi Linnaranta;

Email: outi.linnaranta@thl.fi

Safe and valid? A systematic review of the psychometric properties of culturally adapted depression scales for use among Indigenous populations

Michelle Yang¹, Quinta Seon^{2,3}, Liliana Gomez Cardona^{2,3}, Maharshee Karia², Gajanan Velupillai³, Valérie Noel^{2,4} and Outi Linnaranta^{2,3,5} 

¹École interdisciplinaire des sciences de la santé/Interdisciplinary School of Health Sciences, Université d'Ottawa/University of Ottawa, Ottawa, ON, Canada; ²Department of Psychiatry, Douglas Mental Health University Institute, Montreal, QC, Canada; ³Department of Psychiatry, McGill University, Montreal, QC, Canada; ⁴ACCESS Open Minds, Centre de recherche Douglas/Perry 3, Montreal, QC, Canada and ⁵Department of Equality, Finnish Institute for Health and Welfare, Helsinki, Finland

Abstract

Background: Implementing culturally sensitive psychometric measures of depression may be an effective strategy to improve acceptance, response rate, and reliability of psychological assessment among Indigenous populations. However, the psychometric properties of depression scales after cultural adaptation remain unclear.

Methods: We screened the Ovid Medline, PubMed, Embase, Global Health, PsycInfo, and CINAHL databases through three levels of search terms: Depression, Psychometrics, and Indigenous, following the PRISMA guidelines. We assessed metrics for reliability (including Cronbach's alpha), validity (including fit indices), and clinical utility (including predictive value).

Results: Across 31 studies included in the review, 13 different depression scales were adapted through language or content modification. Sample populations included Indigenous from the Americas, Asia, Africa, and Oceania. Most cultural adaptations had strong psychometric properties; however, few and inconsistent properties were reported. Where available, alphas, inter-rater and test-retest reliability, construct validity, and incremental validity often indicated increased cultural sensitivity of adapted scales. There were mixed results for clinical utility, criterion validity, cross-cultural validity, sensitivity, specificity, area under the receiver operating characteristic curve, predictive value, and likelihood ratio.

Conclusions: Modifications to increase cultural relevance have the potential to improve fit and acceptance of a scale by the Indigenous population, however, these changes may decrease specificity and negative predictive value. There is an urgent need for suitable tools that are useful and reliable for identifying Indigenous individuals for clinical treatment of depression. This awaits future work for optimal specificity and validated cut-off points that take into account the high prevalence of depression in these populations.

Impact statement

The present study suggests that modifying depression scales to fit the Indigenous context through changes to language or question structure is a culturally sensitive strategy that increases acceptance of psychological evaluation and treatment in communities. However, increasing acceptability must be balanced with maintaining clinical utility of instruments. The high prevalence of depression in these populations must be taken into account when developing culturally sensitive but specific tools.

Introduction

Cultural safety is a combination of cultural awareness (acknowledging the differences between cultures), cultural sensitivity (respecting other cultures), and cultural competency (effectively working with diverse populations through appropriate behaviors, attitudes, measures, and policies; Marsella et al., 1985; Simon and Catherine, 2009). Research has shown that culturally competent care can improve communication between minority groups and health care professionals, including recognition of mental illness and assessment of its severity among minority groups (Schouten and Meeuwesen, 2006). These considerations are especially important in mental health care for Indigenous peoples, as Indigenous peoples share cultural backgrounds and ideas about health and concern that are unique from other ethnic groups (Mayberry et al.,

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



2000). As a response to these unique perspectives, cultural adaptation of psychometric tools has become a commonly used method to increase cultural safety of psychometric evaluation and to reduce communication problems (Gomez Cardona et al., *Submitted*). As a consequence, it is hoped to lead to reduced risk of harm, and to improved access of Indigenous peoples to psychiatric treatment for depression. Screening for symptoms with appropriate psychometric scales in community services is an approach to improve the quality of treatment for some Indigenous populations (Esler et al., 2007).

Measurement-based psychiatric care uses standardized measures to guide treatment and subsequently evaluate treatment outcomes (Aboraya et al., 2018). Validated psychometric measures can be used in first point of contact settings (i.e., primary care) to screen for psychiatric disorders, assess the need for treatment, monitor symptom severity, and track treatment outcomes (Porcerelli and Jones, 2017). Indigenous communities often lack specialized resources, including staff with specialized psychiatric skills (Boksa et al., 2015). Many studies have reported that there is a dearth of trained mental health workers of local Indigenous origin; moreover, a high turnover of non-Indigenous health workers leads to a lack of continuity of services and a lack of connection to specialized services or excessively long wait lists for Indigenous with severe mental illness (Boksa et al., 2015). This means limited access to an interview-based psychiatric diagnosis and specialist follow-up for relapses after treatment (Boksa et al., 2015). Moreover, trauma-informed care is rare, yet would be needed for safe and valid psychiatric assessment and interventions for Indigenous peoples. This is important as it is known that Indigenous people's health promotion and health seeking behaviors are largely influenced by a colonialistic past which has caused intergenerational social inequities; a lack of trust and confidence in many governments still prevails among Indigenous populations, making them less likely to be screened and causing an overall resistance to engaging with the healthcare system (Leung, 2016). Because compromised quality of mental health care is evident in many Indigenous communities, there is a need for effective symptom screening and monitoring with stable psychometric measures (Chan et al., 2021).

In addition to evaluating the need for treatment at the individual level, having culturally safe and accurate measurement of a population's mental health is essential for appropriate resource allocation and service planning at population and community levels (Chan et al., 2021). The use of psychometric screens at a community level can raise awareness of mental health needs and crises among stakeholders in Indigenous health and provide tools for evaluating the efficacy of interventions (Chapla et al., 2019). To make reliable conclusions on the efficacy of screens at community and population levels, it is important to ensure that psychometric tools are culturally safe and trauma informed, but also clinically useful, reliable when used with a specific population, and have a high quality as compared to traditional gold standards (Chan et al., 2021).

Standard qualities are requested from psychometric screens and outcome measures, and are also essential for culturally adapted depression scales. First, *reliability* represents a test's consistency across test questions (Andrade, 2018). Second, *validity* represents a test's accuracy such that test items are reported to be meaningful and relevant to the population they are used with (Andrade, 2018). Finally, the *clinical utility* of a scale indicates its utility for clinicians to diagnose and determine content of treatment (Labrique and Pan, 2010). In particular, the *sensitivity* of a screen indicates its capacity to correctly identify people who are most likely to benefit from a

clinical diagnostic interview (Parikh et al., 2008). Accordingly, an optimal culturally adapted psychometric scale should not exclude from clinical interviews those who are depressed (*sensitivity*), but should also guide an efficient use of clinician resources for an interview with individuals who require treatment (*specificity*).

The prevalence of depression and anxiety, and incidence of suicide among Indigenous peoples, is commonly high in comparison to Western cohorts (Shen et al., 2018). It is known that a heightened presence of environmental stress and distress disproportionately raises the sensitivity of a diagnostic tool even where psychiatric care is not indicated or appropriate (Simon, 2015). As such, in Indigenous communities, the proportion of positive cases screened by a highly sensitive scale may misleadingly indicate a need to provide a clinical interview for the entire population (Parikh et al., 2008). Therefore, the optimal balance of different psychometric properties must be thoroughly tested in Indigenous communities before they are used to guide treatment.

Recently, more studies have explored different methods of culturally adapting and developing measures that reflect mental health conditions (Haswell et al., 2010). We recently reviewed the methods of cultural adaptation of measures for depression, identifying the modifications and adaptations made and evaluating their acceptability by target Indigenous populations (Gomez Cardona et al., *Submitted*). Here, we continue the work and investigate the psychometric properties of the previously identified culturally adapted psychometric measures. In this review, we assess the reported quality of the psychometric characteristics of adapted depression scales and their utility for psychological evaluation among Indigenous groups.

Methods

The methods for the systematic search, including the search strategy, were reported previously (Gomez Cardona et al., *Submitted*). This study followed the outline set forth by the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA; Figure 1). We searched the Ovid Medline, PubMed, PsycInfo, Embase, CINAHL, and Global Health databases for articles using three levels of search terms related to: a) Depression, b) Psychometrics, and c) Indigenous (*Supplementary Material 1*). After an initial search to capture articles from the inception of the databases to April 2021, we extended the search to the end of August 2022. Any reports with information on psychometric properties reported after extraction of these original studies were added to the current analysis as gray literature.

We extracted the information on utility of depression scales following cultural adaptation. This information included the "gold-standards" they were measured against, and the optimal cut-off point(s) determined after cultural adaptation (Table 1). We also extracted information on psychometric properties of adapted scales (Table 2), which were assessed through a quality criteria checklist (*Supplementary Material 2*). We extracted data on reliability (internal consistency, test-retest reliability, and inter-rater reliability), criterion validity (concurrent and predictive validity), construct validity (convergent and discriminant validity), cross-cultural validity (measurement invariance), incremental validity, and clinical utility (sensitivity, specificity, area under the ROC curve [AUC], positive predictive value [PPV], negative predictive value [NPV], and likelihood ratio [LR]). Detailed results on trends found across unique adaptation processes are presented in *Supplementary Material 3*.

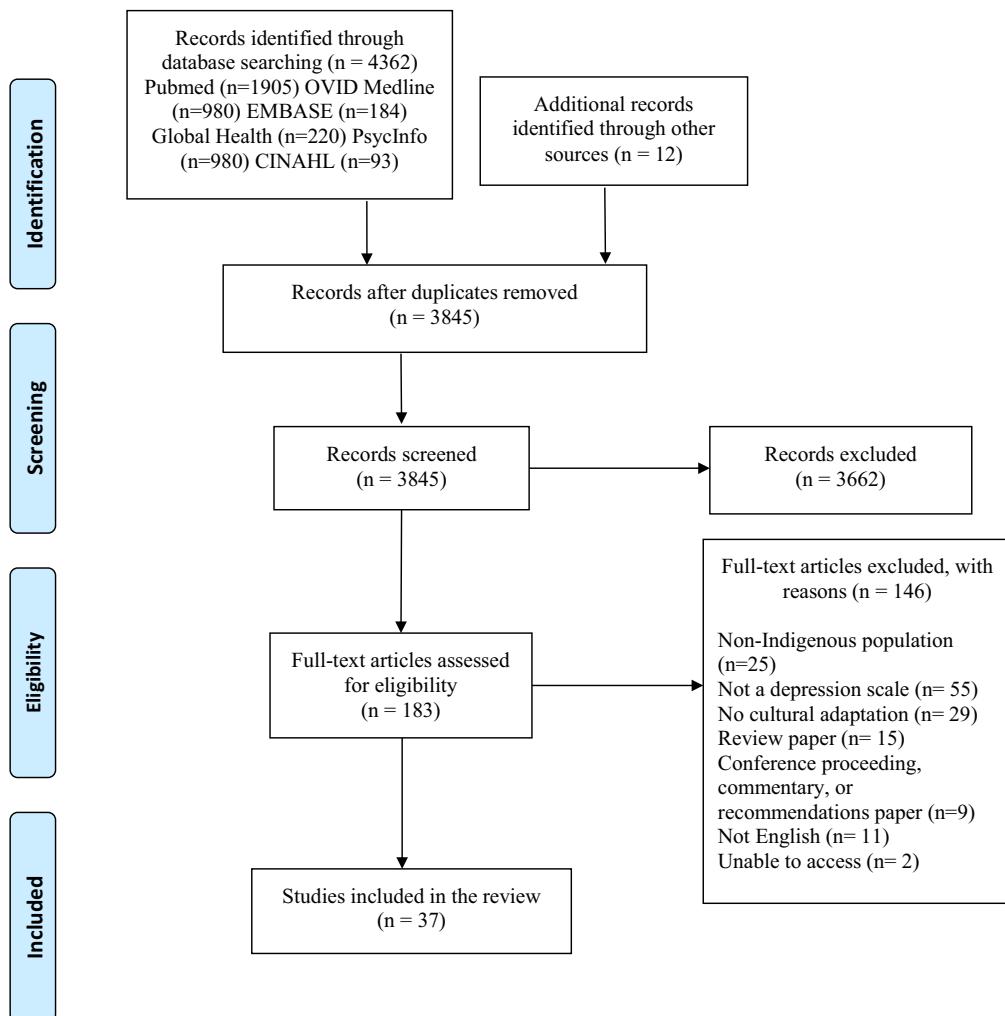


Figure 1. PRISMA flow diagram.

Assessment of quality

We followed the guidelines of the ROBIS tool to assess the quality and risk of bias of this review (Whiting et al., 2016; Supplementary Material 4). Here, we report potential biases across several domains: a) study eligibility criteria, b) identification and selection of studies, c) data collection and study appraisal, and d) synthesis and findings.

Results

Description of the adapted scales

Originally, we had identified 37 studies that met criteria in the systematic search (Gomez Cardona et al., 2021). Thirty-one (83.8%) of these studies reported results on the validation of their scales' psychometric propert(ies). Cohort sizes ranged from $n = 97$ to $n = 4,767$. Target Indigenous peoples were from native to Canada or the United States (5/31), Latin America (3/31), Asia (8/31), Africa (8/31), and Australia or New Zealand (7/31); many populations lived in rural settings. Each of the 31 studies produced a unique culturally adapted scale based on their individual methods. These 31 scales were, or are variations of, the following scales, listed in decreasing number of adaptations per scale: Center for

Epidemiologic Studies Depression Scale (CES-D), $n = 8$; Patient Health Questionnaire (PHQ-9), $n = 7$; Edinburgh Postnatal Depression Scale (EPDS), $n = 4$; Kessler Psychological Distress Scale, $n = 4$; Hopkins Symptom Checklist Depression Scale (HSCL), $n = 3$; Geriatric Depression Scale (GDS), $n = 2$; Kimberley Indigenous Cognitive Assessment of Depression (KICA-dep), $n = 1$; Ndetei-Othieno-Kathuku scale (NOK), $n = 1$; International Depression Symptom Scale-General (IDSS-G), $n = 1$; Dar-es-Salaam Symptom Questionnaire (DSQ), $n = 1$; Kimberley Mum's Mood Scale (KMMS), $n = 1$; Self-Reporting Questionnaire (SRQ), $n = 1$; Functioning Assessment Instrument (FAI), $n = 1$.

Reliability

Cronbach's alpha was the most commonly analyzed psychometric property, reported by 77% (24/31) of adapted depression measurement scales (Ganguli et al., 1999; Tiburcio Sainz and Natera Rey, 2007; Bass et al., 2008; Campbell et al., 2008; Esler et al., 2008; Kaaya et al., 2008; Ekeroma et al., 2012; Gelaye et al., 2013; Armenta et al., 2014; Haroz et al., 2014; McNamara et al., 2014; Schneider et al., 2015; Bougie et al., 2016; Baron et al., 2017; Denckla et al., 2017; Haroz et al., 2017; Marley et al., 2017; Schantz et al., 2017; Gallis et al., 2018; Harry and Crea, 2018; Kilburn et al., 2018; Ashaba et al., 2019; Chapla et al., 2019; Hackett et al., 2019).

Table 1. Characteristics of culturally adapted scales

Reference	Original scale	Adapted scale	Measured against ("Gold-standard")	Clinical use	Optimal culturally safe cut-off	Population that optimal cut-off is determined for
Armenta et al., 2014	N/A	CES-D-20	Diagnosis of MDD through the DSM-IV	○ Negative affect and somatic difficulties items are useful to predicting MDD	N/A	○ Young North American Indigenous (early to late adolescence)
Baron et al., 2017	CES-D-20	CES-D-10	Diagnosis of major depression using MINI 6.0 (major depressive episode)	○ Can indicate high risk of depression ○ Optimized cut-off is more acceptable for use in South Africa (compared to the Western cut-off) ○ Length is more feasible to use in research and clinical settings	11–13 (mixed-results)	○ Zulu, colored Afrikaans and Xhosa-speaking populations
Chapla et al., 2019	CES-DC	Gujarati version of CES-DC	N/A	○ Can be used to check prevalence of depression among youth	15	○ School-going adolescents from Gujarati medium schools
Chapleski et al., 1997	N/A	CES-D	N/A	○ Is robust and useful to measuring depression factors (depressive affect, somatic, interpersonal, well-being)	N/A	○ American Indian Elders over 55 years (urban, rural off-reservation, reservation)
Harry and Crea, 2018	CES-D	Modified long and short form CES-D	N/A	○ Has measurement invariance properties and can be used to assess mental health needs of marginalized groups	N/A	○ American Indian young adults
Kilburn et al., 2018	CES-D-20	CES-D-10	N/A	○ Is a low resource screener for adolescent depression ○ Scores are related to individual and household determinants of depression (i.e., increasing age)	N/A	○ Sub-Saharan African youth ○ Low socioeconomic status, rural communities
Tiburcio Sainz and Natera Rey, 2007	N/A	CES-D-19	N/A	○ Are reliable measures of coping, emotional regulation, and negative affect	N/A	○ Indigenous populations in Mexico
Schantz et al., 2017	CES-D-20 PHQ-9	Short form CES-D PHQ-8	N/A	○ Can measure nonspecific psychological distress and symptoms of clinical depression	10 10	○ Indigenous in Andean Latin America ○ Patient population from urban and low-income setting
Caneo et al., 2020	PHQ-9	PHQ-2		○ Comparable to the PHQ-9 in screening for major depressive syndrome	3	○ Agricultural/rural Chilean population
Esler et al., 2008	PHQ-9	Modified PHQ-9	Diagnoses of major depression if meeting all DSM-IV criteria; diagnosis of minor depression if meeting 2–4 of the DSM-IV criteria	○ Is valid and acceptable as a part of clinical assessment for depression ○ Has stronger concordance with DSM-IV diagnosis criteria than other tools in this region	9	○ Australian Aboriginals with ischemic heart disease
Gallis et al., 2018	PHQ-9	Urdu version of the PHQ-9	Diagnosis of depression on SCID	○ Is a valid tool to be used in low-income settings where depressive disorders are prevalent	10	○ Pregnant women in Pakistan ○ Rural community settings
Gelaye et al., 2013	PHQ-9	Amharic version of the PHQ-9	Diagnosis of depression on the SCAN	○ Has high discrimination power and is comparable to diagnoses via clinical interviews	10	○ Ethiopian and sub-Saharan African adults
Hackett et al., 2019	PHQ-9	aPHQ-9	Diagnosis of major depressive episodes on the MINI 6.0.0	○ Has utility in assessment of depression levels, epidemiology studies, treatment monitoring, and outcome assessment	10	○ Australian Aboriginal adults at high risk of depression ○ Rural and urban setting Primary care setting

(Continued)

Table 1. (Continued)

Reference	Original scale	Adapted scale	Measured against ("Gold-standard")	Clinical use	Optimal culturally safe cut-off	Population that optimal cut-off is determined for
Husain et al., 2006	PHQ	Urdu version of PHQ	Diagnosis of depressive disorder on the PAS	<ul style="list-style-type: none"> ○ Is reliable for screening depressive disorders ○ Can detect somatic and psychological problems 	5	<ul style="list-style-type: none"> ○ Rural Pakistani population ○ People with little or no education
	SRQ	Urdu version of SRQ			7	
Bougie et al., 2016	N/A	K-10	N/A	<ul style="list-style-type: none"> ○ Can measure non-specific psychological distress (DSM-IV criteria for anxiety and mood disorders) ○ Can be used as a single factor score in the APS 	10–11 (mixed results)	<ul style="list-style-type: none"> ○ First Nations living off reserve, Métis, and Inuit
McNamara et al., 2014	K-10	K-5	N/A	<ul style="list-style-type: none"> ○ Is a good measure for detecting likelihood of psychological distress ○ Can be used in population health surveys and clinical practice 	N/A	<ul style="list-style-type: none"> ○ Older Aboriginal and Torres Strait Islanders ○ Clinical setting
Mitchell and Beals, 2011	K-10	K-6	Diagnosis of depression on CIDI	<ul style="list-style-type: none"> ○ Is an effective screener for psychological disorder ○ Can be used as the first stage in a two-stage sampling process that screens for potential cases before a lengthy diagnostic interview 	13	<ul style="list-style-type: none"> ○ American Indians living on or near reservations
Fernandes et al., 2011	N/A	K10	Diagnosis of depression through the MINIPlus	<ul style="list-style-type: none"> ○ Are sensitive and specific instruments for screening during prenatal check-ups 	6	<ul style="list-style-type: none"> ○ Prenatal women in South India ○ Rural primary care setting
Campbell et al., 2008	EPDS	Translated EPDS			13	
Ekeroma et al., 2012	EPDS	Samoan and Tongan languages translated EPDS	Diagnosis of serious depression on WHO-CIDI v3	<ul style="list-style-type: none"> ○ Can measure the same construct as the standard EPDS ○ Can identify participants at risk of depression better than the standard EPDS 	12	<ul style="list-style-type: none"> ○ Aboriginal and Torres Strait Islander women (antenatal and postnatal)
Bass et al., 2008	EPDS and HSCL-D hybrid	Scale to detect Maladi ya Souci	Diagnosis of a depression case by a local non-professional	<ul style="list-style-type: none"> ○ Can detect a local post-partum syndrome consistent with the DSM-IV diagnosis of MDD 	10–17 (mixed results)	<ul style="list-style-type: none"> ○ Postnatal Samoan and Tongan women
Ashaba et al., 2019	BDI-II; HSCL	"20-item depression scale"	Diagnosis of MDD through the DSM-IV, MINI-KID	<ul style="list-style-type: none"> ○ Can measure affective and cognitive symptoms of MDD 	10	<ul style="list-style-type: none"> ○ Adolescents living with HIV in rural Uganda
Haroz et al., 2014	HSCL; HTQ; AUDIT	MHAP-I	N/A	<ul style="list-style-type: none"> ○ Is reliable in assessing psychosocial problems and functioning 	N/A	<ul style="list-style-type: none"> ○ Adult Burmese survivors of systemic violence in Thailand ○ Displaced community setting
Denckla et al., 2017	NOK	Kiswahili and Kikamba translated NOK	N/A	<ul style="list-style-type: none"> ○ Can detect for depression, anxiety and somatic factors of depression 	N/A	<ul style="list-style-type: none"> ○ Rural school aged children
Ganguli et al., 1999	GDS	GDS-H	N/A	<ul style="list-style-type: none"> ○ Can detect cognitive and functional impairment symptoms of depression ○ Has higher distribution of scores in this population than what is reported from the original scale 	11	<ul style="list-style-type: none"> ○ Illiterate elderly population in India

(Continued)

Table 1. (Continued)

Reference	Original scale	Adapted scale	Measured against ("Gold-standard")	Clinical use	Optimal culturally safe cut-off	Population that optimal cut-off is determined for
Sarkar et al., 2015	GDS-15	Tamil version of GDS-15	Diagnosis of depression on ICD-10	<ul style="list-style-type: none"> ○ Is reliable to give a point prevalence rate of geriatric depression ○ Indicates those at high-risk for clinical depression 	7	<ul style="list-style-type: none"> ○ Elderly people in South India Rural community
Haroz et al., 2017	HSCL-25	IDSS-G	Diagnosis of depression on SCID	<ul style="list-style-type: none"> ○ Can detect functional impairment and suicidal ideation ○ Has utility for detecting DSM defined disorders ○ Can be used by non-specialist providers 	0.56	<ul style="list-style-type: none"> ○ Non-Western populations Medical settings
Kaaya et al., 2008	HSCL-25	DSQ	N/A	<ul style="list-style-type: none"> ○ Is reliable to measure depression and anxiety ○ Can measure symptom constellation related to depressive disorder 	1.75	<ul style="list-style-type: none"> ○ Indigenous Tanzanian population Primary care setting
Marley et al., 2017	EPDS	KMMS	Diagnosis of depression on DSM-IV	<ul style="list-style-type: none"> ○ Is a safe and reliable screener for identifying people at clinically high risk of anxiety and depressive disorders ○ Is more feasible for use than the original scale 	9	<ul style="list-style-type: none"> ○ Aboriginal Australian perinatal women
Schneider et al., 2015	N/A	FAI	Diagnosis of depression on MINI 6.0	<ul style="list-style-type: none"> ○ Can be used to reliably screen maternal depression ○ Can complement or potentially replace (as a short instrument) other measures of functioning 	N/A	<ul style="list-style-type: none"> ○ Woman who are pregnant or have young babies in Cape Town
Almeida et al., 2014	N/A	KICA-dep	Diagnosis of a major depressive episode through ICD-10 and DSM-IV	<ul style="list-style-type: none"> ○ Useful to screen for presence of depression ○ High NPV is excellent at screening out non-depression cases 	8	<ul style="list-style-type: none"> ○ Older Indigenous Australians

Abbreviations: AUDIT, Alcohol Use and Disorders Identification Test; BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; CIDI, Composite International Diagnostic Interview; CQ, Coping Questionnaire; DSM, Diagnostic and Statistical Manual of Mental Disorders; DSQ, Dar-es-Salaam Symptom Questionnaire; EPDS, Edinburgh Postnatal Depression Scale; FAI, Functioning Assessment Instrument; GDS, Geriatric Depression Scale; HIV, human immunodeficiency virus infection; HSCL, Hopkins Symptom Checklist Depression Scale; HTQ, Harvard Trauma Questionnaire; ICD-10, International Classification of Diseases-10; IDSS-G, International Depression Symptom Scale-General; K-5/6/10, Kessler Psychological Distress Scale 5-/6-/10 item; KICA-dep, Kimberley Indigenous Cognitive Assessment of Depression; KMMS, Kimberley Mum's Mood Scale; MDD, major depressive disorder; MHAP-I, Mental Health Assessment Project Instrument; MINI(-KID), Mini International Neuropsychiatric Interview (-Children and Adolescents); NOK, Ndetei-Othieno-Kathuku Scale; PAS, psychiatric assessment schedule; PHQ, Patient Health Questionnaire; SCAN, Semi-Structured Schedules for Clinical Assessment in Neuropsychiatry Interview; SCID, Structured Clinical Interview for DSM Disorders; SRQ, Self-Reporting Questionnaire; SRT, Symptom Rating Test; WHO-CIDI v3, World Health Organization Composite International Diagnostic Interview.

Scales: light blue: CES-D; dark blue: EPDS; light purple: PHQ; dark purple: NOK; light gray: FAI; dark gray: SRQ; light orange: KICA; dark orange: Kessler; light pink: HSCL; dark pink: KMMS; light turquoise: IDSS-G; dark turquoise: GDS; brown: DSQ.

Table 2. Psychometric properties of adapted scales

Reference	Modifications	Reliability			Validity						Clinical utility							
		Internal consistency	Test-retest reliability	Inter-rater reliability	Concurrent (criterion) validity	Predictive (criterion) validity	Convergent (construct) validity	Discriminant (construct) validity	Cross-cultural validity	Incremental validity	Sensitivity	Specificity	Discrimination (AUC)	PPV	NPV	LR		
Baron et al., 2017	-h, t	$\alpha = 0.69\text{--}0.89$				$r(\text{FL}) = 0.30\text{--}0.83$					71.4%–84.6%	72.6%–95%	0.81–0.94	16.1%–54.8%				
Armenta et al., 2014	+l, +s	$\alpha = 0.85\text{--}0.88$			$p < 0.001$	RMSEA = 0.04–0.08, SRMR = 0.04–0.08												
Chapla et al., 2019	t	$\alpha = 0.71$				$r = 0.60, p < 0.001$	$r = -0.25, p < 0.001$											
Chapleski et al., 1997	*					$r(\text{FL}) = 0.26\text{--}0.84$, RMSE = 0.04–0.05, CFI = 0.86–0.93			RMSE = 0.06, CFI = 0.89									
Harry and Crea, 2018	+i	$\alpha = 0.84\text{--}0.87$				RMSEA = 0.03, CFI = 0.97, TLI = 0.97, WRMR = 0.99			RMSEA = 0.02, WRMR = 2.10, TLI = 1.0 CFI = 1.0									
Kilburn et al., 2018	-h, +s, t	$\alpha = 0.70\text{--}0.76$				$r(\text{FL}) = 0.37\text{--}0.86$			RMSEA = 0.07, TLI = 0.89, CFI = 0.91, SRMR = 0.05									
T Tiburcio Sainz and Natera Rey, 2007	f	$\alpha = 0.91$				$r(\text{factor}) = 0.61\text{--}0.90$												
Schantz et al., 2017	+i, +m, t	$\alpha = 0.81\text{--}0.83$			$r = 0.76$		$r(\text{FL}) = 0.40\text{--}0.74$, EVA = 3.21–3.44											
Caneo et al., 2020	t										74.6%	93.9%	0.92					
Esler et al., 2008	+a, f	$\alpha = 0.80$									80%–100%	12.5%–71.4%		30%–58.3%	70.8%–95.2%			
Gallis et al., 2018	+m, t	$\alpha = 0.84$									94.7%	88.9%	0.96	75.2%	97.9%			
Gelaye et al., 2013	+s, +d, t	$\alpha = 0.92$	ICC = 0.92		$p < 0.0001$		$r(\text{FL}) = 0.39\text{--}0.78$, EVA > 1				86%	67%	0.77	23%	96.4%	3.0		
Hackett et al., 2019	+a	$\alpha = 0.88$									81%–87%	72%–82%	0.88	43%–56%	94%–97%	2.8–4.6		
Husain et al., 2006	t										69.6%–93.1%	80.8%–85.2%		77.9%–78.9%	77.8%–94.1%			
Bougie et al., 2016	t	$\alpha = 0.84\text{--}0.88$, $r(\text{item-item}) = 0.21\text{--}0.71$				RMSEA = 0.05, CFI = 0.99, WRMR = 1.20–1.90												
McNamara et al., 2014	-h, +s, +m	$\alpha = 0.88\text{--}0.93$, $r(\text{item-total}) = 0.57\text{--}0.83$			$p < 0.0001$			CFI = 0.99, TLI = 0.99, RMSEA = 0.04										
Mitchell and Beals, 2011	*							$p < 0.05$						0.73–0.83				
Fernandes et al., 2011	t				$p < 0.01$						100%	81.3%–84.9%	0.95					

(Continued)

Table 2. (Continued)

Reference	Modifications	Reliability			Validity						Clinical utility					
		Internal consistency	Test-retest reliability	Inter-rater reliability	Concurrent (criterion) validity	Predictive (criterion) validity	Convergent (construct) validity	Discriminant (construct) validity	Cross-cultural validity	Incremental validity	Sensitivity	Specificity	Discrimination (AUC)	PPV	NPV	LR
Campbell et al., 2008	+m	$\alpha = 0.89\text{--}0.92$, r (item-total) = 0.48–0.88			$p < 0.05$											
Ekeroma et al., 2012	t	$\alpha = 0.86$			$\kappa = 0.57$						80%	80%	0.83–0.89	82%	86%	
Bass et al., 2008	t	$\alpha = 0.76\text{--}0.88$				$r = 0.34$, $p = 0.0001$	$p < 0.0001$						0.83–0.87			
Ashaba et al., 2019	+i, +l, +s, +a, t	$\alpha = 0.85\text{--}0.91$					$r(FL) = 0.40\text{--}0.68$				81%	78%	0.84			4.2
Haroz et al., 2014	+s	$\alpha = 0.79$	$r = 0.86$	$r = 0.86$	$p = 0.30\text{--}0.93$											
Denckla et al., 2017	t	$\alpha = 0.77\text{--}0.84$, r (item-item) = 0.17–0.53			$r = 0.15\text{--}0.45$			$r(FL) = 0.27\text{--}0.8$, RMSEA = 0.04, CFI = 0.93; TLI = 0.92								
Ganguli et al., 1999	t	$\alpha = 0.92$			$p < 0.05$			$r(FL) = 0.10\text{--}0.78$, EVA > 1								
Sarkar et al., 2015	t										80%	47.6%	0.66			
Haroz et al., 2017	+l, +s, +a	$\alpha = 0.92$	$r = 0.87$	$ICC = 0.90$	$p < 0.05$		$r = 0.56\text{--}0.78$			$VIF = 2.8$, $p = 0.001$	73%–77%	67%	0.72–0.75			
Kaaya et al., 2008	+d	$\alpha = 0.84$, $ICC = 0.89$, r (item-total) = 0.29–0.76	$ICC = 0.82$	$ICC = 0.89$	r (item-to-subscale) = 0.45–0.79		$r(FL) = 0.24$, EVA ≥ 1 , $r = -0.57\text{--}-0.37$	$r = -0.35\text{--}-0.49$								
Marley et al., 2017	p, f	$\alpha = 0.89$									82.6%–87%	75%–86.8%	0.86–0.90	54.1%–67.9%	93.7%–94.4%	
Schneider et al., 2015	*	$\alpha = 0.77$				$p < 0.0$										
Almeida et al., 2014	+i	$ICC = 0.88$									72–78%	82–90%	0.88	39%–50%	96%	

Abbreviations: AUC, area under the receiving operating curve; EVA, eigenvalue(s); ICC, intraclass correlation coefficient; FL, factor loading(s); LR, likelihood ratio; NPV, negative predictive value; PPV, positive predictive value.

Quality of values (Supplementary Material 2): red: poor psychometric qualities; yellow: moderate psychometric qualities; green: strong psychometric qualities.

Scales: light blue: CES-D; dark blue: EPDS; light purple: PHQ; dark purple: NOK; light gray: FAI; dark gray: SRQ; light orange: KICA; dark orange: Kessler; light pink: HSCL; dark pink: KMMS; light turquoise: IDSS-G; dark turquoise: GDS; brown: DSQ.

Modifications to scale: +/−, added/deleted; i, suicidal ideation items; h, hope items; l, loneliness items; s, somatic difficulty items; a, anger items; m, simplified language; d, local idioms of distress; p, changed scale administration protocol; t, translated; f, rephrased; *, other.

High alpha values were reported by all but four of the studies; moderate alphas corresponded with the CES-D, HSCL, and FAI (Haroz et al., 2014; Schneider et al., 2015; Kilburn et al., 2018; Chapla et al., 2019). These findings indicate good consistency in the types of questions used in the adapted measure as a whole.

Interclass correlation (ICC), item-item correlation, and item-total correlation are an alternative to alpha as a measure for reliability following adaptation. Six studies showed moderate properties of these metrics within the KICA-dep, K-10, EPDSb, NOK, DSQ, and K-5 (Campbell et al., 2008; Kaaya et al., 2008; Almeida et al., 2014; McNamara et al., 2014; Bougie et al., 2016; Denckla et al., 2017). Four studies demonstrated high properties of inter-rater and test-retest reliability (2 days to 1 week; Kaaya et al., 2008; Gelaye et al., 2013; Haroz et al., 2014, 2017).

Cross-cultural validity

Measurement invariance testing was conducted in 4/31 studies (12.9%; Chapleski et al., 1997; McNamara et al., 2014; Harry and Crea, 2018; Kilburn et al., 2018). These tests used multigroup confirmatory factor analysis (CFA) to provide evidence for cross-cultural validity. These studies showed that some adapted scales were invariant across different groups of peoples, such as the same Indigenous Nation living in distinct residential locations (i.e., urban, rural off-reservation, and reservation; Chapleski et al., 1997).

Criterion (concurrent and predictive) validity

Concurrent and predictive (criterion) validity was determined in 11/31 (35.5%) studies. Evidence for concurrent and predictive validity were reported through correlation analysis between scale ratings and ratings from the Diagnostic Interview Schedule or a well-established depression measure in the respective location (Ganguli et al., 1999; Campbell et al., 2008; Kaaya et al., 2008; Fernandes et al., 2011; Ekeroma et al., 2012; Gelaye et al., 2013; Armenta et al., 2014; Haroz et al., 2014; Denckla et al., 2017; Haroz et al., 2017; Schantz et al., 2017). About 7/11 studies (63.4%) measuring criterion validity demonstrated that the scale adaptations resulted in high levels of concordance with a psychiatric diagnosis of major depressive disorder (MDD) or a gold-standard measure of distress (Campbell et al., 2008; Kaaya et al., 2008; Fernandes et al., 2011; Gelaye et al., 2013; Armenta et al., 2014; Haroz et al., 2017; Schantz et al., 2017). However, 4/11 (36.4%) reports showed poor evidence for criterion validity (Kaaya et al., 2008; Ekeroma et al., 2012; Haroz et al., 2014; Denckla et al., 2017).

Construct (convergent and discriminant) validity

Convergent and discriminant (construct) validity was measured through CFA, exploratory factor analysis (EFA), correlation analysis, or multivariate regression in 18/31 (58%) studies (Chapleski et al., 1997; Ganguli et al., 1999; Tiburcio Sainz and Natera Rey, 2007; Bass et al., 2008; Kaaya et al., 2008; Gelaye et al., 2013; Armenta et al., 2014; McNamara et al., 2014; Schneider et al., 2015; Bougie et al., 2016; Baron et al., 2017; Denckla et al., 2017; Haroz et al., 2017; Schantz et al., 2017; Harry and Crea, 2018; Kilburn et al., 2018; Ashaba et al., 2019; Chapla et al., 2019). The majority of these studies showed evidence for a high level of construct

validity. Most of the adaptation processes improved the scales' ability to capture globally meaningful constructs of depression.

Incremental validity

Only 2/31 (6.5%) studies examined incremental validity through regression modeling (Mitchell and Beals, 2011; Haroz et al., 2017). Strong incremental validity suggested that these adapted scales were able to predict the severity of outcomes among the Indigenous population better than existing measures. In the two studies, the predicted outcomes included lifetime mood disorders, physical diagnosis, alcohol use, and impaired functioning (Mitchell and Beals, 2011; Haroz et al., 2017).

Clinical utility

Sensitivity and specificity were measured by 14/31 (45.2%) studies (Husain et al., 2006; Esler et al., 2008; Fernandes et al., 2011; Ekeroma et al., 2012; Gelaye et al., 2013; Almeida et al., 2014; Sarkar et al., 2015; Baron et al., 2017; Haroz et al., 2017; Marley et al., 2017; Gallis et al., 2018; Ashaba et al., 2019; Hackett et al., 2019; Caneo et al., 2020). Sensitivity and specificity are dependent on cut-off point(s) of the adapted scale. As such, 7/14 (50%) studies showed that the determined cut-off point of their adapted scales yielded high sensitivity (Esler et al., 2008; Fernandes et al., 2011; Gelaye et al., 2013; Baron et al., 2017; Marley et al., 2017; Gallis et al., 2018; Hackett et al., 2019), and in 7/14 (50%) cases, there was moderate sensitivity (Husain et al., 2006; Ekeroma et al., 2012; Almeida et al., 2014; Sarkar et al., 2015; Haroz et al., 2017; Ashaba et al., 2019; Caneo et al., 2020). In contrast, only 4/14 (28.5%) studies showed that the determined cut-off point yielded high specificity (Almeida et al., 2014; Baron et al., 2017; Gallis et al., 2018; Caneo et al., 2020), whereas in 10/14 (71.4%) cases, there was only low or moderate specificity (Husain et al., 2006; Esler et al., 2008; Fernandes et al., 2011; Ekeroma et al., 2012; Gelaye et al., 2013; Sarkar et al., 2015; Haroz et al., 2017; Marley et al., 2017; Ashaba et al., 2019; Hackett et al., 2019).

The cut-off score of an adapted scale also dictates the strength of the PPV, NPV, and AUC. These values must be strong in order to adapt scales to capture the true prevalence of depression among the Indigenous group. Fourteen (14/31, 45.1%) studies reported discrimination properties following cultural adaptation of the scale, determined by AUC (Tiburcio Sainz and Natera Rey, 2007; Bass et al., 2008; Fernandes et al., 2011; Ekeroma et al., 2012; Gelaye et al., 2013; Almeida et al., 2014; Sarkar et al., 2015; Baron et al., 2017; Haroz et al., 2017; Marley et al., 2017; Gallis et al., 2018; Ashaba et al., 2019; Hackett et al., 2019; Caneo et al., 2020). Of these, 4/14 (28.6%) scales had a strong discrimination and were able to discriminate between cases and non-cases of depression among the specific community (Fernandes et al., 2011; Baron et al., 2017; Gallis et al., 2018; Caneo et al., 2020). 9/31, 29.0%). Nine studies reported the PPV and NPV of the adapted instrument (Husain et al., 2006; Esler et al., 2008; Ekeroma et al., 2012; Gelaye et al., 2013; Almeida et al., 2014; Baron et al., 2017; Marley et al., 2017; Gallis et al., 2018; Hackett et al., 2019). Of these, only one scale had a high PPV (Ekeroma et al., 2012), but 8/9 (88.9%) had a high NPV (Husain et al., 2006; Esler et al., 2008; Ekeroma et al., 2012; Gelaye et al., 2013; Almeida et al., 2014; Marley et al., 2017; Gallis et al., 2018; Hackett et al., 2019).

Three studies reported the LR, and found moderate levels of LR following adaptation. (Gelaye et al., 2013; Ashaba et al., 2019; Hackett et al., 2019).

Scale performance

Differences in adaptation methods across studies yielded differences in scale performance after adaptation (Table 2). There is evidence that the adaptation methods impacted the psychometric properties of the scales, evidenced by several patterns across scales. The adapted CES-D ($n = 8$) was a particularly strong scale; there was evidence for its reliability, validity, and clinical utility across adaptations. The PHQ ($n = 7$) performed poorly on specificity and PPV, however, it had excellent internal consistency, sensitivity, NPV, and construct validity. The Kessler scales ($n = 4$) performed well across all validity tests and had high internal consistency. The EPDS ($n = 4$) had good criterion validity, PPV, and NPV, but only moderate sensitivity, specificity, and discrimination. The HSCL ($n = 3$) had good internal consistency and construct validity, but poor criterion validity and clinical utility metrics. The GDS ($n = 2$) did not have high sensitivity or specificity. The NOK ($n = 1$) had moderate internal consistency and high construct validity but poor criterion validity. The IDSS-G ($n = 1$) had excellent reliability metrics and incremental validity, but did not have high specificity, and discrimination. The DSQ ($n = 1$) had high reliability but poor validity. The KMMS ($n = 1$) had high internal consistency, sensitivity, specificity, and NPV, but poor PPV. The FAI ($n = 1$) had high construct validity and moderate internal consistency. The adapted KICA-DEP ($n = 1$) showed acceptable internal consistency, specificity, and NPV.

Discussion

In this review, we synthesized the global evidence for the psychometric properties of depression scales culturally adapted for Indigenous peoples. Many processes taken to develop and adapt such instruments were successful in improving the measures' reliability (internal consistency, test-retest reliability, and inter-rater reliability), convergent validity (construct and discriminant), and incremental validity. These processes included adding or deleting items, translation, and incorporating local idioms of distress. However, cultural adaptation methods had less success in improving criterion validity (concurrent and predictive) and cross-cultural validity (measurement invariance). Additionally, the adapted screening instruments were typically highly sensitive, which means they were useful for identifying individuals who might be depressed. Conversely, the specificity of many instruments was low. Despite most scales being acceptable among the population, for clinicians, it is possible that low specificity tools show no added value for screening within populations with a high prevalence of depression.

Some depression scales might be more globally suitable for use among Indigenous peoples than others. Among the studies, researchers chose 13 different original depression scales to adapt. Information on selection criteria and processes remain insufficient to conclude whether a higher number of reports indicates positive or negative characteristics of a scale. In this study, adapted CES-D outperformed other scales after adaptation, as it showed the most uniformly high psychometric properties.

A gap in reporting characteristics

Testing for quality metrics was sporadic; only 2/31 studies tested for the majority of quality metrics (Gelaye et al., 2013; Haroz et al., 2017). Out of 15 psychometric characteristics to assess the effects of cultural adaptation, reliability and construct validity were two of the most commonly tested psychometric properties. In contrast,

clinical utility metrics, including sensitivity, specificity, PPV, NPV, and LR, were less commonly measured. These results indicate a continued gap in the knowledge around the performance of adapted scales with Indigenous populations, namely, how different adaptation processes produce benefits for psychological testing in these communities. As such, we cannot draw strong conclusions about the clinical use of culturally adapted scales, nor about individual methods of cultural adaptation that maintain reliability and clinical utility.

Increasing reliability and validity through cultural adaptations

Most reliability metrics of adapted scales were excellent, particularly in alpha values. This finding suggests that following adaptations, scale items were relatively consistent and that instructions for scoring scales were unambiguous. Yet, a limitation to using alpha to determine internal consistency is that higher alpha's do not necessarily indicate higher quality of a new scale (Sijtsma, 2009). Exceptionally high alpha values (>0.9) could have resulted when the adaptation process adds length or redundant items to the original scale, or when the modifications constrict the crucial constructs which are measured by the scale (Panayides and Walker, 2013). Some studies assumed higher reliability of the developed scale following a high alpha value. In some cases, researchers even relied on alpha values to remove items that did not correlate well; however, this is not the developed purpose of alpha values (Cartagena-Ramos et al., 2018).

Most studies did not cover multiple types of validity. Based on few positive results, some adaptation processes improved understanding and acceptance among the Indigenous population. Many scales had high construct validity, showing that factors of the scales represented true constructs of depression known to the specific Indigenous group, such as affective or somatic symptoms. A few studies also found acceptable criterion validity, showing that the scales accurately responded to established criterion of depression used by gold-standard instruments with populations where the original scale is used. Measuring incremental validity of scales was important to understand if adapting scales, such as through incorporating local idioms of distress, predicted outcomes above and beyond previously established Western measures. A few scales with high incremental validity were useful to predict functional impairment above the scores on non-adapted measures; they showed which scores most accurately indicated mental health concern in that population.

Measurement invariance

A scale's measurement invariance, or cross-cultural validity, indicates how well a new scale minimizes the inter-rater differences, such as how different populations endorse scale items (Bader et al., 2021). Some scales proved to be measurement invariant following adaptation processes, meaning different cultural groups interpreted the constructs of depression in the scale in a conceptually similar way. Unfortunately, most studies did not compare the results of the scale's properties between Indigenous and non-Indigenous groups, or between different Indigenous groups (Bougie et al., 2016; Kilburn et al., 2018). In fact, without invariance testing, results might not even generalize to the same Indigenous group living on-reserves or in rural areas (Bougie et al., 2016). Therefore, it remains inconclusive whether these adaptation processes increased cultural safety without sacrificing the accuracy of the original scale. An adapted

scale should ultimately balance measurement invariance and diagnostic ability for it to be useful.

Clinical utility

For clinical utility, the fact that the prevalence of depression is high in many Indigenous communities should be considered in future studies. Increasing the acceptance of a screening tool through cultural adaptation may inadvertently increase its sensitivity by lowering the threshold to be considered a positive case (Shen et al., 2018). Specifically, we found that the sensitivity of adapted scales was often considerably higher than its specificity. Accordingly, PPV and NPV represent the percentage of individuals who truly do or do not have a depressive disorder (respectively), and a trend of high NPV and low PPV among the scales may be explained by the prevalence rates (Simon, 2015). The findings suggest that processes to culturally adapt scales can make them more effective at detecting positive cases of depression rather than screening out negative cases in the population (Labrique and Pan, 2010).

Moreover, a pattern of low PPV suggests the added or modified scale items may not have been disorder specific after cultural adaptation. Our results on the NPV and PPV of adapted scales align with previous literature stating that predictive values are one of the most important metrics to guide treatment of clinical practitioners (Labrique and Pan, 2010). Predictive values indicate the diagnostic capability of the test in the real-world and are thus referred to as the scale's *real-world performance* or *clinical relevance* to screening individuals who may benefit from a more accurate diagnostic interview and treatment guided by a diagnosis (Labrique and Pan, 2010). Although predictive value metrics were seldom reported by studies, studies have stated that this psychometric evaluation should not be forgotten. Predictive value scores reflect congruity with Western-based ideas about the definition of depression, how it manifests and should optimally be treated; community knowledge may enhance clinical utility and safety in presentations that do not match Western ideas (Haswell et al., 2010).

The studies indicate that to increase the *clinical relevance* of the adapted scales, it is necessary to increase NPV (minimizing false negatives), and to increase PPV (minimizing false positives). The findings in the original studies suggest that to increase NPV after adaptation, it is best to avoid or exclude scale items which are likely to be endorsed easily by the entire population (i.e., not measuring a locally specific symptom of depression) as may inaccurately increase the sensitivity of the scale for a true depressive case (Mitchell and Beals, 2011; Sarkar et al., 2015; Simon, 2015). Additionally, to increase PPV and to increase specificity, it is recommended to focus questions on motivational, cognitive and affective components of depression rather than on symptoms that may arise from non-psychiatric medical conditions, as these can inflate scores and make the scale less specific for a depressive case (Ganguli et al., 1999; Simon, 2015). The studies noted that increasing specificity is particularly necessary in low-resource settings, such as the settings where most adaptations were conducted. The value of a screen with high specificity in Indigenous communities is that it allows clinicians to effectively allocate resources such that the most in need receive timely treatment.

The AUC of a few studies was high, showing that the cut-off points determined after cultural adaptation had discriminative capacity (Schwarzbold et al., 2014). However, because few studies assessed AUC, we cannot reach a strong conclusion on the discrimination properties of most types of adapted depression scales. Similarly, the LRs of most adapted scales in the studies were

moderate, but there is a lack of data to reach conclusions on their ability to screen for probable depression. A high LR is the likelihood for having a high risk for depression – as indicated by the cut-off point of the adapted scale – for a person with current MDD compared to someone who does not (Hackett et al., 2019).

When could a cultural adaptation be useful?

Ideally, cultural adaptations would be useful when balancing two goals: a) increase acceptability through translating Western constructs of psychological problems to the Indigenous context, and b) support treatment (Kohrt et al., 2014; Chan et al., 2021). The strength of cultural adaptation is that it can improve scale items' reliability and validity through changes to language or representation of depression constructs. At the same time, there are limitations inextricably linked to these changes, including their impact on clinical utility.

Symptom measures are a Western concept, and the Western medicine has been built to treat individuals based on symptom severity (Bredström, 2019). Doctors need certain criteria to evaluate the need for treatment and to define recovery. Researchers have also used scale cut-offs for intake and to define how many benefitted from a certain intervention. An existing gold standard for a diagnosis is essential for validation after cultural adaptation, and to validate diagnostic cut-offs for a screen. Without tools available to have an accurate prevalence measure or diagnostic tool, it is difficult to ascertain true clinical utility of adapted scales. Defining a gold-standard for depression is harder, and a gold standard measure is often not available, in Indigenous areas (Kaaya et al., 2008; Haroz et al., 2014).

Future lines of work

Action is needed to address knowledge gaps around Indigenous mental health constructs as well as to understand how interventions, policies, or programs can support unmet needs of these populations. For this to be possible, there must first be an understanding about Indigenous concepts of emotional or mental distress and wellbeing. This forms a basis for treatments as well as tools to identify individuals and communities in need of support and interventions. However, in some instances, cultural adaptation might not be the primary approach in building trust and supporting empowerment (Gomez Cardona et al., 2021). As an alternative to cultural adaptation, methods of administration could also increase cultural safety. This includes at least considering the setting, language, use of community members at administration, and visual elements (Gomez Cardona et al., Submitted).

It is not rare for Indigenous communities to question the concept of a symptom focused approach that uses distinct cut-offs (Gomez Cardona et al., 2021). In fact, research teams have worked to co-design culture-specific tools for a culturally based, often community-targeted interventions supporting empowerment rather than symptom reduction to meet cut-offs (Haswell et al., 2010; Gomez Cardona et al., 2021). Thus, if you accept a non-symptom-based approach for interventions, you will not need symptom-based measures. This should be a main consideration when evaluating the necessity to culturally adapt a depression measure for a particular Indigenous group.

In the future, studies should include a needs assessment prior to developing new screening tools. If modifications are warranted, qualitative methods may be beneficial to understand the community's needs which can be addressed by a novel or adapted scale

(Gomez Cardona et al., Submitted). Concurrently, there must be emphasis on ensuring adaptation processes yield stable and clinically useful tools. This represents a shortcoming observed across the studies in this review; there was limited comprehensive testing of multiple psychometric domains. It is necessary to not only evaluate reliability and validity following changes to the scale, but also the utility of the new scale to accurately detecting the risk of depression.

Similar to what has been found by other researchers who adapted or developed new scales for use with Indigenous peoples, we advocate for future studies to examine scales' sensitivity to change (Haswell et al., 2010). Sensitivity to change reflects the competency of the scale to detect changes in mood, and as such, is an essential characteristic to evaluate treatment response. Evaluation and treatment of mental health relies on stable and measurement invariant tools for screening those in need of treatment, which should simultaneously signal the direction and magnitude of effect sizes of treatments to understand their efficacy for a particular population (Fitzpatrick et al., 2019).

Limitations of the current study

Five databases were systematically searched. Hence, it is possible that not all relevant studies were identified through systematic search strategies. To reduce the risk, gray literature found through hand-searching was incorporated in the review; this includes searching on open access repositories and through checking references of included articles. Several key terms do not have globally accepted terms, most importantly, *Indigenous* or *cultural adaptation*, which may have limited the scope of our screening. The conclusions made for validity and utility are limited by the fact that the included reports on cultural adaptation described a limited number of characteristics. This limits this study's ability to draw strong conclusions to guide a specific selection of tools for clinical use or for research.

We only considered studies published in the English language. This ensured that data for the studies' validation analyses were uniform, however, it is possible that some existing adaptation processes were left out of this review if they were reported in a different language. Further validation, analysis, and adaptation of new and existing measures of depression is needed and will confirm the applicability of such instruments in culturally distinct populations.

Risk of bias

The four domains of ROBIS were completed, which indicated this review was completed with a low risk of bias. This pertains to study selection, data collection and study appraisal, and data synthesis. The conclusions of the review are supported by the evidence presented and included consideration of the relevance of included studies. The methodology of the synthesis was driven by the nature of the studies and our research objectives, however, since a meta-analysis was not conducted, no statistical synthesis methods were undertaken.

Conclusion

Through a review of the literature, we found evidence that cultural adaptation may increase the validity of depression scales and their reliability to be used in mental health assessment of Indigenous populations across the world. The current review supports the use

of adapting scales to fit the Indigenous context, increasing its acceptability to community members and overall consistency. At the same time, psychometric testing of adapted scales highlights a potential caveat of losing clinical utility with too high sensitivity and low specificity. Cultural adaptation of depression assessments for Indigenous populations would be clinically useful when balancing two goals: 1) increase acceptability through translating Western constructs of psychological problems to the Indigenous context, and 2) support treatment.

Open peer review. To view the open peer review materials for this article, please visit <http://doi.org/10.1017/gmh.2023.52>.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/gmh.2023.52>.

Acknowledgments. We would like to thank Ms. Andrea Quaiattini from the McGill University Library for her assistance with our comprehensive systematic search procedure. We also express our gratitude for Indigenous co-authors in other publications and some other community members in Montreal and Kahnawà:ke for introducing us to some basic concepts of their view of health and wellbeing as opposed to symptom measurement.

Author contribution. Conceptualization: L.G.C., O.L.; Formal analysis: M.Y., Q.S., L.G.C., V.N., M.K., G.V., O.L.; Funding acquisition: O.L., L.G.C.; Investigation: M.Y., Q.S., V.N., M.K., O.L.; Methodology: L.G.C., Q.S., M.Y., O.L., V.N.; Project administration: O.L.; Resources: O.L.; Supervision: O.L.; Validation: O.L.; Visualization: M.Y., O.L.; Writing – original draft: M.Y., Q.S., L.G.C., O.L.; Writing – review and editing: M.Y., Q.S., L.G.C., V.N., M.K., G.V., O.L.

Financial support. O.L. was funded by grants from FSISSS (#8400958), CIHR Institute of Indigenous Peoples' Health (#426678), FRSQ (#252872 and #265693O), and the Réseau Québécois sur le suicide, les troubles de l'humeur et les troubles associés. L.G.C. had financial support from the Réseau universitaire intégré de santé et services sociaux (RUISSS) McGill and CIHR Institute of Indigenous Peoples' Health (#430331). Q.S. had funding from McGill University's Healthy Brains Healthy Lives fellowship. The funders had no role in study design or in later collection and data analysis.

Competing interest. The authors report no competing interests. The authors alone are responsible for the content and writing of the paper.

References

- Aboraya A, Nasrallah HA, Elswick DE, Ahmed E, Estephan N, Aboraya D and Dohar S (2018) Measurement-based care in psychiatry-past, present, and future. *Innovations in Clinical Neuroscience* **15**(11–12), 13–26.
- Almeida OP, Flicker L, Fenner S, Smith K, Hyde Z, Atkinson D, Skeaf L, Malay R and LoGiudice D (2014) The Kimberley assessment of depression of older Indigenous Australians: Prevalence of depressive disorders, risk factors and validation of the KICA-dep scale. *PLoS One* **9**(4), e94983. <https://doi.org/10.1371/journal.pone.0094983>
- Andrade C (2018) Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine* **40**(5), 498–499. https://doi.org/10.4103/ijpsym.Ijpsym_334_18
- Armenta BE, Sittner Hartshorn KJ, Whitbeck LB, Crawford DM and Hoyt DR (2014) A longitudinal examination of the measurement properties and predictive utility of the Center for Epidemiologic Studies Depression Scale among North American Indigenous adolescents. *Psychological Assessment* **26**(4), 1347–1355. <https://doi.org/10.1037/a0037608>
- Ashaba S, Cooper-Vince C, Vořechovská D, Maling S, Rukundo GZ, Akena D and Tsai AC (2019) Development and validation of a 20-item screening scale to detect major depressive disorder among adolescents with HIV in rural Uganda: A mixed-methods study. *SSM - Population Health* **7**, 100332. <https://doi.org/10.1016/j.ssmph.2018.100332>
- Bader M, Jobst IJ, Zettler I, Hilbig BE and Moshagen M (2021) Disentangling the effects of culture and language on measurement noninvariance in cross-cultural research: The culture, comprehension, and translation bias (CCT)

- procedure. *Psychological Assessment* **33**(5), 375–384. <https://doi.org/10.1037/pas0000989>
- Baron EC, Davies T and Lund C** (2017) Validation of the 10-item Centre for Epidemiological Studies Depression Scale (CES-D-10) in Zulu, Xhosa and Afrikaans populations in South Africa. *BMC Psychiatry* **17**(1), 6. <https://doi.org/10.1186/s12888-016-1178-x>
- Bass JK, Ryder RW, Lammers M-C, Mukaba TN and Bolton PA** (2008) Post-partum depression in Kinshasa, Democratic Republic of Congo: Validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine & International Health* **13**(12), 1534–1542. <https://doi.org/10.1111/j.1365-3156.2008.02160.x>
- Boksa P, Joober R and Kirmayer LJ** (2015) Mental wellness in Canada's Aboriginal communities: Striving toward reconciliation. *Journal of Psychiatry & Neuroscience* **40**(6), 363–365. <https://doi.org/10.1503/jpn.150309>
- Bougie E, Arim RG, Kohen DE and Findlay LC** (2016) Validation of the 10-item Kessler Psychological Distress Scale (K10) in the 2012 Aboriginal peoples survey. *Health Reports* **27**(1), 3–10.
- Bredström A** (2019) Culture and context in mental health diagnosing: Scrutinizing the DSM-5 revision. *Journal of Medical Humanities* **40**(3), 347–363. <https://doi.org/10.1007/s10912-017-9501-1>
- Campbell A, Hayes B and Buckley B** (2008) Aboriginal and Torres Strait Islander women's experience when interacting with the Edinburgh Postnatal Depression Scale: A brief note. *Australian Journal of Rural Health* **16**(3), 124–131. <https://doi.org/10.1111/j.1440-1584.2007.00930.x>
- Caneo C, Toro P and Ferreccio C** (2020) Validity and performance of the Patient Health Questionnaire (PHQ-2) for screening of depression in a rural Chilean cohort. *Community Mental Health Journal* **56**(7), 1284–1291. <https://doi.org/10.1007/s10597-020-00605-8>
- Cartagena-Ramos D, Fuentealba-Torres M, Rebustini F, Leite A, Alvarenga WA, Arcencio RA, Dantas RAS and Nascimento LC** (2018) Systematic review of the psychometric properties of instruments to measure sexual desire. *BMC Medical Research Methodology* **18**(1), 109. <https://doi.org/10.1186/s12874-018-0570-2>
- Chan AW, Reid C, Skeffington P and Marriott R** (2021) A systematic review EPDS cultural suitability with Indigenous mothers: A global perspective. *Archives of Women's Mental Health* **24**(3), 353–365. <https://doi.org/10.1007/s00737-020-01084-2>
- Chapla A, Prabhakaran A, Ganjiwale J, Nimbalkar S and Kharod N** (2019) Validation of the Gujarati version of Center for Epidemiological Studies Depression Scale for Children (CES-DC) and prevalence of depressive symptoms amongst school going adolescents in Gujarat, India. *Journal of Clinical and Diagnostic Research* **13**(12), VC06–VC11. <https://doi.org/10.7860/JCDR/2019/41001.1337>
- Chapleski EE, Lamphere JK, Kaczynski R, Lichtenberg PA and Dwyer JW** (1997) Structure of a depression measure among American Indian elders: Confirmatory factor analysis of the CES-D scale. *Research on Aging* **19**(4), 462–485. <https://doi.org/10.1177/0164027597194004>
- Denckla CA, Ndetei DM, Mutiso VN, Musyimi CW, Musau AM, Nandoya ES, Anderson KK, Milanovic S, Henderson D and McKenzie K** (2017) Psychometric properties of the Ndetei-Othieno-Kathuku (NOK) scale: A mental health assessment tool for an African setting. *Journal of Child and Adolescent Mental Health* **29**(1), 39–49. <https://doi.org/10.2989/17280583.2017.1310729>
- Ekeroma AJ, Ikenasio-Thorpe B, Weeks S, Kokaua J, Puniani K, Stone P and Foliaki SA** (2012) Validation of the Edinburgh Postnatal Depression Scale (EPDS) as a screening tool for postnatal depression in Samoan and Tongan women living in New Zealand. *New Zealand Medical Journal* **125**(1355), 41–49.
- Esler D, Johnston F, Thomas D and Davis B** (2008) The validity of a depression screening tool modified for use with Aboriginal and Torres Strait Islander people. *Australian and New Zealand Journal of Public Health* **32**(4), 317–321. <https://doi.org/10.1111/j.1753-6405.2008.00247.x>
- Esler DM, Johnston F and Thomas D** (2007) The acceptability of a depression screening tool in an urban, Aboriginal community-controlled health service. *Australian and New Zealand Journal of Public Health* **31**(3), 259–263.
- Fernandes MC, Srinivasan K, Stein AL, Menezes G, Sumithra R and Ramchandani PG** (2011) Assessing prenatal depression in the rural developing world: A comparison of two screening measures. *Archives of Women's Mental Health* **14**(3), 209–216. <https://doi.org/10.1007/s00737-010-0190-2>
- Fitzpatrick SA, Haswell MR, Williams MM, Nathan S, Meyer L, Ritchie JE and Jackson Pulver LR** (2019) Learning about Aboriginal health and well-being at the postgraduate level: Novel application of the growth and empowerment measure. *Rural and Remote Health* **19**, 2(2019), 4708. <https://doi.org/10.22605/RRH4708>
- Gallis JA, Maselko J, O'Donnell K, Song K, Saqib K, Turner EL and Sikander S** (2018) Criterion-related validity and reliability of the Urdu version of the patient health questionnaire in a sample of community-based pregnant women in Pakistan. *PeerJ (San Francisco, CA)* **6**, e5185–e5185. <https://doi.org/10.7717/peerj.5185>
- Ganguli M, Dube S, Johnston JM, Pandav R, Chandra V and Dodge HH** (1999) Depressive symptoms, cognitive impairment and functional impairment in a rural elderly population in India: A Hindi version of the Geriatric Depression Scale (GDS-H). *International Journal of Geriatric Psychiatry* **14**(10), 807–820.
- Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, Wondimagegn D, Lemenhe A, Fann JR, Vander Stoep A and Andrew Zhou XH** (2013) Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry Research* **210**(2), 653–661. <https://doi.org/10.1016/j.psychres.2013.07.015>
- Gomez Cardona L, Brown K, McComber M, Outerbridge J, Parent-Racine E, Phillips A, Boyer C, Martin C, Splicer B, Thompson D, Yang M, Velupillai G, Laliberté A, Haswell M and Linnaranta O** (2021) Depression or resilience? A participatory study to identify an appropriate assessment tool with Kanien'kéha (Mohawk) and Inuit in Quebec. *Social Psychiatry and Psychiatric Epidemiology* **56**(10), 1891–1902. <https://doi.org/10.1007/s00127-021-02057-1>. Epub 2021 Mar 8. PMID: 33683413.
- Gomez Cardona L, Yang M, Seon Q, Karia M, Velupillai G, Noel V and Linnaranta O** (Submitted) *The Methods of Improving Cultural Sensitivity of Depression Scales for Use among Indigenous Populations: A Systematic Scoping Review*. Montreal, QC: Douglas Mental Health University Institute.
- Hackett ML, Teixeira-Pinto A, Farnbach S, Glozier N, Skinner T, Askew DA, Gee G, Cass A and Brown A** (2019) Getting it right: Validating a culturally specific screening tool for depression (aPHQ-9) in Aboriginal and Torres Strait Islander Australians. *Medical Journal of Australia* **211**(1), 24–30. <https://doi.org/10.5694/mja2.50212>
- Haroz EE, Bass J, Lee C, Oo SS, Lin K, Kohrt B, Michalopoulous L, Nguyen AJ and Bolton P** (2017) Development and cross-cultural testing of the International Depression Symptom Scale (IDSS): A measurement instrument designed to represent global presentations of depression. *Global Mental Health (Cambridge)* **4**, e17. <https://doi.org/10.1017/gmh.2017.16>
- Haroz EE, Bass JK, Lee C, Murray LK, Robinson C and Bolton P** (2014) Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology* **2**(1), 31–31. <https://doi.org/10.1186/s40359-014-0031-6>
- Harry ML and Crea TM** (2018) Examining the measurement invariance of a modified CES-D for American Indian and non-Hispanic White adolescents and young adults. *Psychological Assessment* **30**(8), 1107–1120. <https://doi.org/10.1037/pas0000553>
- Haswell MR, Kavanagh D, Tsey K, Reilly L, Cadet-James Y, Laliberte A, Wilson A and Doran C** (2010) Psychometric validation of the growth and empowerment measure (GEM) applied with Indigenous Australians. *Australian & New Zealand Journal of Psychiatry* **44**(9), 791–799. <https://doi.org/10.3109/00048674.2010.482919>
- Husain N, Gater R, Tomenson B and Creed F** (2006) Comparison of the Personal Health Questionnaire and the Self Reporting Questionnaire in rural Pakistan. *The Journal of the Pakistan Medical Association* **56**(8), 366–370.
- Kaaya SF, Lee B, Mbawando JK, Smith-Fawzi MC and Leshabari MT** (2008) Detecting depressive disorder with a 19-item local instrument in Tanzania. *International Journal of Social Psychiatry* **54**(1), 21–33. <https://doi.org/10.1177/0020764006075024>
- Kilburn K, Prencipe L, Hjelm L, Peterman A, Handa S and Palermo T** (2018) Examination of performance of the Center for Epidemiologic Studies Depression Scale Short Form 10 among African youth in poor, rural households. *BMC Psychiatry* **18**(1), 201. <https://doi.org/10.1186/s12888-018-1774-z>
- Kohrt BA, Rasmussen A, Kaiser BN, Haroz EE, Maharjan SM, Mutamba BB, de Jong JT and Hinton DE** (2014) Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global

- mental health epidemiology. *International Journal of Epidemiology* **43**(2), 365–406. <https://doi.org/10.1093/ije/dyt227>
- Labrique AB and Pan WKY** (2010) Diagnostic tests: Understanding results, assessing utility, and predicting performance. *American Journal of Ophthalmology* **149**, e872–881. <https://doi.org/10.1016/j.ajo.2010.01.001>
- Leung L** (2016) Diabetes mellitus and the Aboriginal diabetic initiative in Canada: An update review. *Journal of Family Medicine and Primary Care* **5**(2), 259–265. <https://doi.org/10.4103/2249-4863.192362>
- Marley JV, Kotz J, Engelke C, Williams M, Stephen D, Coutinho S and Trust SK** (2017) Validity and acceptability of Kimberley Mum's mood scale to screen for perinatal anxiety and depression in remote Aboriginal health care settings. *PLoS One* **12**(1), e0168969. <https://doi.org/10.1371/journal.pone.0168969>
- Marsella A, Sartorius N and Jablensky A** (1985) Cross-cultural studies of depressive disorders: An overview. In Kleinman A and Good B (eds), *Culture and Depression; Studies in the Anthropology and Cross-Cultural Psychiatry of Affect and Disorder*. Berkeley: University of California Press, pp. 299–324.
- Mayberry RM, Mili F and Ofili E** (2000) Racial and ethnic differences in access to medical care. *Medical Care Research and Review* **57**(4), 108–145. <https://doi.org/10.1177/107755800773743628>
- McNamara BJ, Banks E, Gubhaju L, Williamson A, Joshy G, Raphael B and Eades SJ** (2014) Measuring psychological distress in older Aboriginal and Torres Strait Islanders Australians: A comparison of the K-10 and K-5. *Australian and New Zealand Journal of Public Health* **38**(6), 567–573. <https://doi.org/10.1111/1753-6405.12271>
- Mitchell CM and Beals J** (2011) The utility of the Kessler screening scale for psychological distress (K6) in two American Indian communities. *Psychological Assessment* **23**(3), 752–761. <https://doi.org/10.1037/a0023288>
- Panayides P and Walker MJ** (2013) Evaluating the psychometric properties of the foreign language classroom anxiety scale for Cypriot senior high school EFL students. *Rasch Measurement Approach* **9**(3), 493–516. <https://doi.org/10.5964/ejop.v9i3.611>
- Parikh R, Mathai A, Parikh S, Chandra Sekhar G and Thomas R** (2008) Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology* **56**(1), 45–50. <https://doi.org/10.4103/0301-4738.37595>
- Porcerelli JH and Jones JR** (2017) Uses of psychological assessment in primary care settings. In Maruish ME (ed.), *Handbook of Psychological Assessment in Primary Care Settings*. New York: Routledge, pp. 75–94. <https://doi.org/10.4324/9781315658407>
- Sarkar S, Kattimani S, Roy G, Premarajan KC and Sarkar S** (2015) Validation of the Tamil version of short form Geriatric Depression Scale-15. *Journal of Neurosciences in Rural Practice* **6**(3), 442–1446. <https://doi.org/10.4103/0976-3147.158800>
- Schantz K, Reighard C, Aikens JE, Aruquipa A, Pinto B, Valverde H and Piette JD** (2017) Screening for depression in Andean Latin America: Factor structure and reliability of the CES-D short form and the PHQ-8 among Bolivian public hospital patients. *International Journal of Psychiatry in Medicine* **52**(4–6), 315–327. <https://doi.org/10.1177/0091217417738934>
- Schneider M, Baron E, Davies T, Bass J and Lund C** (2015) Making assessment locally relevant: Measuring functioning for maternal depression in Khayelitsha, Cape Town. *Social Psychiatry and Psychiatric Epidemiology* **50**(5), 797–806. <https://doi.org/10.1007/s00127-014-1003-0>
- Schouten BC and Meeuwesen L** (2006) Cultural differences in medical communication: A review of the literature. *Patient Education and Counseling* **64**(1), 21–23. <https://doi.org/10.1016/j.pec.2005.11.014>
- Schwarzbold ML, Diaz AP, Nunes JC, Sousa DS, Hohl A, Guarneri R, Linhares MN and Walz R** (2014) Validity and screening properties of three depression rating scales in a prospective sample of patients with severe traumatic brain injury. *Brazilian Journal of Psychiatry* **36**(3), 206–212. <https://doi.org/10.1590/1516-4446-2013-1308>
- Shen YT, Radford K, Daylight G, Cumming R, Broe TGA and Draper B** (2018) Depression, suicidal behaviour, and mental disorders in older Aboriginal Australians. *International Journal of Environmental Research and Public Health* **15**(3), 447. <https://doi.org/10.3390/ijerph15030447>
- Sijtsma K** (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* **74**(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Simon B and Catherine W** (2009) Cultural safety: Exploring the applicability of the concept of cultural safety to Aboriginal health and community wellness. *Journal of Aboriginal Health* **5**(2), 6.
- Simon R** (2015) Sensitivity, specificity, PPV, and NPV for predictive biomarkers. *Journal of the National Cancer Institute* **107**(8), djv153. <https://doi.org/10.1093/jnci/djv153>
- Tiburcio Sainz M and Natera Rey G** (2007) Adaptación al contexto náhuatl del Cuestionario de Enfrentamientos (CQ), la Escala de Síntomas (SRT) y la Escala de Depresión del Centro de Estudios Epidemiológicos (CES-D). *Salud Mental (México)* **30**(3), 48–58.
- Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R, ROBIS group and ROBIS group** (2016) ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology* **69**, 225–234. <https://doi.org/10.1016/j.jclinepi.2015.06.005>