


REVIEW ARTICLE

Psychometric properties of outcome measures in non-pharmacological interventions of persons with dementia in low-and middle-income countries: A systematic review

Bharath DU ¹, Monisha LAKSHMINARAYANAN,² Murali KRISHNA,³ Sridhar VAITHESWARAN,⁴ Mina CHANDRA,⁵ Shaji KUNNUKATTIL SIVARAMAN,⁶ Satyapal Puri GOSWAMI,⁷ Thara RANGASWAMY,⁸ Aimee SPECTOR⁹ and Charlotte R. STONER¹⁰

¹Foundation for Research and Advocacy in Mental health (FRAMe) and ⁷All India Institute of Speech and Hearing, Mysore, ²Schizophrenia Research Foundation, Chennai, ⁵Atal Bihari Vajpayee Institute of Medical Sciences and Dr Ram Manohar Lohia Hospital, New Delhi and ⁶Government Medical College, Thrissur, India and ⁹University College London and ¹⁰University of Greenwich, London, UK

Correspondence: Dr Bharath Du, MD Psychiatry, No 98, Kapila Cross, Gururaja Layout, Mysore-11, India. Email: bharathdu@gmail.com

Disclosure: We declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Global Alliance for Chronic Diseases (GACD) MH02
Indian Council of Medical Research (ICMR)
UK Medical Research Council (MRC)

Received 19 August 2020; revision received 7 November 2020; accepted 26 November 2020.

Key words: dementia, low middle income countries, non-pharmacological therapy, outcome measures, psychometric properties.

INTRODUCTION

Demographic ageing is a global phenomenon and the most important social transformation of the 21st century.¹ Of all the chronic non-communicable diseases (NCDs) related to ageing, dementia and cognitive impairment are the leading contributors to disability, and particularly, dependence among older people worldwide.²

Abstract

Despite high burden of dementia in low-and middle-income countries (LMICs), only a small number of clinical trials of psychosocial interventions for persons with dementia (PwD) have been conducted in these settings. It is essential that such trials use appropriate outcome measures that are methodologically robust and culturally appropriate to evaluate the effectiveness of interventions. We carried out a systematic review to examine the evidence base and psychometric properties of measures employed in these studies in LMICs. A systematic search of published literature on randomised controlled trials (RCT) of psychosocial interventions for PwD in LMICs between 2008 and April 2020 was carried out. Measures employed in each of the eligible studies were identified and through a focused search, we further explored the evidence base and psychometric properties employing Terwee criteria. Data extraction and quality appraisal were conducted by two independent reviewers. The review identified 41 measures from 17 RCTs which fulfilled eligibility criteria and they examined effectiveness across the domains of cognition ($n = 16$), behaviour and psychological symptoms ($n = 11$) and quality of life ($n = 8$). Of these 41, we were able to access relevant literature only for 18 and they were subject to psychometric analysis. Psychometric properties of these 18 instruments were at best modest, with Terwee scores ranging from 3 (low) to 15 (moderate). A majority of the studies were from China ($n = 5$) and Brazil ($n = 6$). The evidence base for the routinely employed measures in RCTs of non-pharmacological interventions for PwD in LMICs is limited. The quality of adaptation and validation of these instruments is variable and studies are largely uninformative about their psychometric properties and cultural appropriateness to the study setting. There is an urgent need to develop scientifically robust instruments in LMIC settings that can be confidently employed to measure outcomes in trials of psychosocial interventions for PwD.

Worldwide, around 50 million people live with dementia, and this is estimated to reach 75 million by 2030.³ Two in three people with dementia live in low- and middle-income countries (LMIC).⁴ This poses a huge challenge for governments to plan and design viable assessment and treatment options for persons with dementia suitable for their countries. In LMIC settings,

dementia is often seen as part of normal ageing, is under-recognised, under-disclosed, under-treated, and under-managed.⁵ These factors make evaluation, treatment and research on dementia in these settings uniquely challenging, with specialist and culturally specific tools, methods for assessment and monitoring of treatment required.

Considering the aforementioned complexities and challenges of evaluation and treatment of dementia, the development of novel, tailor-made therapeutic interventions is required for LMIC settings. Among all the interventions available, psychosocial interventions are particularly important and suitable as they are typically low cost and less resource intensive. They are more relevant to those settings where access to medicines and specialists is restricted. However, in this era of evidence-based medicine, these novel psychosocial interventions need to be tested for their feasibility, efficacy and applicability in local contexts using gold standard randomised controlled trials (RCTs). Selecting appropriate outcome measures is a critical step in designing valid and useful clinical trials for persons with dementia, as the use of an unreliable measure may result in important information about the effectiveness of an intervention being lost or distorted.⁶

Choosing an appropriate outcome measure is even more important in LMIC settings as a significant number of measures used in intervention trials for persons with dementia were originally developed in high income countries (HICs). As there is little standardisation of methods for adaptation of these measures, their current 'adaptation' varies from cross-cultural adaptation with adequate methodology to informal verbatim translation. There is no consensus as to which measures are most appropriate or psychometrically robust for use in persons with dementia.

The aims of this systematic review are to:

- 1 Identify outcome measures that are used to evaluate the effectiveness of psychosocial interventions for persons with dementia in LMICs
- 2 Conduct a quality appraisal of the psychometric properties of each of the outcome measures
- 3 Provide recommendations for use of outcome measures, based on their psychometric robustness.

METHODS

Design

A systematic search of published literature from 2008 to 2019 on psychosocial interventions delivered to persons with dementia in LMICs was previously conducted by authors of this team.⁷ Results from this search consisted of 17 studies, describing 11 interventions in six countries. A repeat search was run in April 2020 using the published search strategy and the process of the systematic review is shown in Figure 1. Each of the studies included in this systematic review was subject to an additional search to identify relevant outcome measures used and focused searches were used to identify articles that described the development or adaptation of these measures for the countries in question. All included measures were subject to a quality appraisal to determine validity and reliability by employing Terwee criteria.⁸ This systematic review followed the standard Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic review and a checklist for the same has been submitted as a Appendix S1 for further reading.⁹

Search strategy

The full search strategy is described in a related review.⁷ Briefly, Cochrane Controlled Register of Trials (CENTRAL), PubMed, EMBASE, PsycINFO and MEDLINE were searched for studies published between 2008 and April 2020. Search terms fell under the categories: psychosocial or non-pharmacological interventions, LMICs and people with dementia. LMICs were designated as such according to their classification by the Organisation for Economic Co-operation and Development.¹⁰ The list of included studies was then examined and all outcome measures for cognitive, psychological and social domains were extracted. The reference list was then examined to identify articles that described the development of these measures for the country in question or, if it was an existing measure, the article that described the translation and adaptation of the measure for the country in question. These articles are subsequently described as 'measure development or adaptation papers' and were included regardless of publication date. In cases where the reference given was for an English language

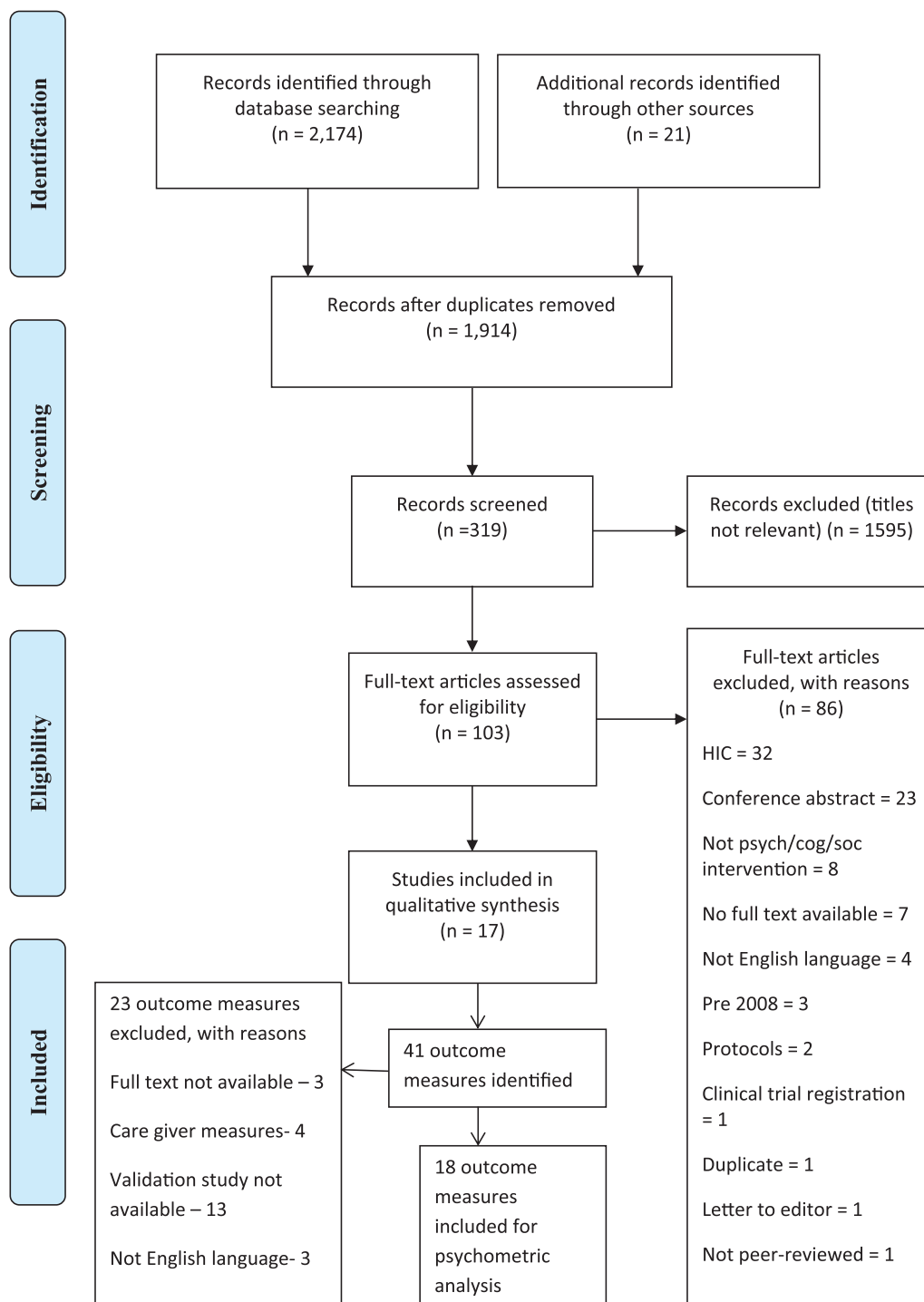


Figure 1 Process of systematic review (flow chart).

measure but not for the translated or adapted version, the corresponding author was contacted to ascertain which version of the measure was used.

Inclusion and exclusion criteria

A measure development or adaptation paper was included if:

- 1 The domain measured was deemed by authors to be cognitive, psychological or social in nature
- 2 The outcome measure was used in an intervention study to examine change over a period of time in persons with dementia, as an indicator of benefit derived from the intervention
- 3 It was published in a peer-reviewed journal.

A measure development or adaptation paper was excluded if:

- 1 It was published in a language other than English and an English translation was not available.

Quality assessment

Included measures were grouped by domain (cognition, behaviour, depression, anxiety and quality of life) and a quality assessment was undertaken independently by two authors (BD and ML) against the Terwee criteria,⁸ used successfully in related reviews.^{11, 12} The Terwee criteria are based on a list of nine common psychometric properties: content validity, internal consistency, criterion validity, construct validity, reproducibility, responsiveness, floor and ceiling effects and interpretability and, for each domain listed, a maximum score of two was awarded if the psychometric properties were correctly evaluated and were within an acceptable range. A score of one per criterion was awarded if the methodology reported was flawed and zero was awarded if no information was reported or psychometric properties reported fell outside the acceptable range. Full criteria are provided in Table 1. After the initial appraisal, authors BD and ML discussed their ratings and any discrepancies until a consensus was reached.

RESULTS

Our search yielded 17 intervention studies from six LMICs. Studies were conducted in Brazil ($n = 6$), China ($n = 5$), India ($n = 2$), Tanzania ($n = 2$), Turkey ($n = 1$) and Argentina ($n = 1$). A wide range of interventions were evaluated: reality orientation, cognitive stimulation therapy (CST), reminiscence therapy, music therapy, tailored rehabilitation programs, games and other activities were used for the treatment of dementia. Forty-one outcome measures were identified, of which 16 were primarily measures of cognition, three measured behavioural and psychological symptoms or distress in dementia, eight

measured depression and anxiety, eight measured quality of life, four measured caregiver burden and two measured disability.

Only 18 outcome measures were included for psychometric analysis, as 23 had to be excluded for the following reasons: there was no evidence of validation or adaptation of the chosen outcome measure for the study setting ($n = 13$), inability to access the full articles ($n = 3$), scale measured other outcomes related to caregivers ($n = 4$) and the validation studies were not in English language ($n = 3$).

Most authors had provided the citation of the original development article of an outcome measure in English, but these studies lacked information related to cultural adaptation of the outcome measure for use in the study setting. For example, many authors referenced Folstein et al., 1975 for Mini Mental Status Examination (MMSE), which is an original development article.¹³ It is likely that many would have used a verbatim translated outcome measure (informal linguistic translation) instead of a systematically translated, adapted and validated measure (cultural adaptation) for study population and settings. Hence, we contacted the corresponding authors of all the eligible studies by email to obtain further clarification about the measures they had employed in their study, including the references for those measures. If no response was received after 2 weeks of initial contact, we sent them another email reminder. However, only three of the 17 authors (Li, Asiret and Camargo) replied. Li and colleagues had used linguistically translated (without formal adaptation and validation) outcome measures which were widely in use in China, while Asiret and Camargo had used culturally adapted and validated scales in Turkish and Portuguese languages respectively but had referenced original English developmental articles of the outcome measure. After discussing as a team, we decided to assume that authors who did not respond were likely to have used either a culturally adapted or verbatim translated version of the original outcome measure. Hence, for the purpose of this review, further searches were undertaken to identify the culturally adapted and validated version of measures specific to each country.

A summary of all the 17 intervention studies is given in Table 2. Psychometric properties of the outcome measures are described below and ratings of these measures based on Terwee criteria ARE

Table 1 Terwee criteria

Property	Definition		Quality criteria	
1	Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire (the extent to which the measure represents all facets of the construct under question).	+2	A clear description of measurement aim, target population, concept(s) that are being measured, and the item selection AND target population (investigators OR experts) were involved in item selection.
			?1	A clear description of the above-mentioned aspects in lacking OR only target population involved OR doubtful design or method.
			-0	No target population involvement.
			00	No information found on target population involvement.
2	Internal consistency	The extent to which items in a (sub)scale are inter-correlated, thus measuring the same construct.	+2	Factor analyses performed on adequate sample size (7* # items and ≥ 100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95
			?1	No factor analysis OR doubtful design or method
			-0	Cronbach's alpha(s) < 0.70 or > 0.95 , despite adequate design and method
			00	No information found on internal consistency
3	Criterion validity	The extent to which scores on a particular questionnaire relate to a gold standard	+2	Convincing arguments that gold standard is 'gold' AND correlation with gold standard ≥ 0.70
			?1	No convincing arguments that gold standard is 'gold' OR doubtful design or method
			-0	Correlation with gold standard < 0.70 , despite adequate design and method
			00	No information found on criterion validity
4	Construct validity	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured	+2	Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses
			?1	Doubtful design or method (e.g. no hypotheses)
			-0	Less than 75% of hypotheses were confirmed, despite adequate design and methods
			00	No information found on construct validity
5	Reproducibility			
5.1	Agreement	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	+2	SDC $<$ MIC OR MIC outside the LOA OR convincing arguments that agreement is acceptable
			?1	Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable)
			-0	MIC \leq SDC OR MIC equals or inside LOA despite adequate design and method
			00	No information found on agreement
5.2	Reliability	The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)	+2	ICC or weighted Kappa ≥ 0.70
			?1	Doubtful design or method
			-0	ICC or weighted Kappa < 0.70 , despite adequate design and method
			00	No information found on reliability
6	Responsiveness	The ability of a questionnaire to detect clinically important changes over time	+2	SDC or SDC $<$ MIC OR MIC outside the LOA OR RR $>$ 1.96 OR AUC ≥ 0.70
			?1	Doubtful design or method
			-0	SDC or SDC \geq MIC OR MIC equals or inside LOA OR RR \leq 1.96 or AUC < 0.70 , despite adequate design and methods
			00	No information found on responsiveness
7	Floor and ceiling effects	The number of respondents who achieved the lowest or highest possible score	+2	$\leq 15\%$ of the respondents achieved the highest or lowest possible scores

Table 1 Continued

Property	Definition	Quality criteria
8 Interpretability	The degree to which one can assign qualitative meaning to quantitative scores	?1 Doubtful design or method
		-0 >15% of the respondents achieved the highest or lowest possible scores, despite adequate design and methods
		00 No information found on interpretation
		+2 Mean and SD scores presented of at least four relevant subgroups of patients and MIC defined
		?1 Doubtful design or method OR less than four subgroups OR no MIC defined
		00 No information found on interpretation

In order to calculate a total score + = 2; ? = 1; - = 0; 0 = 0 (scale of 0–s18). SDC, smallest detectable difference (this is the smallest within person change, above measurement error. A positive rating is given when the SDC or the limits of agreement are smaller than the MIC); MIC, minimal important change (this is the smallest difference in score in the domain of interest which patients perceive as beneficial and would agree to, in the absence of side effects and excessive costs); SEM, standard error of measurement; AUC, area under the curve; RR, responsiveness ratio; LOA, limits of agreement.

tabulated in Table 3 and a further summary of their validation studies is given in Appendix S1.

OUTCOME MEASURES RELATED TO COGNITION

A total of eight scales that measure cognition were included. The Chinese Montreal Cognitive Assessment Basic (MoCA- BC) scored most robustly on psychometric properties with a score of 15/18. Alzheimer's Disease Assessment Scale – Cognitive sub-Saharan Africa (ADAS-Cog SSA - 10/18), the Revised Turkish MMSE (r MMSE –Turkish- 10/18) and the Short Cognitive Test (SKT Brazil version – 9/18) showed moderate score on psychometric analysis while Toronto Alexithymia Scale 20 Chinese (TAS 20 Chinese - 7/18), Consortium to Establish a Registry for Alzheimer's Disease- neuropsychological battery (Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Brazil - 7/18), MMSE-Brazil (6/18) and Chinese adapted MMSE (6/18) scored poorly.

Chinese version of MoCA-BC

The MoCA basic (MoCA-B) was developed by Nasreddine (1996) in Canada to screen for MCI in illiterate individuals and those with little education.^{31, 49} The Chinese version of the MoCA-B (MoCA-BC) was translated from the original English version. The MoCA-BC was reported to have good content validity and criterion-related validity (Pearson correlation coefficient of MoCA-BC vs. MMSE = 0.787) and reliable

internal consistency (Cronbach's $\alpha = 0.807$). This scale showed good responsiveness, with the area under the receiver operating characteristic (ROC) curve more than 0.8 across all education levels in Chinese older adults. Inter-rater reliability was also excellent with intraclass coefficient value of 0.96 ($P < 0.001$).³¹

Alzheimer's Disease Assessment Scale – Cognitive subscale for sub-Saharan Africa- ADAS-Cog SSA

The ADAS-Cog was developed in the 1980s to assess the level of cognitive dysfunction in Alzheimer's disease (AD), but its use has extended into pre-dementia studies despite concerns about its ability to detect important changes at these milder stages of disease progression.^{32, 50} One team adapted the ADAS-Cog for use in sub-Saharan African settings with low literacy levels. The area under the ROC curve as 0.973 (95% CI = 0.936–1.00) for dementia, indicating good responsiveness of the scale. Internal consistency was high (Cronbach's $\alpha = 0.884$) and inter-rater reliability was excellent (intraclass correlation coefficient (ICC) 0.905, 95% CI 0.804–0.964). The scale also showed excellent content and criterion validity with convincing arguments and demonstration of strong correlation with severity of dementia measured with Clinical Dementia Rating Scale (CDR).³²

Short Cognitive Performance Test Brazilian version-SKT Brazil

The SKT is a bedside cognitive screening battery designed to detect memory and attention deficits.^{34, 51}

Table 2 Summary of intervention studies

Sr. no	Study authors, country and language of study setting	Study description	Outcome measures used in the study	Evidence of adaptation or validation of outcome measure for study setting	Comments
1	Machado et al., 2009 ¹⁴ Brazil Portuguese	RCT to determine effect of participation of AD patients in a multidisciplinary rehabilitation program on cognition, depression and QOL	MMSE – Mini Mental State Examination GDS-Geriatric Depression Scale QOL-AD - Quality of life in AD	Yes (C) Yes (C) Yes (A)	No significant change in cognition from pre to post intervention in both experimental and control group. No significant change in depression scores from pre to post intervention in both experimental and control groups. No significant change in QOL from pre to post intervention in both experimental and control groups.
2	Niu et al., 2010 ¹⁵ China Chinese	RCT to evaluate efficacy of cognitive stimulation therapy on individual dimensions of neuropsychiatric symptoms in AD patients	NPI - Neuropsychiatric inventory MMSE – Mini Mental State Examination	Yes (C) Yes (C)	Total score showed a significant improvement (-2.06 points, SE = 0.35) compared with a slight decline (0.00 points, SE = 0.26) in the control group ($t_{1/4} = -4.766, P < 0.001$) Significant improvement in score by 0.81 points, (SE = 0.28) in the treatment group and decline by 0.19 points (SE = 0.16) in the control group ($t = 3.106, P = 0.004$). Significant increase in total scores of control group ($P = 0.05$) indicating deterioration of cognition in control group. Significant improvement in scores ($P = 0.004$) in experimental group post intervention. No significant change pre and post intervention in both groups.
3	Viola et al., 2011 ¹⁶ Brazil Portuguese	To evaluate the effect of a multifunctional stimulation program on cognition, neuropsychiatric symptoms, and QOL in patients with mild AD in a controlled, single-blind design	SKT - Short Cognitive Test QOL-AD - Quality of life in AD NPI - Neuropsychiatric inventory MMSE – Mini Mental State Examination GDS - Geriatric Depression Scale	Yes (A) Yes (A) Yes (C) Yes (C)	No significant change pre and post intervention in both groups. Significant reduction in scores ($P = 0.001$) in experimental group post intervention.
4	Azcurra, 2012 ¹⁷ Argentina Spanish	RCT to evaluate the efficacy of reminiscence therapy in improving QOL in dementia patients	CDR – clinical dementia rating scale CPS – cognitive performance scale SES - Social Engagement Scale SRQoL - the resident self-reported RAID - rating anxiety in dementia The Zarit Burden Interview Minimum Data Set - bad depression rating scale Well-being/ ill-being scale	No No No No No No No	No significant change pre and post intervention in both groups. Significant reduction in scores ($P = 0.001$) in experimental group post intervention.

Table 2 Continued

Sr. no.	Study authors, country and language of study setting	Study description	Outcome measures used in the study	Evidence of adaptation or validation of outcome measure for study setting	Comments
5	Kumar <i>et al.</i> , 2014 ¹⁸ India Hindi	RCT to evaluate the impact of a novel occupational therapy program on QOL of patients with mild to moderate dementia	MMSE – Mini Mental State Examination TSI – test for severe impairment WHOQOL – Bref WHO quality of life brief	No No Yes (A)	Significant improvement of scores ($P < 0.001$) post intervention in experimental group.
6	Lin <i>et al.</i> , 2015 ¹⁹ China Chinese	To evaluate the impacts of a GO- game (Chinese chess) intervention on AD in a Northeast Chinese population and follow up evaluation at 6 months	MADRS- Montgomery-Asberg depression rating scale KICA Depression - Kimberley Indigenous Cognitive Assessment of Depression HADS- Hospital anxiety and depression scale	No No Yes (C)	A reduction of HADS mean score by 1.75 points (95% CI, 0.17–3.68) post intervention in experimental group.
7	Santos <i>et al.</i> , 2015 ²⁰ Brazil Portuguese	To evaluate the effects of a multidisciplinary rehabilitation program on cognitive ability, quality of life and depression symptoms in patients with AD and cognitive impairment without dementia (CIND).	GAF - Global assessment of functioning RAND 36 TAS 20 - Toronto alexithymia scale MMSE – Mini Mental State Examination GDS - Geriatric Depression Scale QOL-AD - Quality of life in AD	No Yes (C) Yes (C) Yes (C) Yes (A) Yes (A)	Statistically significant increase of 4.61 points (95% CI, -2.75–11.32) when compared with a control group ($P < 0.05$), providing evidence for efficacy of GO game program No statistically significant difference between the experimental and control groups Significant increase in mean MMSE scores in both mild AD ($P = 0.021$) and CIND patients ($P = 0.005$). Mild AD patients ($P < 0.001$) and CIND patients ($P = 0.011$) in Experimental group had reduction in depressive symptoms Mild AD patients who received the intervention had improvements in quality of life with significant increase in mean scores of QOL-AD Brazil ($P = 0.003$)
8	Camargo <i>et al.</i> , 2015 ²¹ Brazil Portuguese	To assess the effectiveness of reality orientation when combined with acetylcholinesterase inhibitors in the treatment of mild and moderate AD dementia.	CERAD neuropsychological battery MMSE – Mini Mental State Examination CDT (clock drawing test)	Yes (B) Yes (B) No	Significantly higher mean CERAD score ($P = 0.03$) compared to control group after 6 months follow up Treatment group scored significantly higher mean MMSE score ($P = 0.03$) compared to control group. Full text article not accessible

Table 2 Continued

Sr. no	Study authors, country and language of study setting	Study description	Outcome measures used in the study	Evidence of adaptation or validation of outcome measure for study setting	Comments
9	Asiret & Kapucu, 2015 ²² Turkey Turkish	To investigate the effect of reminiscence therapy on the cognitive status, depression, and daily living activities of institutionalised patients with mild and moderate AD	MMSE – Mini Mental State Examination GDS - Geriatric Depression Scale	Yes (B) Yes (B)	Significant increase ($P < 0.05$) in mean r MMSE-T score of therapy group compared to the control group Validation study in Turkish language, English translation could not be traced.
10	Raguraman et al., 2017 ²³ India Tamil	To culturally adapt, validate, and test the feasibility of delivering UK-based CST as an acceptable intervention in Chennai, India	Feedback forms		Not a standard outcome measure
11	Paddick et al., 2017 ²⁴ Tanzania Swahili	To conduct a trial of CST in a rural setting in Tanzania and evaluate its usefulness as a treatment for dementia using a stepped-wedge design with randomisation	WHOQOL-Bref - WHO Quality of Life measure - Brief WHODAS 2 – WHO disability assessment scale ADAS-COG – Alzheimer's Disease Assessment Scale – Cognitive subscale Barthel index - activities of daily living HADS- Hospital anxiety and depression scale Zarit Burden Interview NPI- Neuropsychiatric Inventory MMSE – Mini Mental State Examination	No No Yes (A) No No - - Yes (C)	Disability assessment scale. Significant improvement in cognition as evidenced by significant reduction in mean ADAS-COG score after 8 weeks of therapy Dependency assessment scale
12	Li & Li, 2017 ²⁵ China Chinese	To investigate the efficacy of a Chinese folk recreational program on symptoms among older people with dementia in China	NPI - Neuropsychiatric Inventory	Yes (C)	Care giver burden scale Carer rating was used in the study Mean scores of MMSE increased significantly from baseline to week 16 ($P < 0.01$) in the experimental group, while for the control group, the mean score of MMSE decreased significantly ($P < 0.01$). Mean score of CNPI - symptom decreased significantly ($P < 0.01$) in the experimental group
13	de Oliveira et al., 2018 ²⁶ Brazil Portuguese	To investigate the effectiveness of the TAP intervention adapted for an outpatient memory clinic (tailored activity program–outpatient version	Barthel index - activities of daily living NPI - Neuropsychiatric Inventory	No Yes (A)	Treatment group had a significant decrease in hallucination ($P = 0.04$), agitation ($P = 0.03$), anxiety ($P = 0.02$), aggression ($P = 0.01$), sleep disorder ($P = 0.02$), aberrant motor behaviour ($P = 0.02$)

Table 2 Continued

Sr. no	Study authors, country and language of study setting	Study description	Outcome measures used in the study	Evidence of adaptation or validation of outcome measure for study setting	Comments
14	Lyu <i>et al.</i> , 2018 ²⁷ China Chinese	(TAP-O) on reducing NPS and caregiver burden in patients with dementia. To explore the effects of music therapy on cognition, BPSD, and ADL of AD patients and their caregiver distress	Zarit Burden Interview MMSE – Mini Mental State Examination WHOUCLA AVLT- The WHO University of California-Los Angeles, Auditory Verbal Learning Test The semantic verbal fluency test NPI - Neuropsychiatric Inventory	No Yes (C) No No Yes (C)	Care giver burden measure No significant difference in scores in experimental group pre and post intervention Validation study in Chinese language was found, but English translation could not be traced. Experimental group scored significantly lower compared to control group post intervention
15	Mkenda <i>et al.</i> , 2018 ²⁸ Tanzania Nigeria	To describe the adaptation and feasibility assessment of CST as a potential low-resource intervention for dementia in Tanzania and Nigeria.	Zarit Burden Interview WHOUCLA - WHO Quality of Life measure ADAS-COG – Alzheimer's Disease Assessment Scale – Cognitive subscale NPI - Neuropsychiatric Inventory	No No Yes (C) Yes (A)	Data were not analysed to assess improvements in outcome due to the small number of participants involved in the study Experimental group had significantly reduced total NPI score ($P = 0.00$; Cohen $d = 0.93$), number ($P = 0.00$; Cohen $d = 0.93$), frequency ($P = 0.00$; Cohen $d = 1.12$), intensity ($P = 0.00$; Cohen $d = 0.77$) of Behavioural and Psychological Symptoms in Dementia (BPSD) Care giver burden measure Caregivers in the experimental group had reported improvement in their own QOL ($P < 0.05$; Cohen $d = 0.57$) and that of the person with dementia ($P < 0.01$; Cohen $d = 0.56$). No differences were found in the ratings of QOL by the person with dementia themselves.
16	Novelli <i>et al.</i> , 2018 ²⁹ Brazil Portuguese	To evaluate the effect of TAP-BR (tailored activity program - Brazil) on the number, frequency, and intensity of BPSD and QOL of persons with dementia	Zarit Burden Interview QOL-AD - Quality of life in AD	Yes (A)	Patients in the music therapy group demonstrated a significant improvement ($P = 0.003$) over patients in the control group Patients in The music therapy group demonstrated a significant improvement in
17	Wang <i>et al.</i> , 2018 ³⁰ China Chinese	To evaluate effects of music therapy on cognitive function and behaviour of mild AD patients receiving pharmacological intervention	MMSE – Mini Mental State Examination NPI - Neuropsychiatric Inventory	Yes (C) Yes (C)	Patients in the music therapy group demonstrated a significant improvement ($P = 0.003$) over patients in the control group Patients in The music therapy group demonstrated a significant improvement in

Table 2 Continued

Sr. no	Study authors, country and language of study setting	Study description	Outcome measures used in the study	Evidence of adaptation or validation of outcome measure for study setting	Comments
			MoCA - Montreal Cognitive Assessment	Yes (C)	NPI scores ($P < 0.01$) over patients in the control group Patients in the music therapy group demonstrated a significant improvement ($P < 0.01$) over patients in the control group

A, provided by the authors in the study; B, provided by authors after request for clarification by Email; C, found by focused search.

Flaks and colleagues have validated a Brazilian version of the SKT and reported the area under ROC ranging between 0.7 and 1, suggesting that the SKT adequately discriminates AD from participants without dementia (MCI and controls), irrespective of education. Inter-rater and test-retest agreement, floor and ceiling effects were not reported by the authors. However, authors have mentioned that the preliminary study in Brazil showed good internal consistency, with Cronbach's α equal to 0.8 and significant correlation with MMSE and the CDT (clock drawing test).³⁴

Toronto Alexithymia Scale 20 Chinese version -TAS 20-Chinese

Bagby and colleagues developed TAS 20 in 1994 from an earlier 26 item version developed by them.^{36, 52} It has three subscales: Difficulty Describing Feelings to others (DDF), Difficulty Identifying Feeling (DIF) and Externally-Oriented Thinking (EOT) designed to measure deficiency in understanding, processing, or describing emotions. Zhu and colleagues translated the TAS to Chinese with involvement of Chinese psychologists and developers of the original English TAS and reported good content validity. Confirmatory factor analysis was conducted, which showed that a three factor model showed best acceptable standards and a Cronbach's alpha >0.7 showed high internal consistency. Test-retest coefficient for the whole scale together and subscales were >0.7 showing good test-retest reliability. However, there was no information on criterion validity, construct validity, responsiveness, floor and ceiling effects or inter-rater agreement in the article.³⁶

CERAD – Brazil version

The CERAD was funded in 1986 by the National Institute on Ageing to develop a standardised assessment tool of AD for use by all Alzheimer Disease Centres established in the United States.^{35, 53} It consists of a clinical battery, neuropsychological battery, neuroimaging battery, family history scale, behavioural problems scale, family history assessment, services assessment, autopsy resources and educational brochures. Bertolucci and colleagues evaluated its validity in Brazil and reported that all the tests in CERAD had good sensitivity and specificity ranging from 73–97% and 67–87% respectively, with the exception of the Boston Naming Test with

Table 3 Psychometric rating of outcome measures based on Tenree criteria

Construct	Outcome measure version	Content validity	Internal consistency	Criterion validity	Construct validity	Agreement	Reliability	Responsiveness	Floor and ceiling effects	Interpretability	Total
Cognition	MoCA-BC Chinese ³¹	+2	?1	+2	+2	?1	+2	+2	+2	?1	15
Cognition	ADAS-Cog SSA ³²	+2	?1	+2	00	00	+2	+2	00	?1	10
Cognition	r MMSE-Turkey ³³	+2	?1	?1	00	+2	+2	+2	00	00	10
Cognition	SKT-Brazil ³⁴	+2	?1	+2	?1	00	00	+2	00	?1	9
Cognition	CERAD-Brazil ³⁵	+2	00	00	+2	00	00	+2	00	?1	7
Cognition	TAS 20-Chinese ³⁶	+2	+2	00	?1	+2	00	00	00	00	7
Cognition	MMSE-Brazil ³⁷	+2	00	?1	00	00	00	+2	00	?1	6
Cognition	CAMSE ³⁸	+2	00	?1	00	+2	00	00	00	?1	6
BPSD	NPI-Brazil ³⁹	+2	?1	00	00	+2	+2	00	?1	?1	9
BPSD	NPI-clinician - Brazil ⁴⁰	+2	00	?1	+2	00	+2	00	00	?1	8
BPSD	NPI-Chinese ⁴¹	+2	+2	00	?1	+2	00	00	00	00	7
Depression and anxiety	HADS Chinese ⁴²	+2	?1	?1	?1	00	00	+2	00	?1	8
Depression	GDS-30 Brazil ⁴³	+2	?1	+2	00	00	00	+2	00	00	7
Depression	GDS-15 Brazil ⁴⁴	+2	00	+2	00	00	00	+2	00	00	6
QOL	RAND Chinese ⁴⁵	+2	+2	?1	+2	+2	00	00	00	?1	10
QOL	WHOQOL-Bref ⁴⁶	+2	?1	?1	+2	+2	00	00	00	00	9
QOL	QOL-AD-Brazil ⁴⁷	+2	?1	+2	+2	00	00	00	00	?1	8
QOL	WHOQOL-Hindi ⁴⁸	+2	?1	00	00	00	00	00	00	00	3

BPSD, Behavioural and Psychological Symptoms of Dementia; QOL, quality of life.

sensitivity of 61% and Constructional Praxis with specificity of 51%. All the tests showed good responsiveness with areas under ROC curve ranging between 0.7 and 0.9. However, internal consistency, criterion validity, test–retest reliability, inter-rater reliability, floor and ceiling effects have not been reported.³⁵

MMSE

Folstein and his colleagues formulated the MMSE, a 30-point psychological tool for measuring cognitive impairment.¹³ Since then it has been adapted to multiple languages and regions and extensively used in clinical and research settings.⁵⁴ In this review, three culturally adapted MMSE scales were evaluated.

Revised MMSE Turkey version – r MMSE-T

The authors reported areas under ROC curve in educated and uneducated older people to be 0.953 and 0.907 respectively, which indicates good responsiveness of the outcome measure in detecting clinically important change in cognitive function over time.³³ The scale had good content validity, internal consistency, inter-rater and intra-rater agreement with Cronbach's α and kappa values higher than 0.7 for both educated and uneducated older people. Cut-off point of 22/23 of r MMSE-T in the educated older people had the highest sensitivity (90.9), specificity (97.0) and positive likelihood ratio (30.3), whereas cut-off point of 18/19 of the test in uneducated older people had the highest sensitivity (82.7%), specificity (92.3%) and positive likelihood ratio (10.7). Construct validity, floor and ceiling effects of the scale have not been reported.³³

MMSE-Brazil version - MMSE-Brazil

A modified translated Portuguese version of the MMSE, proposed by Bertolucci and colleagues in 1994 and Almeida and colleagues in 1998 was used in this validation study.^{37, 55, 56} The authors involved geriatricians in item selection during measure adaptation and reported good content validity. Sensitivity, specificity, positive and negative predictive values were 80.8%, 65.3%, 44.7% and 90.7% respectively for a cut-off point of 23/24. The area under the ROC curve was 0.807, indicating good responsiveness. Criterion validity has been tested with diagnosis of dementia by geriatricians using structured interviews based on Diagnostic and Statistical Manual 4th

edition (DSM-IV) and International Classification of Diseases Edition 10 (ICD-10). However, information on other psychometric measures such as internal consistency, inter-rater agreement, test–retest reliability, construct validity, responsiveness and floor and ceiling effects was lacking.³⁷

Chinese adapted MMSE – CAMSE

The CAMSE was adapted from the original MMSE with some changes in test items to minimise literacy dependency and render them compatible with Chinese culture, while the main structures of the original test were kept intact and similar principles for scoring were used as much as possible.³⁸ This suggests that the CAMSE tests the same cognitive functions as the original MMSE. Literate participants scored a higher CAMSE total score than illiterate participants ($P < 0.05$) to yield optimal cut-off points of 22 for literates and 20 for illiterates with a sensitivity of 83.87% and a specificity of 84.48%. Corresponding positive predictive value (PPV) was 0.65, and negative predictive value (NPV) was 0.94. The test–retest reliability tested after 4–6 weeks for total scores was 0.75 ($P < 0.01$). However, the article lacked information on internal consistency, criterion validity, construct validity, responsiveness, floor and ceiling effects and interpretability of the scale.

OUTCOME MEASURES RELATED TO BEHAVIOURAL AND PSYCHOLOGICAL SYMPTOMS IN DEMENTIA (BPSD)

NPI - Neuro-Psychiatric Inventory

The NPI is a tool which measures behavioural disturbances in dementia using two separate scales for rating the severity of each symptom and the distress caused to the caregiver respectively.⁵⁷ It was originally developed by Cummings *et al.* in 1994.⁵⁷ This scale was used in eight studies across three countries – Brazil, China and Tanzania. Adaptation studies of NPI Brazil and China versions are reviewed here, while adaptation to Tanzania could not be traced.

NPI - Brazil

This tool received a Terwee score of 9 and reported test–retest reliability (Spearman's rho for total severity = 0.82), internal consistency (Cronbach's $\alpha = 0.7$ for both severity and distress scales) and inter-rater

reliability (ICC severity = 0.98, distress = 0.96).³⁹ It also provided ample information on content validity for the Portuguese translation, and both ceiling and floor effects for some items. It was one of the few papers that provided some information on floor and ceiling effects. However, it was uninformative on criterion validity, construct validity or responsiveness.³⁹

NPI - Brazil- clinician version

This adapted tool scored 8/18 on the Terwee scale as it lacked information on internal consistency, agreement, responsiveness and floor and ceiling effects.⁴⁰ The validation focused mostly on inter-rater reliability (ICC of 0.923) and convergent validity with seven other scales, each of which measure various behavioural problems in dementia, with a sample of 156 participants. Convergent validity with the Apathy Inventory, Cohen-Mansfield agitation index, Cornell Scale for depression in dementia and Brief Psychiatric Rating scale – delusions was high (Pearson correlation $r \geq 0.7$) but was poor with Brief Psychiatric Rating Scale – hallucinations ($r = 0.432$). Even though the authors mentioned conducting test–retest reliability analysis, the results were not reported in the paper.⁴⁰

Chinese NPI – CNPI

This tool scored 7 on the Terwee scale and had clear information on content validity, internal consistency (Cronbach's $\alpha = 0.69$ for the severity and 0.72 for the caregiver distress scale) and agreement (test–retest correlation coefficient between 0.66 and 0.98).⁴¹ Construct validity was also analysed through the Kaiser-Meyer-Olkin value which confirmed that there were five common factors present within the tool. Of note, there were no clear hypotheses tested in the paper.⁴¹

OUTCOME MEASURES RELATED TO DEPRESSION AND ANXIETY

Hospital Anxiety and Depression Scale - Chinese version - HADS Chinese

The HADS was originally developed by Zigmond and Snaith (1983) to screen for depression and anxiety in general hospital patients.⁴² Leung and colleagues validated a Chinese-Cantonese version of the HADS against the Hamilton Rating Scale of Depression

(HRSD) and Hamilton Rating Scale of Anxiety (HRSA) and reported good internal consistency (Cronbach's $\alpha = 0.86$) and concurrent validity (Pearson's coefficient = 0.67 and 0.63, respectively; $P < 0.001$) with favourable sensitivity (0.79; 95% CI = 0.66–0.90) and specificity (0.80; 95% CI = 0.69–0.91) for screening for psychiatric disorders. However, its performance was marginally inferior to that of the HRSD. The authors did not report test–retest reliability, inter-rater agreement, floor and ceiling effect and hence scored moderately (8/18) on Terwee criteria. As the validation has been done in a general population, this questions its applicability in dementia research.⁴²

Geriatric Depression Scale Brazil - 30 item version - GDS 30 Brazil and 15 item version - GDS 15 Brazil

GDS was originally developed in 1983 by Yesavage.^{43, 44, 58} The original 30 item version of the GDS has been shortened and separately adapted and validated into scales with 15, 10, four and one item(s) across many languages and cultures.

Two studies from Brazil have validated the GDS for use in the local community. Paradela and colleagues validated the shortened GDS-15 version with a geriatric population.⁴⁴ This study obtained a Terwee score of 6/18, while Castelo *et al.* validated GDS 30 and scored 7/18 points on the Terwee scale. Both studies reported on the content validity adequately (with description on translation and back translation by experts), criterion validity (against DSM-IV criteria based diagnosis provided by a trained clinician) and responsiveness (area under the curve value above 0.9).^{43, 44} Castelo *et al.* validated all versions of the GDS (30, 15, 10, four and one) and additionally reported on internal consistency (Cronbach's $\alpha = 0.7$ or above in all of the tools).⁴³ Both lacked information on construct validity, test–retest and inter-rater agreement, floor and ceiling effects and interpretability.^{43, 44}

OUTCOME MEASURES RELATED TO QUALITY OF LIFE (QOL)

Four measures examined QOL in persons with dementia. The Chinese Short Form health survey-36 (SF-36) scored 10/18 and World Health Organization QOL assessment scale brief (WHOQOL- BREF)

scored 9/18, while the Brazilian version of the QOL - Alzheimer's disease (QOL-AD) scored 8/18 and WHOQOL-Hindi scored only 3/18.

Chinese Short Form health survey-36 - SF-36/ RAND 36 Chinese

The SF-36 was developed as part of a medical outcomes study.^{45, 59} Li *et al.* in 2003 adapted and validated it for Chinese use. The content validity was found to be good with a clear description of measurement aim, target population, concept being measured and involvement of target population in item selection. Convergent validity and discriminant validity were satisfactory for all except the social functioning scale. The Cronbach's α coefficient ranged from 0.72 to 0.88 except 0.39 for the social functioning scale and 0.66 for the vitality scale. Test-retest reliability coefficients (at 2 weeks) ranged from 0.66 to 0.94. Factor analysis identified two principal components explaining 56.3% of the total variance. Interrater reliability, responsiveness and floor and ceiling effects were not reported.⁴⁵

QOL for patients with AD Brazilian version - QOL-AD Brazil

Logsdon *et al.* proposed the QOL-AD, which has three versions: two addressing the patient's QOL: one for the patient himself/herself (PQOL) and another for the caregiver perception of patient's QOL- CPQOL), and a third related to the QOL of the Caregiver- (CQOL).^{47, 60} The QOL-AD has been translated and adapted to Portuguese by Novelli *et al.* Authors reported Cronbach's α of more than 0.8 for all the three versions. Content validity and construct validity were found to be good with convincing arguments for the same. Criterion validity was not determined as there was no instrument available for evaluation of QOL in dementia in Portuguese. The authors did not report test-retest and inter-rater reliability, responsiveness, floor and ceiling effects.⁴⁷

WHO QOL assessment scale brief - WHOQOL- BREF and WHOQOL-Hindi

WHOQOL-BREF has been derived from the WHOQOL-100 tool, which was developed by the WHOQOL Group in 15 international field centres as a cross-culturally applicable QOL assessment tool.^{46, 48, 61, 62} The authors reported high correlations ranging from

0.89 to 0.95 between domain scores based on the WHOQOL-100 and WHOQOL-BREF. Cronbach's α ranged from 0.66 to 0.84 demonstrating good internal consistency. Content validity and test-retest reliability (range from 0.66 to 0.87) was good, while discriminant validity was excellent. However, inter-rater reliability, responsiveness and floor and ceiling effects were not reported.⁴⁶

WHOQOL-Bref Hindi was developed in Delhi, one of the 15 centres in the WHOQOL study. The authors reported that the Hindi version and other national versions were compatible and comparable, as the WHOQOL was developed simultaneously in many centres across the world. However, the article was uninformative about the psychometric properties of WHOQOL-Bref Hindi.⁴⁸

DISCUSSION

Eighteen outcome measures related to persons with dementia were identified (covering the constructs of cognition, behavioural and psychological symptoms, QOL, anxiety and depression) from 17 psychosocial intervention studies in LMICs. All of these were culturally adapted and validated versions from an original English measure, indicating a lack of indigenously developed measures in the native language/s of LMIC. Most measures achieved a modest score on their adaptation procedures, with the MoCA-Chinese version scoring highest (15/18) and the WHOQOL-Bref Hindi scoring the lowest (3/18) on Terwee criteria.

In intervention studies involving persons with dementia, the most commonly employed indicators of effectiveness are measures of cognition. Of the nine outcome measures for cognition, the MoCA-BC (Chinese) was the most robustly developed, while the SKT Brazil version, ADASCOG-SSA and r MMSE-T gave moderate results on psychometric analysis. These tools appear to be adequate measures of cognition in patients with dementia. The TAS 20 Chinese version, CERAD Portuguese version, CAMSE and MMSE-Brazil version scored low on psychometric analysis and need further psychometric examination before they can be used routinely. All the cognition measures were validated in geriatric populations except TAS 20 Chinese version, which has been validated in undergraduate students.

BPSD form another important dimension of dementia research. The NPI is one of the most widely used tools for evaluating BPSD and all the three versions - NPI Brazil clinician version, NPI Brazil version and NPI Chinese version - have been developed with moderate robustness and are adequate to detect and measure BPSD. However, further adaptation and validation of NPI to other languages and regions of LMIC is essential. The HADS Chinese, GDS 30 Brazil and GDS 15 Brazil used to measure anxiety, depression in hospital patients and depression in geriatric population respectively, have been developed with moderate robustness. However, the HADS Chinese is validated for general hospital patients and its validity for research in dementia is questionable and requires further psychometric examination before it can be routinely used with confidence.

Quality of life is a more recent but firmly established theme in dementia research, facilitating an integrative model for dementia treatment and care. The QOL-AD Brazil, WHOQOL-Bref and Chinese SF-36 appear to be adequate measures of QOL, while WHOQOL-Bref Hindi appears to be a poor measure of QOL as the authors did not report most psychometric parameters. The Chinese SF-36 and WHOQOL have been validated in general populations and their validity for research in dementia is questionable and requires further psychometric examination before they are routinely used.

Many studies had not used adequate methodology for transcultural adaptation of an outcome measure, instead used an informally translated measure for validation. Cultural adaptation of a tool involves the production of an equivalent instrument for a target population, one that measures the same phenomenon in the original and the target cultures, rather than a verbatim translation. The first phase of the process includes a translation of words and sentences from the original language to another and then further linguistic adaptation to the cultural context of the target population to ensure that the new version is conceptually and culturally pertinent. The second phase of the cultural adaptation includes a validation phase during which the instrument is proven to be psychometrically equivalent to the original version.^{63, 64} Even when translated versions are in a population's native language, there can be cultural differences in the verbal expression of concepts, in meaning, and in relevance that may affect confidence

in the validity of results obtained using the translation.⁶⁵ Furthermore, a verbatim translated measure of cognition would increase the possibility of false positive rates of dementia as participants undergoing the test might skip or give wrong answers due to lack of understanding of the questions and alien concepts of the test, rather than cognitive deficits. This highlights the need for use of transcultural adaptation of outcome measures with adequate methodology in place of informal linguistic translations.

METHODOLOGICAL ISSUES AND LIMITATIONS

All the measures included here failed to define minimal important change, which is a requisite of Terwee criteria for interpretability and responsiveness. Except for MoCA-BC, no other validation study reported the floor and ceiling effects. This meant scoring the measures for interpretability and floor and ceiling effect was nearly impossible. Even though most authors reported Cronbach's alpha, they failed to report information on factor analysis performed on adequate sample size, leading to poor scores on internal consistency. We also noted that many authors had reported sensitivity, specificity, PPV and NPV in their validation papers, but these statistical tools are not included in the Terwee criteria. This suggests that researchers consider sensitivity and specificity as important tools to be tested in a validation study and further hints toward the need for a more inclusive and comprehensive psychometric criterion, which includes sensitivity and specificity of outcome measures in the psychometric analysis.

Referencing in scientific literature is very important as it gives the readers an understanding of the source of the information and also enables them to find the source of information for further reading if necessary. However, if the standard guidelines for referencing are not adhered in scientific articles, it undermines the purpose of referencing. In this review, we found many researchers citing the reference of an original development article of the outcome measure instead of the actual culturally adapted and validated version used in the research work in the country in question. Furthermore, some validation studies for these outcome measures were difficult to locate and could only be located by extensive searching. Also, many outcome measures had

to be excluded from the review as adapted versions could not be found despite exhaustive searching. This warrants a need to promote and sensitise researchers about standard referencing guidelines.

Although we employed broad search criteria to identify potentially eligible studies, it is still possible that we may have missed out some studies due to heterogeneous nature of reporting changes in psychosocial interventions studies among persons with dementia.

IMPLICATIONS FOR RESEARCH AND PRACTICE

Our review highlights the need for researchers to examine and ensure appropriate psychometric properties of outcome measures to be included in their research, while designing the research protocol and use outcome measures designed for a specific population, for a particular age group, region, culture and language to avoid skewed results and for better applicability of results in the population in question. Researchers should also provide references to the specific adapted version of an outcome measure correctly, in addition to referencing an original outcome measure developed in a different study setting. This review highlights limited availability of indigenously developed, culturally appropriate and validated outcome measures in LMIC, which may have inadvertently led the investigators of the studies included in this systematic review to use verbatim translated instruments. Even though most studies included in this review reported statistically significant effect of the intervention across domains of cognition, psychological symptoms and QOL, little is known about its clinical effectiveness.

This review indicates that MoCA-BC (for cognition) and Chinese SF-36 (for QOL), SKT Brazil version (for cognition) and NPI Brazil (for BPSD), ADASCOG-SSA (for cognition) and r MMSE-T (for cognition) can be used in dementia research with confidence in China, Brazil, sub-Saharan Africa and Turkey respectively. Researchers should be aware of lack of psychological robustness of other outcome measures evaluated here. We suggest researchers exercise caution about the psychometric properties of outcome measures while choosing outcome measures for their research pursuits and, also while interpreting results of an intervention study from a LMIC setting. LMICs are

characterised by populations with distinctively different cultures and spoken languages that are specific to a region within a country, which limits the generalisability and applicability of outcome measures and results of an intervention study beyond the study setting. Therefore, the first step in planning an intervention study for persons with dementia in LMICs should be to develop culture and context specific measures in their language/s and establish their psychometric properties.

CONCLUSION

The evidence base for the routinely employed measures in RCTs of non-pharmacological interventions for persons with dementia in LMICs is limited. The quality of adaptation and validation of these instruments is variable and studies are largely uninformative about their psychometric properties and cultural appropriateness to the study setting. There is an urgent need to develop scientifically robust instruments in LMIC settings that can be confidently employed to measure outcomes in trials of psychosocial interventions for persons with dementia.

ACKNOWLEDGMENTS

We acknowledge the following Global Alliance for Chronic Diseases (GACD) funding agencies: Indian Council of Medical Research (ICMR) and UK Medical Research Council (MRC) for funding the project Cognitive Stimulation Therapy-International (CST-International) during which the systematic review has been conducted. [Correction added on 25 February, after first online publication: The list of funders on the first page and the Acknowledgments section have been updated and rewritten, respectively, to include 'UK Medical Research Council (MRC)'.]

REFERENCES:

- 1 Desa UN. United nations department of economic and social affairs, population division. world population prospects: The 2015 revision, key findings and advance tables. In Technical Report: Working Paper No. ESA/P/WP 241 2015.
- 2 Alzheimer's Disease International. World Alzheimer Report 2019: attitudes to dementia. 2019;
- 3 Prince MJ. *World Alzheimer Report 2015: the Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends*. England: Alzheimer's Disease International, 2015.

- 4 Dementia. 2020. [Cited 1 Aug 2020]. Available from URL: <https://www.who.int/news-room/fact-sheets/detail/dementia>
- 5 Prince MJ. World Alzheimer Report 2016 - Improving healthcare for people living with dementia: Coverage, quality and costs now and in the future. 2016 [Cited 7 Jul 2020]. Available from URL: <https://www.alz.co.uk/research/world-report-2016>
- 6 Coster WJ. Making the best match: selecting outcome measures for clinical trials and outcome studies. *Am J Occup Ther* 2013; **67**: 162–170.
- 7 Stoner CR, Lakshminarayanan M, Durgante H, Spector A. Psychosocial interventions for dementia in low- and middle-income countries (LMICs): a systematic review of effectiveness and implementation readiness. *Aging Ment Health* 2019 Dec 9: 1–12.
- 8 Terwee C, Bot S, Boer M *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; **60**: 34–42.
- 9 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009 Jul 21; **6**: e1000097.
- 10 DAC List of ODA Recipients - OECD. 2020. [Cited 1 Aug 2020]. Available from URL: <http://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/daclist.htm>
- 11 Stansfeld J, Stoner CR, Wenborn J, Vernooij-Dassen M, Moniz-Cook E, Orrell M. Positive psychology outcome measures for family caregivers of people living with dementia: a systematic review. *Int Psychogeriatr* 2017; **29**: 1281–1296.
- 12 Stoner CR, Orrell M, Spector A. Review of positive psychology outcome measures for chronic illness, traumatic brain injury and older adults: adaptability in dementia? *Dement Geriatr Cogn Disord* 2015; **40**: 340–357.
- 13 Folstein MF, Folstein SE, McHugh PR. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975; **12**: 189–198.
- 14 Machado F, Nunes PV, Viola LF, Santos FS, Forlenza OV, Yassuda MS. Quality of life and Alzheimer's disease: influence of participation in a rehabilitation center. *Dement Neuropsychol* 2009; **3**: 241–247.
- 15 Niu Y-X, Tan J-P, Guan J-Q, Zhang Z-Q, Wang L-N. Cognitive stimulation therapy in the treatment of neuropsychiatric symptoms in Alzheimer's disease: a randomized controlled trial. *Clin Rehabil* 2010; **24**: 1102–1111.
- 16 Viola LF, Nunes PV, Yassuda MS *et al*. Effects of a multidisciplinary cognitive rehabilitation program for patients with mild Alzheimer's disease. *Clinics* 2011; **66**: 1395–1400.
- 17 Azcurra DJLS. A reminiscence program intervention to improve the quality of life of long-term care residents with Alzheimer's disease. A randomized controlled trial. *Rev Bras Psiquiatr* 2012; **34**: 422–433.
- 18 Kumar P, Tiwari SC, Goel A *et al*. Novel occupational therapy interventions may improve quality of life in older adults with dementia. *Int Arch Med* 2014; **7**: 26.
- 19 Lin Q, Cao Y, Gao J. The impacts of a GO-game (Chinese chess) intervention on Alzheimer disease in a northeast Chinese population. *Front Aging Neurosci* 2015; **7**: 163.
- 20 Santos GD, Nunes PV, Stella F *et al*. Multidisciplinary rehabilitation program: effects of a multimodal intervention for patients with Alzheimer's disease and cognitive impairment without dementia. *Arch Clin Psychiatry* 2015; **42**: 153–156.
- 21 Camargo CHF, Justus FF, Retzlaff G. The effectiveness of reality orientation in the treatment of Alzheimer's disease. *Am J Alzheimers Dis Other Demen* 2015; **30**: 527–532.
- 22 Asjret G, Kapucu S. The effect of reminiscence therapy on cognition, depression, and activities of daily living for patients with Alzheimer disease. *J Geriatr Psychiatry Neurol* 2015; **29**: 31–37.
- 23 Raghuraman S, Lakshminarayanan M, Vaitheswaran S, Rangaswamy T. Cognitive stimulation therapy for dementia: pilot studies of acceptability and feasibility of cultural adaptation for India. *Am J Geriatr Psychiatry* 2017; **25**: 1029–1032.
- 24 Paddick S-M, Mkenda S, Mbowe G *et al*. Cognitive stimulation therapy as a sustainable intervention for dementia in sub-Saharan Africa: feasibility and clinical efficacy using a stepped-wedge design. *Int Psychogeriatr* 2017; **29**: 979–989.
- 25 Li D-M, Li X-X. The effect of folk recreation program in improving symptoms: a study of Chinese elder dementia patients. *Int J Geriatr Psychiatry* 2017; **32**: 901–908.
- 26 de Oliveira AM, Radanovic M, Homem de Mello PC *et al*. An intervention to reduce neuropsychiatric symptoms and caregiver burden in dementia: preliminary results from a randomized trial of the tailored activity program-outpatient version. *Int J Geriatr Psychiatry* 2019; **34**: 1301–1307.
- 27 Lyu J, Zhang J, Mu H *et al*. The effects of music therapy on cognition, psychiatric symptoms, and activities of daily living in patients with Alzheimer's disease. *J Alzheimers Dis* 2018; **64**: 1347–1358.
- 28 Mkenda S, Olakehinde O, Mbowe G *et al*. Cognitive stimulation therapy as a low-resource intervention for dementia in sub-Saharan Africa (CST-SSA): adaptation for rural Tanzania and Nigeria. *Dementia (London)* 2018; **17**: 515–530.
- 29 Novelli MMPC, Machado SCB, Lima GB *et al*. Effects of the tailored activity program in Brazil (TAP-BR) for persons with dementia: a randomized pilot trial. *Alzheimer Dis Assoc Disord* 2018; **32**: 339–345.
- 30 Wang Z, Li Z, Xie J, Wang T, Yu C, An N. Music therapy improves cognitive function and behavior in patients with moderate Alzheimer's disease. *Int J Clin Exp Med* 2018; **11**: 4808–4814.
- 31 Chen K-L, Xu Y, Chu A-Q *et al*. Validation of the Chinese version of Montreal cognitive assessment basic for screening mild cognitive impairment. *J Am Geriatr Soc* 2016; **64**: e285–e290.
- 32 Paddick S-M, Kisoli A, Mkenda S *et al*. Adaptation and validation of the Alzheimer's Disease Assessment Scale - Cognitive (ADAS-cog) in a low-literacy setting in sub-Saharan Africa. *Acta Neuropsychiatr* 2017 Aug; **29**: 244–251.
- 33 Keskinoglu P, Ucku R, Yener G, Yaka E, Kurt P, Tunca Z. Reliability and validity of revised Turkish version of mini mental state examination (rMMSE-T) in community-dwelling educated and uneducated elderly. *Int J Geriatr Psychiatry* 2009 Nov; **24**: 1242–1250.
- 34 Flaks MK, Forlenza OV, Pereira FS, Viola LF, Yassuda MS. Short cognitive performance test: diagnostic accuracy and education bias in older Brazilian adults. *Arch Clin Neuropsychol* 2009 May; **24**: 301–306.
- 35 Bertolucci PHF, Okamoto IH, Brucki SMD, Siviero MO, Toniolo Neto J, Ramos LR. Applicability of the CERAD neuropsychological battery to Brazilian elderly. *Arq Neuropsiquiatr* 2001 Sep; **59**: 532–536.
- 36 Zhu X, Yi J, Yao S, Ryder AG, Taylor GJ, Bagby RM. Cross-cultural validation of a Chinese translation of the 20-item Toronto alexithymia scale. *Compr Psychiatry* 2007 Oct; **48**: 489–496.
- 37 Lourenço RA, Veras RP. Mini-Exame do Estado Mental: características psicométricas em idosos ambulatoriais. *Rev Saude Publica* 2006; **40**: 712–719.
- 38 Xu G, Meyer JS, Huang Y, Du F, Chowdhury M, Quach M. Adapting mini-mental state examination for dementia screening

- among illiterate or minimally educated elderly Chinese. *Int J Geriatr Psychiatry* 2003 Jul; **18**: 609–616.
- 39 Camozzato AL, Kochhann R, Simeoni C *et al.* Reliability of the Brazilian Portuguese version of the neuropsychiatric inventory (NPI) for patients with Alzheimer's disease and their caregivers. *Int Psychogeriatr* 2008; **20**: 383–393.
- 40 Stella F, Forlenza OV, Laks J *et al.* The Brazilian version of the neuropsychiatric inventory-clinician rating (NPI-C): reliability and validity in dementia. *Int Psychogeriatr* 2013; **25**: 1503–1511.
- 41 Wang T, Xiao S, Li X *et al.* Reliability and validity of the Chinese version of the neuropsychiatric inventory in mainland China. *Int J Geriatr Psychiatry* 2012; **27**: 539–544.
- 42 Leung CM, Wing YK, Kwong PK, Lo A, Shum K. Validation of the Chinese-Cantonese version of the hospital anxiety and depression scale and comparison with the Hamilton rating scale of depression. *Acta Psychiatrica Scand* 1999; **100**: 456–461.
- 43 Castelo MS, Coelho-Filho JM, Carvalho AF *et al.* Validity of the Brazilian version of the geriatric depression scale (GDS) among primary care patients. *Int Psychogeriatr* 2010; **22**: 109–113.
- 44 Paradelo EMP, Lourenço RA, Veras RP. Validation of geriatric depression scale in a general outpatient clinic. *Rev Saude Publica* 2005; **39**: 918–923.
- 45 Li L, Wang HM, Shen Y. Chinese SF-36 health survey: translation, cultural adaptation, validation, and normalisation. *J Epidemiol Community Health* 2003; **57**: 259–263.
- 46 Group TW. Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychol Med* 1998; **28**: 551–558.
- 47 Novelli MMPC, Nitrini R, Caramelli P. Validation of the Brazilian version of the quality of life scale for patients with Alzheimer's disease and their caregivers (QOL-AD). *Aging Ment Health* 2010; **14**: 624–631.
- 48 Saxena S, Chandiramani K, Bhargava R. WHOQOL-Hindi: a questionnaire for assessing quality of life in health care settings in India. World Health Organization quality of life. *Natl Med J India* 1998; **11**: 160–165.
- 49 Julayanont P, Tangwongchai S, Hemrungronj S *et al.* The Montreal cognitive assessment-basic: a screening tool for mild cognitive impairment in illiterate and low-educated elderly adults. *J Am Geriatr Soc* 2015; **63**: 2550–2554.
- 50 Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's disease assessment scale-cognitive subscale (ADAS-cog): modifications and responsiveness in pre-dementia populations. A narrative review. *J Alzheimers Dis* 2018; **63**: 423–444.
- 51 Leheld H, Erzigkeit H. The SKT-a short cognitive performance test for assessing deficits of memory and attention. *Int Psychogeriatr* 1997; **9**: 115–121.
- 52 Bagby RM, Parker JDA, Taylor GJ. The twenty-item Toronto alexithymia scale—I. Item selection and cross-validation of the factor structure. *J Psychosom Res* 1994; **38**: 23–32.
- 53 Fillenbaum GG, van Belle G, Morris JC *et al.* CERAD (consortium to establish a registry for Alzheimer's disease) the first 20 years. *Alzheimers Dement* 2008; **4**: 96–109.
- 54 Tombaugh TN, McIntyre NJ. The mini-mental state examination: a comprehensive review. *J Am Geriatr Soc* 1992; **40**: 922–935.
- 55 Bertolucci PH, Brucki SM, Campacci SR, Juliano Y. The mini-mental state examination in a general population: impact of educational status. *Arq Neuropsiquiatr* 1994; **52**: 1–7.
- 56 Almeida OP. Mini mental state examination and the diagnosis of dementia in Brazil. *Arq Neuropsiquiatr* 1998; **56**: 605–612.
- 57 Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia. *Neurology* 1994; **44**: 2308–2314.
- 58 Yesavage JA, Brink TL, Rose TL *et al.* Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res* 1982; **17**: 37–49.
- 59 Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473–483.
- 60 Logsdon RG, Gibbons LE, McCurry SM, Teri L. Assessing quality of life in older adults with cognitive impairment. *Psychosom Med* 2002; **64**: 510–519.
- 61 Kuyken W, Orley J, Hudelson P, Sartorius N. Quality of life assessment across cultures. *Int J Ment Health* 1994; **23**: 5–27.
- 62 Group TW. The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med* 1998; **46**: 1569–1585.
- 63 Brislin RW. Comparative research methodology: cross-cultural studies. *Int J Psychol* 1976; **11**: 215–229.
- 64 Bracken BA, Barona A. State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychol Int* 1991; **12**: 119–132.
- 65 Solano-Flores G, Nelson-Barber S. On the cultural validity of science assessments. *J Res Sci Teach* 2001; **38**: 553–573.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website: <http://onlinelibrary.wiley.com/doi/supinfo>.

Appendix S1: Supplementary Information.