Electronic supplementary material:
The online version of this article contains supplementary material.

journal of global health

Cite as: Zhao Y, Summers R, Gathara D, English M. Conducting cross-cultural, multi-lingual or multi-country scale development and validation in healthcare research: A 10-step framework based on a scoping review. J Glob Health 2024;14:04151.

Conducting cross-cultural, multi-lingual or multi-country scale development and validation in health care research: A 10-step framework based on a scoping review

Yingxi Zhao[1], Richard Summers[2], David Gathara[3,4], Mike English[1,3]

[1]Nuffield Department of Medicine Centre for Global Health Research, University of Oxford, Oxford, UK

[2]School of Social Policy, University of Birmingham, Birmingham, UK

[3]KEMRI-Wellcome Trust Research Programme, Nairobi, Kenya

[4]Centre for Maternal, Adolescent, Reproductive, and Child Health (MARCH), London School of Hygiene and Tropical Medicine, London, UK

Background Valid, reliable and cross-cultural equivalent scales and measurement instruments that enable comparisons across diverse populations in different countries are important for global health research and practice. We developed a 10-step framework through a scoping review of the common strategies and techniques used for scale development and validation in a cross-cultural, multi-lingual, or multi-country setting, especially in health care research.

Methods We searched MEDLINE, Embase, and PsycINFO for peer-reviewed studies that collected data from two or more countries or in two or more languages at any stages of scale development or validation and published between 2010–22. We categorised the techniques into three commonly used scale development and validation stages (item generation, scale development, and scale evaluation) as well as during the translation stage. We described the most commonly used techniques at each stage.

Results We identified 141 studies that were included in the analysis. We summarised 14 common techniques and strategies, including focus groups or interviews with diverse target populations, and involvement of measurement experts and linguists for item content validity expert panel at the item generation stage; back-and-forth translation, collaborative team approach for the translation stage; cognitive interviews and different recruitment strategies and incentives in different settings for scale development stage; and three approaches for measurement invariance (multigroup confirmatory factor analysis, differential item functioning and multiple indicator multiple causes) for scale evaluation stage.

Conclusions We provided a 10-step framework for cross-cultural, multi-lingual or multi-country scale development and validation based on these techniques and strategies. More research and synthesis are needed to make scale development more culturally competent and enable scale application to better meet local health and development needs.

Correspondence to:
Yingxi Zhao
Nuffield Department of Medicine
Centre for Global Health Research,
University of Oxford
S Parks Road, Oxford
UK
yingxi.zhao@ndm.ox.ac.uk

Scales measure latent constructs and specifically 'behaviours, attitudes, and hypothetical scenarios we expect to exist as a result of our theoretical understanding of the world, but cannot assess directly' [1]. Developing a new scale could help us measure a more specific behaviour and experience. However, scale development and validation processes can be long and complex, usually involving three stages: item development, where an initial item pool is produced via deductive, inductive, or combined approaches; scale development, where individual items are constructed into harmonious constructs; and scale evaluation, where the validity and reliability of the scale are tested [2–4].

Global health often needs valid, reliable, and universally applicable tools and measures that enable comparisons across culturally diverse populations in different countries [5–7]. However, measurement instruments developed in one context are often translated into other languages or taken into other contexts without proper equivalence assessment [8]. When evaluated, the psychometric properties in new settings often suggest poor performances [9]. Poor translation quality can contribute to this. Still, cultural differences exist in the conceptual and operational definitions of many behaviours and experiences being measured, which can lead to measurement inequivalence [10,11]. If the aim is to produce comparable and generalisable evidence across contexts, we need to consider how to develop a scale that performs well in different cultures, countries and languages right from the start.

Here, we extend Boateng et al. [1] framework for scale development and validation to encompass such cross-context concerns through a scoping review. Our research question was, 'What are common strategies and techniques that have been used for scale development and validation in a cross-cultural, multi-lingual or multi-country setting, especially in health care research?' We then extended the Boateng et al. framework to 10 steps to incorporate these considerations. We hope this 10-step framework may help researchers who aim to develop a new scale for use in different settings or those who aim to adapt an existing scale to a new language or country.

METHODS

We followed the five steps of the Arksey and O'Malley method [12] for scoping reviews to identify studies that conducted scale development and/or validation in a cross-cultural, multi-lingual, or multi-country setting.

Search strategy and screening

We conducted a systematic search using MEDLINE, Embase, and PsycINFO to obtain relevant articles in health care research, using search terms such as multi-country, cross-cultural, and multigroup (Table S1 in the **Online Supplementary Document**). We included empirical studies published between 2010–22 in English due to time and resource constraints. We included studies if they reported data collection from two or more countries or in two or more languages at any stages of scale development or validation, for example, if a scale was first developed in English and then validated in several other languages or countries. We excluded studies that validated a previously developed scale in a new sample but only in one single language or country; we also excluded studies that focused on measurement invariance of different population groups with different culture or socio-demographic characteristics within one country, despite that some of our findings would still be of relevance for those studies. After deduplication, one reviewer conducted two stages of title/abstract and full-text screening in Abstrackr (Brown University Center for Evidence Synthesis in Health, Providence, Rhode Island, USA) [13] and Microsoft Excel, version 16 (Microsoft Corporation, Redmond, Washington, USA), respectively. Uncertainty was resolved through discussions with other co-authors.

Data charting, collation and reporting

We further charted data from the articles that were included and entered them into a Microsoft Excel spreadsheet. We charted the following data: title, authors, survey sample size, survey country number, survey language number, and technique for cross-cultural, multi-lingual, or multi-country scale development and validation. We categorised the technique into three stages, as listed in Boateng et al. [1], which included item development, scale development, and scale evaluation. We also charted the strategy used for translation when more than one language was used. When available, we extracted these data by looking into the methods, results, tables/figures, and questionnaire appendices. For reporting of our findings, we described the most commonly used techniques at each stage, i.e. used more than once in our included studies, and where relevant, traced back to the cited methodology papers when the authors referenced specific techniques they used. We organised these techniques and strategies and extended the framework by Boateng et al. [1] into a 10-step framework for cross-cultural, multi-lingual, or multi-country scale development and validation.

RESULTS

Of the 1985 citations identified after deduplication, 141 met inclusion criteria after the full-text review. The included studies had an average 5220 sample size in their survey administration stage, 134 included more than one country in their sample, and 102 used more than one language (Figure S2 and Table S3 in the **Online Supplementary Document**).

Table 1 summarises the technique or strategy used by more than one study at each scale development and validation stage. Figure 1 further extends the scale development and validation framework by Boateng et al. [1] to incorporate these techniques and strategies for cross-cultural, multilingual, or multi-country settings. In the following section, we explain each technique and provide examples of their use. Readers could also refer to the references listed in **Table 1** for more detail.

Table 1. Commonly used techniques and strategies for cross-cultural, multi-lingual or multi-country scale development and validation

Technique and strategy	Description	Publications that reported its use (n)	Examples
Item development stage			
Literature based reviews to capture existing tools or constructs	Literature search in databases to capture relevant validated instruments across different countries and settings.	8	Bloemeke et al. [14], Perkmen et al. [15]
Individual concept elicitation or in-depth interviews with target population	Exploratory interviews in different countries and settings.	6	Kerrigan et al. [16], Chen et al. [17]
Focus group discussions with target population	Focus group discussions in different countries and settings to allow respondents to explore and clarify individual and shared perspectives.	9	Aizpitarte et al. [18], Luquiens et al. [19]
Expert panel or consensus group	Inputs from subject experts, measurement experts, and linguists to provide or review items have cross-cultural validity and will be easily translatable.	8	Nackers et al. [20], Abraham et al. [21]
Translation stage			
Back-and-forth translation	First translate from source language to target language, then back translate to source language by a second translator, and lastly compare and resolve inconsistency.	63	Mezquita et al. [22], Roberts et al. [23]
Expert review	Translated items reviewed by bilingual subject experts, measurement experts and linguists.	11	Wilson et al. [24], Vogel et al. [25], Vaingankar et al. [26]
Collaborative and iterative translation	A collaborative approach through parallel or double translation, pretesting, and revise. No back-translation.	3	Hakim and Liu [27], Sproesser et al. [28]
Scale development stage			
Cognitive debriefing or interview	Pilot participants asked their understanding of each instruction, item and response options to evaluate the interpretation and acceptability; sometimes followed by questions on general perception of the draft scale.	8	Luquiens et al. [19], McCoy et al. [29]
Different ways of recruitment	Survey administration including recruitment strategy and motivation should be adapted to local context and logistics feasibility.	2	Korf et al. [30], O'Brien et al. [31]
Separate reliability test in each sample	Cronbach's α-based reliability analysis and item-to-scale correlational analyses in each sample.	3	Littrell et al. [32], Whiting-Collins et al. [33]
Separate factor analysis in each sample	Separate exploratory and/or confirmatory factor analysis in each sample to understand the factor structure patterns between different samples. For confirmatory factor analysis, commonly reported model fit indices are CFI>0.90 or CFI>0.95, RMSEA<0.08, SRMR<0.08, and Tucker–Lewis index >0.90.	30	Encantado et al. [34], Stevelink et al. [35]
Scale evaluation stage			
MGCFA	Technique under classical test theory. Three most commonly used measurement invariance are configural invariance (same number of factors and pattern of loading), metric invariance (factor loading across groups), scalar invariance (same item intercepts); and commonly reported invariance indices are ΔCFI (commonly below 0.01), ΔRMSEA (commonly below 0.015), ΔSRMR (commonly for metric level below 0.03) whereas others also used the χ^2 χ^2 test.	84	Datu et al. [36], Lopez-Fernandez et al. [37]
Rasch analysis and DIF	Technique under item response theory. DIF is useful to discover with item function differently across sub-groups, through having each item as the dependent variable and the total score and sub-group as well as their interaction as covariates, and significant changes in coefficient of determination indicate response to the examined item is affected by language or country.	19	Erhart et al. [38], Lau et al. [39], Geyh et al. [40]
MIMIC	Confirmatory factor analysis with sub-group (country or language) as covariate, and significant direct effect on model indices suggest inequivalence.	3	Boudjemadi et al. [41], Pendergast et al. [42]

CFI – comparative fit Index, DIF – differential item functioning, MGCFA – multi-group confirmatory factor analysis, MIMIC – multiple indicator multiple causes, RMSEA – root-mean-square error of approximation, SRMR – standardised root mean square residual, ΔCFI – change in comparative fix index, ΔRMSEA – change in root-mean-square error of approximation, ΔSRMR – change in standardised root mean square residual

Item development stage (steps one and two)

For step one, which is generating the domain and individual items, techniques that considered cross-cultural, multi-lingual, or multi-country equivalence included literature reviews to capture constructs of interest or existing tools that measure similar phenomena or focus groups and individual interviews to generate items inductively with the target population, which were usually conducted in multi-country settings to generate items that could be broadly generalisable. For example, Chen et al. [17], when developing a Family Role Performance Scale, conducted interviews with 15 Israeli and 11 United States respondents with a diverse array of family structures. They started with general questions about family roles and then moved on to specific questions, using a funnel approach. The interview transcripts were translated and coded for concepts of interest, further informing their categorisation of 17 items.

At the content validity step, which focused on whether the items adequately measure the domain (step two), some reported using expert panels or consensus groups to review their items, and subject experts were deemed important in ensuring all relevant topics were covered in the items. In Abraham et al. [21] development of an oestrogen plus progestin therapies-related breast symptoms questionnaire, after items were generated from concept elicitation interviews with the target population, a group of measurement and clinical experts and linguists reviewed their items. The inclusion of linguists specifically ensured that items had cross-cultural validity, could be easily translated into the four countries of study, and were conceptually equivalent in other potential languages.

Other approaches at this stage reported in literature included: 1) Delphi technique by Michaud et al. [43] in their item generation stage, where participants were asked to discuss and rate different concepts of interest, 2) Benschop et al. [44] when designing a new tool for understanding new psychoactive substance use in six European countries, conducted extensive literature-based review, country reports and expert consultation to understand the policy and context of substance use, which informed their item generation, 3) O'Brien et al. [31] also reported that their research team had members with diverse culture, different ages and experiences which helped to reduce biases.

Translation (step three)

Out of 118 studies that reported using more than one language for their measurement scale, 80 reported their translation strategy. While very few studies used one-way translation (from the source language to the target language), most used back-and-forth translation techniques, and some also specifically referenced the Brislin procedure for back-and-forth translation, which included three steps of translating into the target language, back-translating into source language by another translator and then resolving inconsistency through discussion [45]. In other studies, researchers reported the use of expert reviews after translation. For example, after translating the draft scale to Mandarin, Wilson et al. [24] invited psychologists with measurement and construct expertise to ensure lexical and construct equivalences were maintained. Some studies also used collaborative and iterative translation, or the committee method, which is considered an alternative to back-translation to ensure better conceptual equivalence [27,28,46]. Several other studies translated their instruments following specific organisational guidelines [47,48].

Scale development stage (steps four to seven)

In the scale development stage, items are selected and translated, and data are collected. However, there are other factors to consider in the data collection process. For pre-testing (step four), several studies conducted cognitive interviews or cognitive debriefings to ensure their scale's face and content validity. For example, Luquiens et al. [19] asked patients in their study countries to complete their draft scale. Then, they asked them questions to assess their understanding of each instruction, item and response option. This led to the removal of 14 items that were considered redundant, ambiguous, or difficult to understand, as well as eight revised items.

For scale administration (step five), two studies reported different ways of recruitment. In O'Brien et al. [31] study, where they recruited South Korean and White mothers, the former received small-amount gift cards for a cup of coffee, whereas the latter had the opportunity to enter a lottery to win a larger prize due to cultural differences regarding incentives. Benschop et al. also used different strategies to recruit new psychoactive substance users via nightclubs, drug services and Facebook groups based on differences in country settings [30,44]. These considerations could also be relevant to earlier stages of recruitment for focus groups or cognitive interviews. Additionally, there are other considerations for sample size, both linked with overall scale development and multi-group confirmatory factor analysis (MGCFA), discussed below at the scale evaluation stage (step eight).

For the psychometric analysis of the scale development stage (steps six and seven), several studies highlighted the first use of internal reliability and item correlational analysis in each sample separately. More commonly, studies reported conducting separate exploratory and/or confirmatory factor analyses in each sample to understand the factor structure patterns between different samples. For example, Bothe et al. [49] ran a confirmatory factor analysis on their five theory-based factors and 19 representing items in four separate samples, all of which had acceptable model fits. This step is usually required before running MGCFA in the evaluation stage. In Stevelink et al. [35], an exploratory factor analysis of participants from six countries was performed per database, suggesting two different factor structures (one-factor vs two-factor model). They further conducted MGCFA using both structures, and the two-factor model suggested a better overall fit. For the studies that conducted confirmatory factor analysis and provided model fit indices, four indices are commonly reported – comparative fit index (CFI>0.90 or CFI>0.95), root-mean-square error of approximation (RMSEA<0.08), standardised root mean square residual (SRMR<0.08) and Tucker–Lewis index (TLI>0.90) (Table S4 in the **Online Supplementary Document**).

Scale evaluation stage (steps eight to 10)

All the studies that reported different techniques used at the evaluation stage focused on cross-country or cross-language measurement invariance, i.e. whether the psychometric properties are generalisable across different sub-groups (step eight). The most commonly used technique is MGCFA. MGCFA usually examines three types of measurement invariance, i.e. configural invariance (same number of factors and pattern of loading), metric invariance (factor loading across groups), and scalar invariance (same item intercepts). However, others also examined strict invariance (invariance of the item residuals) [36,37]. For studies that reported invariance indices, three indices are commonly used: change (Δ) in CFI<0.01, ΔRMSEA<0.015, and ΔSRMR<0.03 (commonly for the metric level and for scalar level there is no consensus). Chen [50] and Cheung and Rensvold [51] are the most commonly cited references for these cut-off values. Several other studies also used $\chi 2$ test statistic or alignment methods. For the 83 studies that provided sample size by group for MGCFA, 70 had more than 150 participants per group, and 63 had more than 200 per group (Table S4 in the **Online Supplementary Document**). For example, in Lopez-Fernandez et al. [37], cross-cultural validation of the Compulsive Internet Use Scale across eight languages, configural invariance was supported. However, ΔCFI and ΔRMSEA between the metric and scalar model exceeded cut-off thresholds. Therefore the factor structure and loading were invariant between eight languages. Still, the latent factor means and residuals could be different.

While MGCFA fits under the classical test theory, others used Rasch analysis and differential item functioning (DIF), which are item-response theory techniques. The Rasch model is most useful in evaluating individual items' functioning as it estimates both scale and item-level fit indices. Within Rasch analysis, DIF is commonly used to examine cross-cultural validity across subgroups. Usually, each item is examined as a dependent variable with the total score and sub-group (language or country) and their interaction as covariates. Significant changes in the coefficient of determination suggest significant DIF, i.e. response to the examined item is affected by language or country which could help select items that should be re-considered for cross-cultural invariance. For example, Geyh et al. [40] conducted DIF analyses using data from four countries. They highlighted two items with DIF in the Satisfaction with Life Scale and, through Tukey-Cramer post-hoc tests, highlighted that data from Israel showed the most frequent differences from the other countries.

Lastly, several studies also reported using multiple indicator multiple causes (MIMIC), a variant of confirmatory factor analysis that further incorporated covariates. One paper also claimed that MIMIC is more appropriate for analyses across many groups [42]. Boudjemadi et al. [41], in their testing of multigroup invariance across countries, added a country covariate to their MGCFA model. They observed a significant direct effect of the country covariate on the model fit indices and concluded that factor means behave differently across countries. Besides these three commonly used techniques, other papers also reported using approaches such as the sequential constraint imposition [52] and Satorra-Bentler $\chi 2$ difference test [27].

We did not identify specific techniques used for steps nine and 10, which are the reliability and validity tests.

DISCUSSION

Conducting cross-cultural, multi-lingual or multi-country scale development and validation represents unique logistics and analytical challenges. In this scoping review, through reviewing and analysing 141 published cross-setting, scale development, and validation articles, we summarised the common techniques

and strategies used to ensure that the scale has cross-lingual or cross-country equivalence. While a protocol has been published on adapting measurement scales to a new context [11], our review further extends this to developing new scales across contexts often needed for global health research and practice.

In **Figure 1**, we have provided a 10-step framework of cross-cultural, multi-lingual or multi-country scale development and validation extending the original nine-step framework by Boateng et al. [1]. This framework is based on published health care research papers but is also relevant to other disciplines such as psychology, management and education. Some steps apply to broader global health research and practices, such as conducting multi-country surveys. While all the techniques listed could further improve scale development and validation rigour, researchers might be constrained by resources such as time and funding and have to decide which steps and techniques to prioritise. In the following sections, we highlight some of the key considerations for researchers.

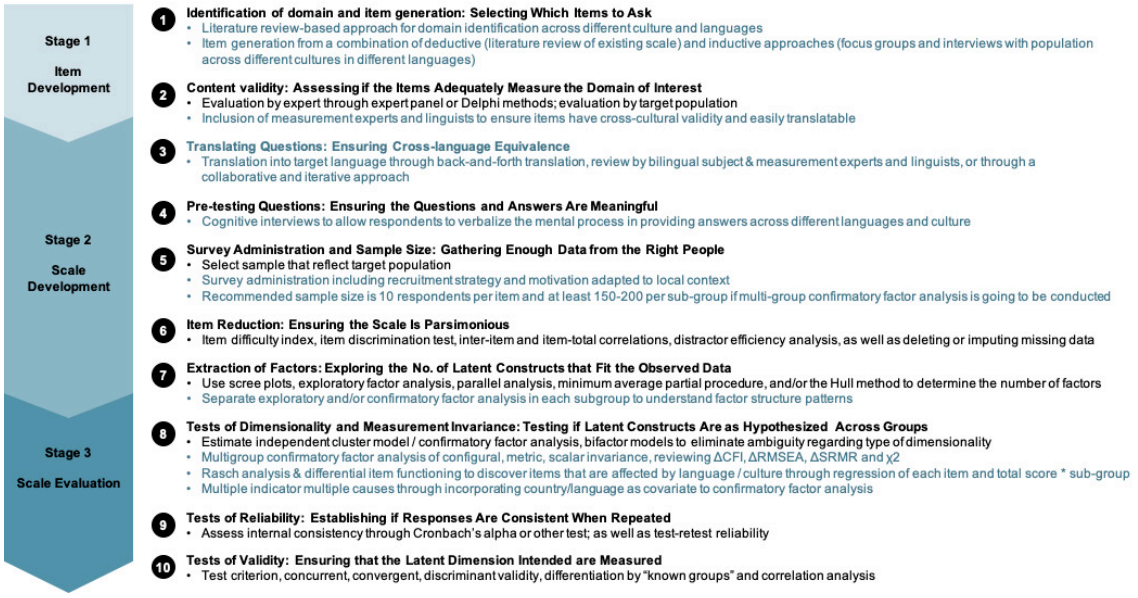

Figure 1. A 10-step cross-cultural, multi-lingual or multi-country scale development and validation framework. Adapted from the nine-step scale development and validation framework by Boateng et al. with added recommendations for cross-cultural, multi-lingual or multi-country scale development and validation in blue.

Prior to scale development for cross-cultural settings, there is a need to ensure that the measure or phenomenon truly exists as a culturally independent construct and that the construct itself can be robustly measured [53]. This could be achieved by techniques listed in steps one and two (**Figure 1**), e.g. literature-based review and qualitative work with the target population across settings. Having a diverse research team [31] and understanding the policy context through examining policy documents and conducting expert consultations are also useful [44]. Occasionally, some measures, constructs or items are not universally relevant but of great importance to certain settings, in those cases slightly different versions of a scale could be developed for different settings.

Commonly draft scales need to be translated into additional target languages (as shown in step three). While most papers reported back-and-forth translation, how developers handled the inconsistency between the source and back-translated versions is unclear. We agree with Douglas et al. that reliance on back-translation could be problematic as bilingual translators could make sense of a poorly written target translation. There are subtle nuances in the use of languages and idioms [54]; therefore, using a team-based approach and the involvement of measurement experts and linguists could help discover these issues [54,55]. We also highlight the importance of conducting cognitive interviews in pilot testing (step four). Cognitive interviews could significantly improve the validity of surveys in global health settings. This is because during the actual survey scale, instructions and items could be understood in unexpected ways, and item responses could be considered inappropriate, compromising the study quality [56]. Unfortunately, only eight studies in our 141 included articles mentioned cognitive interviews in different settings, which might be related to the lack of emphasis on this methodology in global health surveys that rely more on quantitative approaches. We strongly recommend using cognitive interviews for cross-setting scale development, and more details on conducting cognitive interviews can be found in Scott et al. [56]

Measurement invariance is often the centrepiece for cross-setting scale development and validation, as most papers reported using one statistical approach for invariance analysis (step eight). Despite many critiques, MGCFA is still the most commonly used and 'the most powerful and versatile' approach [57,58]; therefore, we recommend using MGCFA if the aim is to produce a universally applicable scale for cross-cultural settings. We acknowledge that there are different invariance indices used in our review, including ΔCFI, ΔRMSEA, and ΔSRMR, as well as slight inconsistency on cut-offs (especially whether, for example, ΔCFI should be <0.01 or ≤0.01) (Table S4 in the **Online Supplementary Document**), and some others also used the more restrictive change in $\chi 2$ test. We recommend reporting three indices. As a relaxed fit, ΔCFI≤0.01, ΔRMSEA≤0.015, and ΔSRMR≤0.03 at the metric level. Aside from MGCFA, if the aim is to identify specific non-equivalent items, researchers should use DIF. If there are many sub-groups in measurement invariance analysis, researchers should consider MIMIC [42].

Regarding sample sizes for step five on survey administration, various rules have been suggested for determining the sample size for the questionnaire survey and scale development. The general rule of thumb for scale development is 10 participants per item [1,59]. For multi-lingual or multi-country scale development where MGCFA is going to be conducted, researchers need to be mindful of whether the sample size in the sub-group is sufficient to obtain an accurate factor solution. Guadagnoli suggested that the actual number is dependent on the number of items per scale construct and component saturation (the magnitude of component loading), and if there are 10 or more items representing each construct, a sample size of 150 should be sufficient [60], and Hair et al. recommended a sample of 200 [61]. Most of our included studies had more than 150 or 200 samples in their smallest subgroup. Therefore, we recommend that researchers use both criteria when determining sample size if MGCFA is to be performed (10 participants per item for the overall sample, and 150–200 samples per sub-group). There are no established guidelines on the sample size required for Rasch and DIF analyses, and 100–200 per subgroup is commonly recommended [62–64].

Developing and administering the scale in multiple countries and languages could also bring logistic and sometimes ethical challenges. Challenges to find and acknowledge the roles of translators, interviewers and survey staff [65], standardisation of recruitment strategies [66], as well as the ethical challenges such as the power asymmetries between the scale development team and frontline staff and whether the scale being developed is relevant and prioritised by local population [67,68] should all be taken into consideration. Our reviewed papers also provided examples of how to recruit in diverse settings and what motivation incentives should be in place [30,31,44]. Researchers could also refer to the broad literature on cross-cultural research when conducting multi-lingual and multi-country scale development [65-67,69–71].

Several limitations should be considered for this review. First, screening and data charting are conducted by only one reviewer. While there could be bias in the selection and extraction process, the reviewer has rich experiences in different review types; thus, the risk of bias is relatively lower. Second, as this is a scoping review, we did not aim to systematically assess the quality of included studies but only reported on what techniques and strategies have been used. Lastly, it should be noted that our inclusion and exclusion criteria might have led to biases. For example, we included studies that included two or more countries or two or more languages in their scale development, and we excluded studies that focused on measurement invariance of different population groups with different cultural or socio-demographic characteristics within one country; inclusion of these studies might have identified additional techniques. An example of examining measurement invariance across culture and socio-demographics could be found in Dong and Dumas [72]. We also only focused on papers published in health care journals in English due to time and resource limitations. While the authors include researchers from different disciplines (health care, medicine, nursing, psychology) and three authors also come from global health backgrounds, which improved the review's relevance and comprehensiveness, we acknowledge that further extension to related fields such as education and psychology journals, non-English language papers, and the inclusion of papers that investigated cultural differences within one country and language could help strengthen our recommendations.

CONCLUSIONS

Scale development and validation in cross-cultural, multi-lingual or multi-country settings is important for global health research and practice, but the process could be challenging. Reviewing current techniques and strategies in published scale development papers, we summarised key recommendations at item generation, translation, scale development, and scale evaluation stages. We produced a newer, expanded 10-step scale development and validation framework (**Figure 1**). As scale development is complicated and requires multiple iterations, and universal and cross-cultural equivalence is always more a goal than a reality, this

review should be considered a 'jumping off point' for anyone interested. More research and synthesis are needed to make scale development more culturally competent and enable scale application to better meet local health and development needs.

Funding: This work is supported by an Africa Oxford travel grant (AfOx-209). YZ is supported by the University of Oxford Clarendon Fund Scholarship, an Oxford Travel Abroad Bursary, and a Keble Association grant. This research was funded in whole or in part by the Wellcome Trust (207522), awarded to ME as a Wellcome Trust Senior Research Fellowship. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Authorship contributions: YZ and ME designed the study. YZ led the study selection, data charting, and collation process and wrote the first draft of the manuscript. RS, DG, and ME provided critical feedback on the first draft. All authors read and approved the final manuscript.

Disclosure of interest: The authors completed the ICMJE Disclosure of Interest Form (available upon request from the corresponding author) and disclose no relevant interests.

Additional material
Online supplementary document

1 Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. Front Public Health. 2018;6:149. Medline:29942800 doi:10.3389/fpubh.2018.00149

2 Carpenter S. Ten Steps in Scale Development and Reporting: A Guide for Researchers. Commun Methods Meas. 2018;12:25–44. doi:10.1080/19312458.2017.1396583

3 Hinkin TR. A Review of Scale Development Practices in the Study of Organizations. J Manage. 1995;21:967–88. doi:10.1177/014920639502100509

4 Morgado FFR, Meireles JFF, Neves CM, Amaral ACS, Ferreira MEC. Scale development: ten main limitations and recommendations to improve future research practices. Psicol Reflex Crit. 2017;30:3. Medline:32025957 doi:10.1186/s41155-016-0057-1

5 World Health Organization. Quality of Life assessment: position paper from the World Health Organization. Soc Sci Med. 1995;41:1403–9. Medline:8560308 doi:10.1016/0277-9536(95)00112-K

6 Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). J Public Health (Oxf). 2006;14:66–70. doi:10.1007/s10389-006-0024-x

7 Sanchez-Niubo A, Forero CG, Wu Y-T, Giné-Vázquez I, Prina M, De La Fuente J, et al. Development of a common scale for measuring healthy ageing across the world: results from the ATHLOS consortium. Int J Epidemiol. 2021;50:880–92. Medline:33274372 doi:10.1093/ije/dyaa236

8 Ford JB, Merchant A, Bartier A-L, Friedman M. The cross-cultural scale development process: The case of brand-evoked nostalgia in Belgium and the United States. J Bus Res. 2018;83:19–29. doi:10.1016/j.jbusres.2017.09.049

9 Clinton-McHarg T, Yoong SL, Tzelepis F, Regan T, Fielding A, Skelton E, et al. Psychometric properties of implementation measures for public health and community settings and mapping of constructs against the Consolidated Framework for Implementation Research: a systematic review. Implement Sci. 2016;11:148. Medline:27821146 doi:10.1186/s13012-016-0512-5

10 Mikkelsen KS, Schuster C, Meyer-Sahling J-H. A cross-cultural basis for public service? Public service motivation measurement invariance in an original survey of 23,000 public servants in ten countries and four world regions. Int Public Manage J. 2021;24:739–61. doi:10.1080/10967494.2020.1809580

11 Ambuehl B, Inauen J. Contextualized Measurement Scale Adaptation: A 4-Step Tutorial for Health Psychology Research. Int J Environ Res Public Health. 2022;19:12775. Medline:36232077 doi:10.3390/ijerph191912775

12 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 2005;8:19–32. doi:10.1080/1364557032000119616

13 Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. New York, USA: Association for Computing Machinery; 2012. p. 819–24.

14 Bloemeke J, Sommer R, Witt S, Bullinger M, Nordon C, Badia FJ, et al. Cross-cultural selection and validation of instruments to assess patient-reported outcomes in children and adolescents with achondroplasia. Qual Life Res. 2019;28:2553-63. Medline:31093848 doi:10.1007/s11136-019-02210-z

15 Perkmen S, Toy S, Caracuel A, Shelley M. Cross-cultural search for Big Five: development of a scale to compare personality traits of pre-service elementary school teachers in Turkey and Spain. Asia Pac Educ Rev. 2018;19:459–68. doi:10.1007/s12564-018-9549-2

16 Kerrigan D, Karver TS, Barrington C, Davis W, Donastorg Y, Perez M, et al. Development of the Experiences of Sex Work Stigma Scale Using Item Response Theory: Implications for Research on the Social Determinants of HIV. AIDS Behav. 2021;25:175–188. Medline:33730252 doi:10.1007/s10461-021-03211-1

17 Chen Y-P, Shaffer M, Westman M, Chen S, Lazarova M, Reiche S. Family role performance: Scale development and validation. Appl Psychol. 2014;63:190–218. doi:10.1111/apps.12005

18 Aizpitarte A, Alonso-Arbiol I, Van de Vijver FJ, Perdomo MC, Galvez-Sobral JA, Garcia-Lopez E. Development of a dating violence assessment tool for late adolescence across three countries: The Violence in Adolescents' Dating Relationships Inventory (VADRI). J Interpers Violence. 2017;32:2626–46. Medline:26160857 doi:10.1177/0886260515593543

19 Luquiens A, Whalley D, Crawford SR, Laramée P, Doward L, Price M, et al. Development of the Alcohol Quality of Life Scale (AQoLS): a new patient-reported outcome measure to assess health-related quality of life in alcohol use disorder. Qual Life Res. 2015;24:1471–81. Medline:25407634 doi:10.1007/s11136-014-0865-7

20 Nackers F, Roederer T, Marquer C, Ashaba S, Maling S, Mwanga-Amumpaire J, et al. A screening tool for psychological difficulties in children aged 6 to 36 months: cross-cultural validation in Kenya, Cambodia and Uganda. BMC Pediatr. 2019;19:108. Medline:30979364 doi:10.1186/s12887-019-1461-3

21 Abraham L, Humphrey L, Arbuckle R, Dennerstein L, Simon JA, Mirkin S, et al. Qualitative cross-cultural exploration of breast symptoms and impacts associated with hormonal treatments for menopausal symptoms to inform the development of new patient-reported measurement tools. Maturitas. 2015;80:273–81. Medline:25542407 doi:10.1016/j.maturitas.2014.11.019

22 Mezquita L, Bravo AJ, Pilatti A, Ortet G, Ibáñez MI. Team C-CAS. Preliminary validity and reliability evidence of the Brief Antisocial Behavior Scale (B-ABS) in young adults from four countries. PLoS One. 2021;16:e0247528. Medline:33617586 doi:10.1371/journal.pone.0247528

23 Roberts ME, Wagner L, Zorjan S, Nèmeth E, van Toor D, Czaplinski M. Testing the Situationism Scale in Europe: scale validation, self-regulation and regional differences. Int J Psychol. 2017;52:264–72. Medline:28703327 doi:10.1002/ijop.12211

24 Wilson CA, Plouffe RA, Saklofske DH, Yan G, Nordstokke DW, Prince-Embury S, et al. A cross-cultural validation of the resiliency scale for young adults in Canada and China. PsyCh J. 2019;8:240–51. Medline:30548571 doi:10.1002/pchj.256

25 Vogel DL, Armstrong PI, Tsai P-C, Wade NG, Hammer JH, Efstathiou G, et al. Cross-cultural validity of the Self-Stigma of Seeking Help (SSOSH) scale: Examination across six nations. J Couns Psychol. 2013;60:303. Medline:23458605 doi:10.1037/a0032055

26 Vaingankar JA, Abdin E, Chong SA, Sambasivam R, Shafie S, Ong HL, et al. Development of the Chinese, Malay and Tamil translations of the positive mental health instrument: cross-cultural adaptation, validity and internal consistency. Transcult Psychiatry. 2021;58:76–95. Medline:33297859 doi:10.1177/1363461520976045

27 Hakim MA, Liu JH. Development, construct validity, and measurement invariance of the parasocial relationship with political figures (PSR-P) scale. Int Perspect Psychol. 2021;10:13–24. doi:10.1027/2157-3891/a000002

28 Sproesser G, Klusmann V, Ruby MB, Arbit N, Rozin P, Schupp HT, et al. The positive eating scale: relationship with objective health parameters and validity in Germany, the USA and India. Psychol Health. 2018;33:313–39. Medline:28641449 doi:10.1080/08870446.2017.1336239

29 McCoy DC, Waldman M, Team CF, Fink G. Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments. Early Child Res Q. 2018;45:58–68. doi:10.1016/j.ecresq.2018.05.002

30 Korf D, Benschop A, Werse B, Kamphausen G, Felvinczi K, Dąbrowska K, et al. How and where to find NPS users: a comparison of methods in a cross-national survey among three groups of current users of new psychoactive substances in Europe. Int J Ment Health Addict. 2021;19:873–90. doi:10.1007/s11469-019-0052-8

31 O'Brien KM, Yoo S-K, Kim YH, Cho Y, Salahuddin NM. The good mothering expectations scale: An international instrument development study. Couns Psychol. 2020;48:162–90. doi:10.1177/0011000019889895

32 Littrell RF, Warner-Soderholm G, Minelgaite I, Ahmadi Y, Dalati S, Bertsch A, et al. Explicit preferred leader behaviours across cultures: instrument development and validation. J Manage Dev. 2018;37:243–57. doi:10.1108/JMD-09-2017-0294

33 Whiting-Collins L, Grenier L, Winch PJ, Tsui A, Donohue PK. Measuring contraceptive self-efficacy in sub-Saharan Africa: development and validation of the CSESSA scale in Kenya and Nigeria. Contracept X. 2020;2:100041. Medline:33145490 doi:10.1016/j.conx.2020.100041

34 Encantado J, Marques MM, Palmeira AL, Sebire SJ, Teixeira PJ, Stubbs RJ, et al. Development and cross-cultural validation of the Goal Content for Weight Maintenance Scale (GCWMS). Eat Weight Disord. 2021;26:2737–48. Medline:33646516 doi:10.1007/s40519-021-01148-x

35 Stevelink SAM, Hoekstra T, Nardi SMT, van Der Zee CH, Banstola N, Premkumar R, et al. Development and structural validation of a shortened version of the Participation Scale. Disabil Rehabil. 2012;34:1596–607. Medline:22372970 doi:10.3109/09638288.2012.656793

36 Datu JAD, Fincham F, Buenconsejo JU. Psychometric validity and measurement invariance of the caring for Bliss Scale in the Philippines and the United States. J Am Coll Health. 2024;72:1394–400. Medline:35623061 doi:10.1080/07448481.2022.2076562

37 Lopez-Fernandez O, Griffiths MD, Kuss DJ, Dawes C, Pontes HM, Justice L, et al. Cross-Cultural Validation of the Compulsive Internet Use Scale in Four Forms and Eight Languages. Cyberpsychol Behav Soc Netw. 2019;22:451–64. Medline:31295025 doi:10.1089/cyber.2018.0731

38 Erhart M, Hagquist C, Auquier P, Rajmil L, Power M, Ravens-Sieberer U, et al. A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. Child Care Health Dev. 2010;36:473–84. Medline:19702637 doi:10.1111/j.1365-2214.2009.00998.x

39 Lau C, Chiesi F, Saklofske DH, Yan G, Li C. How essential is the essential resilience scale? Differential item functioning of Chinese and English versions and criterion validity. Pers Individ Dif. 2020;155:109666. doi:10.1016/j.paid.2019.109666

40 Geyh S, Fellinghauer BA, Kirchberger I, Post MW. Cross-cultural validity of four quality of life scales in persons with spinal cord injury. Health Qual Life Outcomes. 2010;8:94. Medline:20815864 doi:10.1186/1477-7525-8-94

41 Boudjemadi V, Chauvin B, Adam S, Indoumou-Peppe C, Lagacé M, Lalot F, et al. Assessing the Cross-Cultural Validity of the Succession, Identity and Consumption (SIC) Scale Across Four French-Speaking Countries. Int Rev Soc Psychol. 2022;35:5. doi:10.5334/irsp.544

42 Pendergast LL, Schaefer BA, Murray-Kolb LE, Svensen E, Shrestha R, Rasheed MA, et al. Assessing development across cultures: Invariance of the Bayley-III Scales Across Seven International MAL-ED sites. Sch Psychol Q. 2018;33:604–14. Medline:30507236 doi:10.1037/spq0000264

43 Michaud J, Lvina E, Galperin BL, Lituchy TR, Punnett BJ, Taleb A, et al. Development and validation of the Leadership Effectiveness in Africa and the Diaspora (LEAD) scale. Int J Cross Cult. 2020;20:361–84. doi:10.1177/1470595820973438

44 Benschop A, Urbán R, Kapitány-Fövény M, Van Hout MC, Dąbrowska K, Felvinczi K, et al. Why do people use new psychoactive substances? Development of a new measurement tool in six European countries. J Psychopharmacol. 2020;34:600–11. Medline:32043399 doi:10.1177/0269881120904951

45 Brislin RW. Back-translation for cross-cultural research. J Cross Cult Psychol. 1970;1:185–216. doi:10.1177/135910457000100301

46 Ndosi M, Tennant A, Bergsten U, Kukkurainen ML, Machado P, de la Torre-Aboki J, et al. Cross-cultural validation of the Educational Needs Assessment Tool in RA in 7 European countries. BMC Musculoskelet Disord. 2011;12:110. Medline:21609481 doi:10.1186/1471-2474-12-110

47 Oort Q, Dirven L, Sikkes SA, Aaronson N, Boele F, Brannan C, et al. Development of an EORTC questionnaire measuring instrumental activities of daily living (IADL) in patients with brain tumours: phase I–III. Qual Life Res. 2021;30:1491–502. Medline:33496902 doi:10.1007/s11136-020-02738-5

48 Wong Riff KW, Tsangaris E, Goodacre T, Forrest CR, Pusic AL, Cano SJ, et al. International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). BMJ Open. 2017;7:e015467. Medline:28077415 doi:10.1136/bmjopen-2016-015467

49 Bőthe B, Potenza MN, Griffiths MD, Kraus SW, Klein V, Fuss J, et al. The development of the Compulsive Sexual Behavior Disorder Scale (CSBD-19): An ICD-11 based screening measure across three languages. J Behav Addict. 2020;9:247–58. Medline:32609629 doi:10.1556/2006.2020.00034

50 Chen FF. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Struct Equ Modeling. 2007;14:464–504. doi:10.1080/10705510701301834

51 Cheung GW, Rensvold RB. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. Struct Equ Modeling. 2002;9:233–55. doi:10.1207/S15328007SEM0902_5

52 Vogel DL, Heath PJ, Engel KE, Brenner RE, Strass HA, Al-Darmaki FR, et al. Cross-cultural validation of the Perceptions of Stigmatization by Others for Seeking Help (PSOSH) Scale. Stigma Health. 2019;4:82–5. doi:10.1037/sah0000119

53 Dembla P, Cornwell B, Keillor B. Scale development in cross-cultural consumer behavior. American Marketing Association. 2000;11:250.

54 Douglas SP, Craig CS. Collaborative and Iterative Translation: An Alternative Approach to Back Translation. J Int Mark. 2007;15:30-43. doi:10.1509/jimk.15.1.030

55 Harkness JA, Villar A, Edwards B. Translation, Adaptation, and Design. In: Harkness JA, Braun M, Edwards B, Johnson TP, Lyberg LE, Mohler PP, et al., editors. Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Hoboken, USA: John Wiley and Sons; 2010. p. 115–40.

56 Scott K, Ummer O, LeFevre AE. The devil is in the detail: reflections on the value and application of cognitive interviewing to strengthen quantitative surveys in global health. Health Policy Plan. 2021;36:982–95. Medline:33978729 doi:10.1093/heapol/czab048

57 Joshanloo M, Lepshokova ZKh, Panyusheva T, Natalia A, Poon W-C, Yeung VW, et al. Cross-Cultural Validation of Fear of Happiness Scale Across 14 National Groups. J Cross Cult Psychol. 2014;45:246–64. doi:10.1177/0022022113505357

58 Steenkamp J-BEM, Baumgartner H. Assessing Measurement Invariance in Cross-National Consumer Research. J Consum Res. 1998;25:78–90. doi:10.1086/209528

59 Nunnally JC. Psychometric theory. New York, USA: McGraw-Hill; 1967.

60 Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. Psychol Bull. 1988;103:265–75. Medline:3363047 doi:10.1037/0033-2909.103.2.265

61 Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate Data Analysis. Upper Saddle River, New Jersey, USA: Pearson; 2009. Available: https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf. Accessed: 11 July 2024.

62 Khan A, Yavorsky C, Liechti S, Opler M, Rothman B, DiClemente G, et al. A rasch model to test the cross-cultural validity in the positive and negative syndrome scale (PANSS) across six geo-cultural groups. BMC Psychol. 2013;1:5. Medline:25566357 doi:10.1186/2050-7283-1-5

63 Zumbo BD. A Handbook on the Theory and Methods of Differential Item Functioning (dif) logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999. Available: https://www.semanticscholar.org/paper/A-Handbook-on-the-Theory-and-Methods-of-Item-(DIF)-Zumbo/7f88fb0ad98645582665532600d7c46406fa2db6. Accessed: 7 August 2023.

64 Lai J-S, Teresi J, Gershon R. Procedures for the Analysis of Differential Item Functioning (DIF) for Small Sample Sizes. Eval Health Prof. 2005;28:283–94. Medline:16123258 doi:10.1177/0163278705278276

65 Squires A. Methodological challenges in cross-language qualitative research: A research review. Int J Nurs Stud. 2009;46:277–87. Medline:18789799 doi:10.1016/j.ijnurstu.2008.08.006

66 Harkness JA. Comparative Survey Research: Goals and Challenges. In: De Leeuw ED, Hox J, Dillman D, editors. International Handbook of Survey Methodology. New York, New York, USA: Routledge; 2008. Available: http://joophox.net/papers/SurveyHandbookCRC.pdf. Accessed: 11 July 2024.

67 Durham J. Ethical challenges in cross-cultural research: a student researcher's perspective. Aust N Z J Public Health. 2014;38:509–12. Medline:25377146 doi:10.1111/1753-6405.12286

68 Humphery K. Dirty questions: Indigenous health and 'Western research.'. Aust N Z J Public Health. 2001;25:197–202. Medline:11494986 doi:10.1111/j.1467-842X.2001.tb00563.x

69 Bloch A. Methodological Challenges for National and Multi-sited Comparative Survey Research. J Refug Stud. 2007;20:230–47. doi:10.1093/jrs/fem002

70 Sperber AD. The challenge of cross-cultural, multi-national research: potential benefits in the functional gastrointestinal disorders. Neurogastroenterol Motil. 2009;21:351–60. Medline:19309414 doi:10.1111/j.1365-2982.2009.01276.x

71 Pinto da Costa M. Conducting Cross-Cultural, Multi-Lingual and Multi-Country Focus Groups: Guidance for Researchers. Int J Qual Methods. 2021;20:16094069211049929. doi:10.1177/16094069211049929

72 Dong Y, Dumas D. Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. Pers Individ Dif. 2020;160:109956. doi:10.1016/j.paid.2020.109956