International comparability of depression scores from self-report scales: opportunities and challenges

Bernd Löwe and Sebastian Kohlmann*

Department of Psychosomatic Medicine and Psychotherapy, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany

The Lancet Regional Health - Europe 2023;31: 100668

Published Online xxx

https://doi.org/10.1016/j.lanepe.2023.100668

Self-report scales are widely used worldwide, particularly in the area of depressive disorders. When comparing study results from different countries, it is often implicitly assumed that the results from the same self-report scale from different countries are comparable. However, even when scales are perfectly translated, linguistic or cultural differences can lead to a different dimensionality of the instrument and to different ratings of the individual items. Differences in depression scale scores between countries therefore do not automatically reflect differences in the severity of depression between countries, but may also reflect differences in the understanding and weighting of individual items by the populations. Reliability, and therefore measurement accuracy, may also vary between countries. Therefore, before comparing the results of depression scales from different countries, it must be proven that the same depression scale in different countries is indeed measuring the same construct and that the measured severity levels are comparable. This is the question addressed by a recent study published in *The Lancet Regional Health - Europe*[1] using one of the most widely used self-report depression scales worldwide, the Patient Health Questionnaire-8 (PHQ-8).[2]

Arias-de la Torre et al. base their evaluations on a data set from 27 European countries with a total of 258,888 participants, which justifies a high representativeness and generalisability of the results for the countries involved.[1] With the PHQ-8, which consists of the first 8 items of the original 9-item Patient Health Questionnaire-9 (PHQ-9),[2,3] the authors have chosen an instrument translated into over 100 languages. The advantage of the PHQ-9 and the PHQ-8 over other depression scales is that their items reflect the diagnostic criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR, DSM-5, DSM-IV), and the International Classification of Diseases (ICD-10, ICD-11), resulting in particularly good criterion validity.[4] Although already demonstrated in countless previous studies, the Arias-de la Torre et al. study provides very solid confirmation that the structure of the PHQ-8 is unidimensional in all countries studied and that the PHQ-8 measures depression severity with high reliability.[1,3,5] This in itself is valuable. However, the unique scientific value of the study is that its results demonstrate the international comparability of the scoring of the PHQ-8. Specifically, this means that, at the group level, results from the PHQ-8 administered in English in the UK are comparable to those administered in Italian in Italy, for example. Thus, epidemiological and clinical studies using the PHQ-8 as an outcome can now be meaningfully conducted across the 27 countries that participated in the study.

Limitations arise with regard to transferability to other self-report instruments: Since the equivalence of the PHQ-8 and PHQ-9 scores has been established in a recent individual patient data meta-analysis,[6] it is very likely that the results found here for the PHQ-8 also largely apply to the PHQ-9. Still, results cannot be transferred to other self-report instruments or to languages or countries not involved in the study. Second, it should be noted that no specific cut-off values of the PHQ-8 have been compared. Thus, no statement can be made as to whether the same cut-off values for the screening of depressive disorders are valid in different countries. Finally, it is important that the authors make clear that uncritical comparison of depression scores across countries is not without problems. The authors themselves had previously compared the prevalence of depression in the 27 countries using the same data set as in the current study.[1,7] Strictly speaking, the international comparability of the PHQ-8 results should have been established here before reporting cross-national prevalences assessed with this instrument.

What do the results mean for clinical practice? In contrast to the use of the PHQ-8 at the group level addressed in the present study,[1] its uncritical use at the individual level, for example as a screening tool for depression, is of course even more problematic, as both false-positive and false-negative results can have adverse consequences. International guidelines on depression screening unfortunately vary widely in whether they recommend general screening for depressive disorders, screening of at-risk groups, or no screening at all – not surprising, since efficacy studies are almost completely lacking in this regard. In any case, it is crucial for effective depression screening that positive screening results lead to clinical consequences, e.g. a subsequent consultation for further diagnosis and therapy referral. Results of a current randomized-controlled trial show that a feedback of the screening result to the patient and the physician can reduce depression severity as compared to

DOI of original article: https://doi.org/10.1016/j.lanepe.2023.100659

*Corresponding author. Department of Psychosomatic Medicine and Psychotherapy, University Medical Centre Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany.

E-mail address: b.loewe@uke.de (B. Löwe).

feedback to physician only.[8] Although this trial is currently being replicated in two larger and independent samples,[9,10] its results point to the importance of actively involving patients in the screening process.

The core message of the article by Arias-de la Torre et al.[1] is that the PHQ-8 is a self-report instrument for depressive disorders that is not only reliable and factor-stable, but also internationally comparable. This opens up new opportunities for comparative international research to substantiate the role of self-report scales in screening, severity assessment and monitoring of psychopathology.

Contributors

Bernd Löwe conceived the manuscript and wrote the first draft. Sebastian Kohlmann critically reviewed and revised the first draft. Both authors produced the final version of the manuscript.

Declaration of interests

Bernd Löwe reports research funding (no personal honoraria) from the German Research Foundation, the German Federal Ministry of Education and Research, the German Innovation Committee at the Joint Federal Committee, the European Commission Horizon 2020 Framework Programme, the European Joint Programme for Rare Diseases (EJP), the Ministry of Science, Research and Equality of the Free and Hanseatic City of Hamburg, Germany and the Foundation Psychosomatics of Spinal Diseases, Stuttgart, Germany. He has received a remuneration as a committee member from Aarhus University, Denmark.

Sebastian Kohlmann reports research funding (no personal honoraria) from the German Research Foundation and the German Federal Ministry of Education and Research.

References

1 Arias-de la Torre JA, Vilagut G, Ronaldson A, et al. Reliability and cross-country equivalence of the 8-Item version of the Patient Health Questionnaire (PHQ-8) for the assessment of depression: results from 27 countries in Europe. *Lancet Reg Health Eur*. 2023. https://doi.org/10.1016/j.lanepe.2023.100659.

2 Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. 2009;114:163–173.

3 Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatr*. 2010;32:345–359.

4 Löwe B, Spitzer RL, Gräfe K, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord*. 2004;78:131–140.

5 Negeri ZF, Levis B, Sun Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ*. 2021;375:n2183.

6 Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med*. 2020;50:1368–1380.

7 Arias-de la Torre J, Vilagut G, Ronaldson A, et al. Prevalence and variability of current depressive disorder in 27 European countries: a population-based study. *Lancet Public Health*. 2021;6:e729–e738.

8 Löwe B, Blankenberg S, Wegscheider K, et al. Depression screening with patient-targeted feedback in cardiology: DEPSCREEN-INFO randomised clinical trial. *Br J Psychiatry*. 2017;210(2):132–139.

9 Kohlmann S, Lehmann M, Eisele M, et al. Depression screening using patient-targeted feedback in general practices: study protocol of the German multicentre GET.FEEDBACK.GP randomised controlled trial. *BMJ Open*. 2020;10:e035973.

10 Sikorski F, König HH, Wegscheider K, Zapf A, Löwe B, Kohlmann S. The efficacy of automated feedback after internet-based depression screening: study protocol of the German, three-armed, randomised controlled trial DISCOVER. *Internet Interv*. 2021;25:100435.