PSYCHOMETRIKA—VOL. 86, NO. 3, 747–753
SEPTEMBER 2021
https://doi.org/10.1007/s11336-021-09778-x

SOLVING THE TOWER OF BABEL PROBLEM FOR PATIENT-REPORTED OUTCOME MEASURES

Comments on: Linking Scores with Patient-Reported Health Outcome Instruments: A Validation Study and Comparison of Three Linking Methods

JAKOB BUE BJORNER

QUALITYMETRIC INCORPORATED, LLC

UNIVERSITY OF COPENHAGEN

NATIONAL RESEARCH CENTRE FOR THE WORKING ENVIRONMENT

The PROsetta Stone Project, summarized in this issue by Schalet et al. (Psychometrika 86, 2021), is a major step forward in enabling comparability between different patient-reported outcomes measures. Schalet et al. clearly describe the psychometric methods used in the PROsetta Stone project and other projects from the Patient-Reported Outcomes Measurement Information System (PROMIS): linking based on unidimensional item response theory (IRT), equipercentile linking, and calibrated projection based on multidimensional IRT. Analyses in a validation data set and simulation studies provide strong support that the linking methods are robust when basic assumptions are fulfilled. The links already established will be of great value to the field, and the methodology described by Schalet et al. will hopefully inspire the next series of linking studies. Among potential improvements that should be considered by new studies are: (1) a thorough evaluation of the content of the measures to be linked to better guide the evaluation of measurement assumptions, (2) improvements in the design of linking studies such as selection of the optimal sample to provide data in the score ranges where linking precision is most critical and using counterbalanced designs to control for order effects. Finally, it may be useful to consider how the linking algorithms are used in subsequent data analyses. Analytic strategies based on plausible values or latent regression IRT models may be preferable to the simple transformation of scores from one patient at the time.

Key words: linking, equating, item response theory, patient-reported outcomes, depression.

1. Introduction

The “Tower of Babel” problem for patient-reported outcome measures (PROMs) was explicitly stated more the 30 years ago (Mor and Guadagnoli 1988; van Knippenberg and Haes 1988). In 1988, according to the authors, PROMs had proliferated without a uniform approach, without a clear conceptual framework, and with only limited agreement on the definition of core concepts. All these factors hindered the comparison of scores. Since then, the proliferation of PROMs has only increased. Luckily, the conditions for overcoming the Tower of Babel problem have improved as well. While the field has not agreed on one overall conceptual framework, there is practical agreement on a number of core concepts, the field has established standardized methods for achieving content validity, and there is increasing alignment on ways to phrase questions and response choices. Last, but not least, methods to link scores from different PROMs have been imported and adapted from educational testing. The excellent paper by Schalet et al. (2021) describes the steps taken when using the current standard: linking based on unidimensional item response theory

Correspondence should be made to Jakob Bue Bjorner, QualityMetric Incorporated, LLC, 1301 Atwood Avenue, Suite 311N, Johnston, RI 02919, USA. Email: jbjorner@qualitymetric.com

(IRT) models. Also, useful comparisons are made with other approaches: equipercentile linking and calibrated projection. The authors' careful empirical analysis convincingly demonstrates that when basic assumptions are fulfilled, the three approaches concur and these results are reasonably robust in a new data set. Further, Schalet et al. use simulation studies to identify situations where equipercentile linking or calibrated projection may be better choices for linking. These excellent analyses and clear results leave little to add regarding the psychometric work. Instead, I will comment on some issues around this solid psychometric foundation: that a content analysis may be a helpful supplement to correlations and factor analyses in evaluating unidimensionality (Sect. 2), that an optimal design of the linking study may help make the results more robust (Sect. 3), and that if the ultimate aim of the linking study is to enable better group comparisons, there may be alternative approaches to linking the score from each individual participant (Sect. 4).

2. Content Analysis

In their analysis of sufficient unidimensionality of the measures to be linked, Schalet et al. rely on score correlations supplemented by confirmatory factor analyses. They cite Dorans (2004) for using 0.866 as a lower bound for an acceptable correlation, but note that correlations in the range of 0.70 to 0.85 may be acceptable if the purpose is to enable group comparisons. I would argue that content analysis may add insight into whether two measures can be linked and in what situations linking may be problematic. Table 1 presents results of a content analysis of the two measures linked by Schalet et al.: the PHQ-9 (Kroenke et al. 2001) and the PROMIS depression item bank (Pilkonis et al. 2011). I used the DSM-IV (2000) depression criteria as the organizing framework, since the PHQ-9 was built to reflect these criteria (Kroenke et al. 2001). However, while the developers of the PROMIS item bank were well aware of the DSM-IV depression criteria, they also relied on other conceptual frameworks and patient interviews in the item bank development. Also, some subdomains, like fatigue and sleep, are covered by other item banks and therefore only covered sparsely in the depression item bank. Finally, the PROMIS depression item bank avoids items concerning somatic symptoms such as weight gain or loss, fatigue, and psychomotor speed, since these symptoms may cause problems for evaluating psychiatric morbidity in patients with somatic disease (see, e.g., Holzapfel et al. 2008). Thus, despite considerable content overlap between PROMIS depression item bank and the PHQ-9, there are also distinct differences, suggesting that the link between the two tools may be different in patients with somatic disease. Confounding by somatic symptoms may explain the discrepancy between cross-walked and actual PROMIS depression scores found in some studies of somatic patients (Katzan et al. 2017; Kim et al. 2017). While Schalet et al. should be applauded for testing the robustness of their linking across gender and age group, testing whether the linking is also valid for patients with somatic disease would be advisable. A content analysis may be useful in identifying such potential problems.

3. Design Considerations

Schalet et al. made excellent use of archival data from general population samples. However, it may be useful to consider the optimal design for a linking study. The PROMIS depression item bank and the PHQ-9 were developed for use in clinical research and diagnosis of depression. The two measures have optimal precision in the T-score range of 45 to 80—the range relevant for assessing clinical depression severity. However, scores between 70 and 80 are rare in the general population. In this range, the linking methods show discrepant results. As Schalet et al. note for

TABLE 1.
Content validity of the PHQ-9 and PROMIS depression items according to DSM-IV depression criteria

DSM-IV depression criteria	PHQ-9 items	PROMIS depression item bank[a]		
1. Depressed Mood	Feeling down, depressed, or hopeless?	I felt hopeless I felt sad	I felt depressed I felt pessimistic	I felt unhappy I felt upset for no reason
2. Markedly diminished interest or pleasure in most or all activities	Little interest or pleasure in doing things?	I felt that nothing could cheer me up I felt that nothing was interesting	I felt that my life was empty	I found that things in my life were overwhelming
3. Significant weight loss (or poor appetite) or weight gain	Poor appetite or overeating?	–		
4. Insomnia or hypersomnia	Trouble falling or staying asleep, or sleeping too much?	–		
5. Psychomotor retardation	Moving or speaking so slowly that other people could have noticed? Or so fidgety or restless that you have been moving a lot more than usual?	-		
6. Fatigue or loss of energy	Feeling tired or having little energy?	I felt emotionally exhausted		
7. Feelings of worthlessness or excessive or inappropriate guilt	Feeling bad about yourself—or that you are a failure or have let yourself or your family down?	I felt like a failure I felt guilty I felt disappointed in myself	I felt worthless I felt helpless	I felt that I was not as good as other people I felt that I was to blame for things

TABLE 1.
Continued

DSM-IV depression criteria	PHQ-9 items	PROMIS depression item bank[a]		
8. Diminished ability to think or concentrate, or indecisiveness	Trouble concentrating on things, such as reading the newspaper or watching television?	I had trouble making decisions		
9. Recurrent thoughts of death (not just fear of dying), or suicidal ideation, plan, or attempt	Thoughts that you would be better off dead, or thoughts of hurting yourself in some way?	I felt I had no reason for living		
–	–	I felt that I had nothing to look forward to	I felt discouraged about the future	I felt that I wanted to give up on everything
–	–	I withdrew from other people	I felt that I was not needed	I felt lonely
		I had trouble feeling close to people	I felt ignored by people	

[a] In practice, only 20 out of the 28 PROMIS items were used for linking. Abbreviations: *DSM* diagnostic and statistical manual of mental disorders, *PHQ* patient heath questionnaire, *PROMIS* patient-reported outcomes measurement information system.

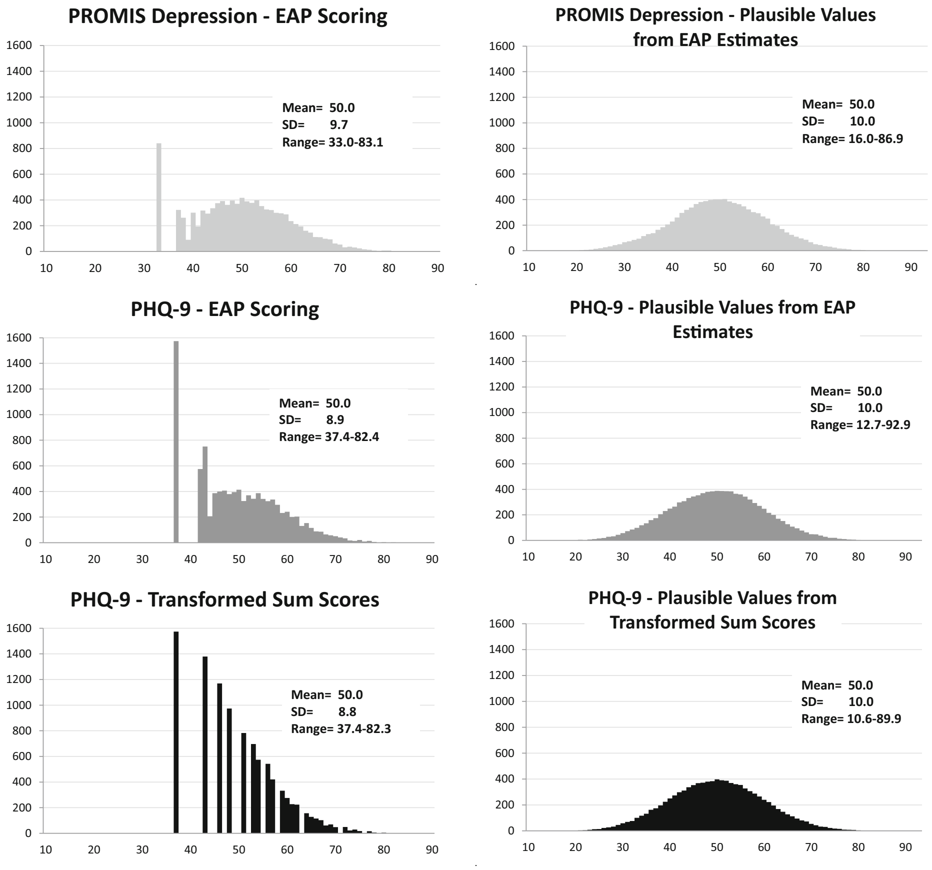

FIGURE 1.
Comparison of distribution of scores from scoring procedures in Schalet et al. (2021) and from a plausible values approach. All simulations were based on latent scores simulated to have a normal distribution with a T-score mean of 50 and a standard deviation of 10. Right column: 10 random values were drawn for each person based on the score estimate and the standard error of measurement (assuming a normal distribution of the error term). Abbreviations: EAP expected a posteriori; PHQ Patient Heath Questionnaire; PROMIS Patient-Reported Outcomes Measurement Information System

the equipercentile scoring method, the discrepancy in this score range is likely to be partly caused by sparse data. A patient sample including participants with high depression scores would provide a more robust link in the severe score range.

Rather than the random-groups design often used in educational linking studies (Kolen and Brennan 2014), the PROsetta Stone project chose a single-group design. This design has advantages in the ability to check the unidimensionality assumption and check for agreement between linked and observed scores. Also, the single-group design has greater statistical power for a given sample size. The main potential problem of the single-group design is the possibility of order effects, e.g., due to respondent fatigue. Schalet et al. note that the order of questionnaire administration should be counterbalanced but suggest that this is less critical in patient-reported outcome (PRO) research since test-taking fatigue is unlikely to be a major factor. However, available evidence shows otherwise. In a PROMIS study of methods of administration, two parallel depression short forms were developed from the PROMIS depression item bank (Bjorner et al. 2014). The forms were administered counterbalanced allowing for an estimation of the order effect. Results showed highly significant order effects: scores for whatever form was administered last were 1.94

to 4.68 T-score points lower, indicating less depression. These results suggest that counterbalancing may also be important for linking studies in the PRO field.

4. Linking Procedures for Group Comparisons

Schalet et al. provide a very useful discussion of the differences between the purposes of linking in educational testing and in PRO research. One difference is that educational testing often involves decisions made on the basis of an individual equated score or cut-off value. In contrast, PROMs are mostly used for comparisons of groups. Given this emphasis on group-based analyses, it may not be wise to apply the linking procedure of Schalet et al. in the most simplistic way: for each person who has answered the PHQ-9, simply estimate a score on the PROMIS metric. To illustrate, I simulated a data set using the same item parameters and sum-score linking procedures (Choi et al. 2014) that were evaluated by Schalet et al. The results from these simulations are presented in Fig. 1. The left column shows the distribution of score estimates by applying expected a posteriori (EAP) estimation based on the PROMIS depression item bank, EAP estimation based on the PHQ-9 items, and applying the sum score transformation algorithm. While all procedures were effective in capturing the correct mean of 50, the standard deviation is underestimated when using the PHQ-9 items and the score distribution is far from normal, due to floor effects that still exist after linking. The right column in Fig. 1 shows "score distributions" using a plausible values approach (Mislevy 1991)—well known in psychometric research. The plots illustrate that this approach was very effective in estimating the correct mean and standard deviation and achieving a score distribution more similar to the generating latent distribution. Thus, the plausible values approach or latent regression IRT models may be useful additions to the linking procedures discussed by Schalet et al. (also, see Fischer and Rose 2019).

5. Conclusion

While the PROsetta Stone project is not the first project to link PROMs (see, e.g., Orlando et al. 2000; Bjorner et al. 2003), it is the largest and most ambitious of such efforts within the field of PROMs. While some details might be improved, the links already established will be of great value to the field. Similarly, the excellent summary by Schalet et al. will be a great help to the next generation of researchers seeking to overcome the Tower of Babel problem.

References

Bjorner, J. B., Kosinski, M., & Ware, J. E, Jr. (2003). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research*, *12*(8), 981–1002.

Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E, Jr. (2014). Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. *Journal of Clinical Epidemiology*, *67*(1), 108–113. https://doi.org/10.1016/j.jclinepi.2013.07.016.

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527. https://doi.org/10.1037/a0035768.

DSM-IV-TR., A.P.A. (2000). *Diagnostic and statistical manual of mental disorders, fourth edition, text revision: DSM-IV-TR* (4th ed., text rev). Washington, DC: American Psychiatric Association.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227–246.

Fischer, H. F., & Rose, M. (2019). Scoring depression on a common metric: A comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivariate Behavioral Research*, *54*(1), 85–99.

Holzapfel, N., Müller-Tasch, T., Wild, B., Jünger, J., Zugck, C., Remppis, A., et al. (2008). Depression profile in patients with and without chronic heart failure. *Journal of Affective Disorders*, *105*(1–3), 53–62.

Katzan, I. L., Fan, Y., Griffith, S. D., Crane, P. K., Thompson, N. R., & Cella, D. (2017). Scale linking to enable patient-reported outcome performance measures assessed with different patient-reported outcome measures. *Value in Health*, *20*(8), 1143–1149.

Kim, J., Chung, H., Askew, R. L., Park, R., Jones, S. M. W., Cook, K. F., et al. (2017). Translating CESD-20 and PHQ-9 scores to PROMIS depression. *Assessment*, *24*(3), 300–307.

Kolen, M. L., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mor, V., & Guadagnoli, E. (1988). Quality of life measurement: A psychometric tower of Babel. *Journal of Clinical Epidemiology*, *41*(11), 1055–1058.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*(3), 354–359.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283.

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika, 86*. https://doi.org/10.1007/s11336-021-09776-z.

van Knippenberg, F. C., & de Haes, J. C. (1988). Measuring the quality of life of cancer patients: Psychometric properties of instruments. [Review]. *J.Clin.Epidemiol., 41*(11), 1043–1053.

Manuscript Received: 16 FEB 2021
Final Version Received: 6 MAR 2021
Accepted: 19 MAY 2021
Published Online Date: 18 JUN 2021