International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores

Terwee, C.B.; Crins, M.H.P.; Roorda, L.D.; Cook, K.F.; Cella, D.; Smits, N.; Schalet, B.D.

DOI
10.1016/j.jclinepi.2021.01.011

Publication date
2021

Document Version
Final published version

Published in
Journal of Clinical Epidemiology

License
CC BY

Link to publication

Citation for published version (APA):
Terwee, C. B., Crins, M. H. P., Roorda, L. D., Cook, K. F., Cella, D., Smits, N., & Schalet, B. D. (2021). International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. *Journal of Clinical Epidemiology*, *134*, 1-13. https://doi.org/10.1016/j.jclinepi.2021.01.011

Journal of Clinical Epidemiology 134 (2021) 1–13

ORIGINAL ARTICLE

International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores

Caroline B. Terwee[a,*], Martine H.P. Crins[b], Leo D. Roorda[b], Karon F. Cook[c], David Cella[c], Niels Smits[d], Benjamin D. Schalet[c]

[a] Amsterdam UMC, Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam Public Health Research Institute, de Boelelaan 1117, Amsterdam, the Netherlands
[b] Amsterdam Rehabilitation Research Center Reade, Amsterdam, the Netherlands
[c] Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[d] Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands

Received 26 June 2020; Received in revised form 13 January 2021; Accepted 18 January 2021; Available online 30 January 2021

Abstract

Objective: PROMIS offers computerized adaptive tests (CAT) of patient-reported outcomes, using a single set of US-based IRT item parameters across populations and language-versions. The use of country-specific item parameters has local appeal, but also disadvantages. We illustrate the effects of choosing US or country-specific item parameters on PROMIS CAT T-scores.

Study design and setting: Simulations were performed on response data from Dutch chronic pain patients ($n = 1110$) who completed the PROMIS Pain Behavior item bank. We compared CAT T-scores obtained with (1) US parameters; (2) Dutch item parameters; (3) US item parameters for DIF-free items and Dutch item parameters (rescaled to the US metric) for DIF items; (4) Dutch item parameters for all items (rescaled to the US metric).

Results: Without anchoring to a common metric, CAT T-scores cannot be compared. When scores were rescaled to the US metric, mean differences in CAT T-scores based on US vs. Dutch item parameters were negligible. However, 0.9%–4.3% of the T-score differences were larger than 5 points (0.5 SD).

Conclusion: The choice of item parameters can be consequential for individual patient scores. We recommend more studies of translated CATs to examine if strategies that allow for country-specific item parameters should be further investigated. © 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Keywords: Questionnaire; Patient-reported outcomes; PROMIS; Computerized Adaptive Test; Item Response Theory; Validation; Outcome measurement; Psychometrics

Conflicts of Interest: Authors Terwee, Cella and Cook are (former) board members of the PROMIS Health Organization. All authors have previously received grants for validation studies on PROMIS measures. All authors declare that they have no competing interests.

[*] Corresponding author. Tel.: +31204444474.
E-mail address: cb.terwee@amsterdamumc.nl (C.B. Terwee).

https://doi.org/10.1016/j.jclinepi.2021.01.011

What is new?

Key findings

- When scores were rescaled to the common US metric, mean differences between CAT T-scores based on US versus Dutch item parameters were negligible. However, 0.9%–4.3% of the T-score differences were larger than 5 points (0.5 SD).

What this adds to what is known?

- This is the first empirical study illustrating the effects of using different sets of country-specific item parameters as compared to the default US parameters on PROMIS CAT T-scores.
- The default PROMIS convention to use a single set of IRT item parameters across populations and language-versions may not have optimal validity in non-US countries, if there is substantial language DIF.

What is the implication, what should change now

- The choice of item parameters in CAT can be consequential for individual patient scores.
- More studies of translated CATs should be performed to examine if strategies that allow for country-specific item parameters should be further investigated.

1. Introduction

Item response theory (IRT) is increasingly used to create item banks as the basis for computerized adaptive testing (CAT) for measuring patient-reported outcomes (PROs) [1–5]. With CAT, after a starting item, subsequent items are selected by the computer based on participants' responses to previous items [6,7]. With CAT, reliable scores can be obtained with only a few relevant questions [8]. The Patient-Reported Outcomes Measurement Information System (PROMIS) is the largest system of PRO item banks administered as CATs [9–12].

Currently, the default PROMIS convention is to use a single set of IRT item parameters across populations and language-versions to express scores on a common scale (T-score metric), unless evidence shows that this is problematic, eg, if items function substantially different across populations or language-versions [9,13]. A common scale is required for international comparisons. The current official set of item parameters comprises the original IRT item parameters estimated in the US calibration sample [9,14]. However, we cannot expect to measure everyone accurately on the same scale. PROMIS guidelines therefore recommend conducting differential item functioning (DIF) analyses to test if people from different groups (eg, language) with the same level of the construct (eg, pain) respond differently to an item [6,15]. DIF by language may occur when items are poorly translated or because of population differences. When DIF is found, country-specific item parameters are needed. Despite rigorous translation methods [16], DIF has been found for PROMIS items, to varying degrees, between language versions [17–25]. If substantial DIF is found, a hybrid approach is recommended, as was used for some Spanish PROMIS measures. Spanish-specific item parameters are used for items with DIF, and the Spanish item parameters of the DIF-items are linked (also called rescaled), to the original PROMIS metric [22,23]. A method adopted from the equating and linking literature, called Stocking-Lord method, was used for this purpose [25–27].

In contrast to the recommended PROMIS approach, the use of country-specific item parameters, when estimated in a large, representative general population sample of the target country, has local appeal. Country-specific calibrations and score centering would allow for improved within-country interpretation and benchmarking. Also, scores obtained would more "purely" reflect the status of respondents. PROMIS uses a T-score metric, where a score of 50 represents the mean score of a reference population (often a general population sample) and 10 is the standard deviation [14]. In the US, a T-score of 50 compares to the US mean; a T-score of 60 is one standard deviation (SD) above the mean. This interpretation, however, does not necessarily apply to other countries because the average PRO level (and SD) may vary across countries [20,24,28,29]. The advantages and disadvantages of using country-specific item parameters for PROMIS measures are ingredients of an important debate because the decision to allow the use of country-specific item parameters or not can have major consequences for the maintenance, distribution, and uptake of PROMIS. We therefore aimed to discuss different options for selecting item parameter sets and to illustrate the effects of using the current PROMIS-recommended strategy vs. different country-specific strategies on PROMIS CAT T-scores in an empirical data set.

2. Methods

2.1. Options for applying PROMIS CATs outside of the US

Six options for choosing item parameter sets are presented and discussed briefly in Table 1 and more extensively in Appendix A. To illustrate the consequences of different options on CAT T-scores, Options 1 through 4 are compared in a simulation study: (1) US parameters; (2) country-specific (Dutch) item parameters; (3) US item parameters for DIF-free items and Dutch item parameters (rescaled to the US metric) for DIF items; (4) Dutch item parameters for all items (rescaled to the US metric). Option 2 is used for illustrative purposes only because with Option 2 different metrics will be used in each country

Table 1. Options for item parameter use in PROMIS CATs outside the US and mean (SD) CAT T-scores based on different options

Option	Description	Metric	Validity	Interpretation	Mean T-score	SD T-score
1	Use US item parameters in all countries. This is the default for PROMIS.	All T-scores across language-versions on the same (US) metric. One set of item parameters.	If there is language-DIF, the US item parameters may not be valid for non-US population.	A T-score of 50 represents the mean score of the US population but may have a different meaning in other countries.	59.58	6.65
2	Use country-specific item parameters.	Different metrics in each country Different sets of item parameters.	Country-specific item parameters have optimal accuracy for the local population.	A T-score of 50 represents the average of the local population in each country.	49.52	9.10
3	Hybrid approach: Use US item parameters for non-DIF items and use country-specific item parameters, rescaled to the US metric, for items with language DIF. This is currently recommended for PROMIS measures in case of substantial DIF.	One (US) metric. For DIF items, different sets of item parameters.	Item parameters for DIF items will be more accurate for the local population.	Similar to Option 1.	59.35	6.36
4	Use country-specific item parameters for all items and rescale all parameters to the US metric.	One (US) metric. Different sets of item parameters.	Item parameters are most accurate for the local population.	Similar to Option 1.	59.17	6.39
5	Use 'world' item parameters.	One (world) metric. One set of item parameters.	Similar to Option 1. Data will likely not be representative for the whole world and there is likely language DIF.	A T-score of 50 represents the mean score of the 'world' population, but may have a different meaning in every country.	Not available	
6	Use country-specific item parameters for local applications (option 2) and rescale scores to the US metric for international comparisons (option 4).	Different metrics (item parameters) across countries.	Risk of misinterpretation of scores if item parameters and rescaling is not (correctly) reported.	Similar to Option 2 for local applications. Similar to Option 4 for international comparisons.	See option 2 and 4	

and T-scores will not be comparable. A graphical representation of these four sets is provided in Fig. 1.

2.2. Dataset used for simulations

For illustrative purposes, we used real response data from a subset of 1,003 patients with no missing responses from a sample of 1,110 Dutch chronic pain patients who completed the full Dutch-Flemish PROMIS v1.1 Pain Behavior item bank [18]. The item bank consists of 39 items, with a recall period of the past 7 days, and response options of Had no pain, Never, Rarely, Sometimes, Often, and Always [30]. Since the US sample responded to only 31 out of the 39 items, the DIF analysis was based on 31 items only. Six items were flagged for DIF [18].

2.3. Analysis Plan

For Option 1 (US CAT T-scores), we obtained the official US v1.1 item parameters used in the US Assessment Center and the API provided by HealthMeasures via enquiry at www.healthmeasures.net. The US calibration sample consisted of participants from the US general population (n = 14,503) and chronic pain patients (n = 967), of which a small subsample completed the full item bank and the majority completed 7 items (each item was completed by about 3,000 participants). A Graded Response

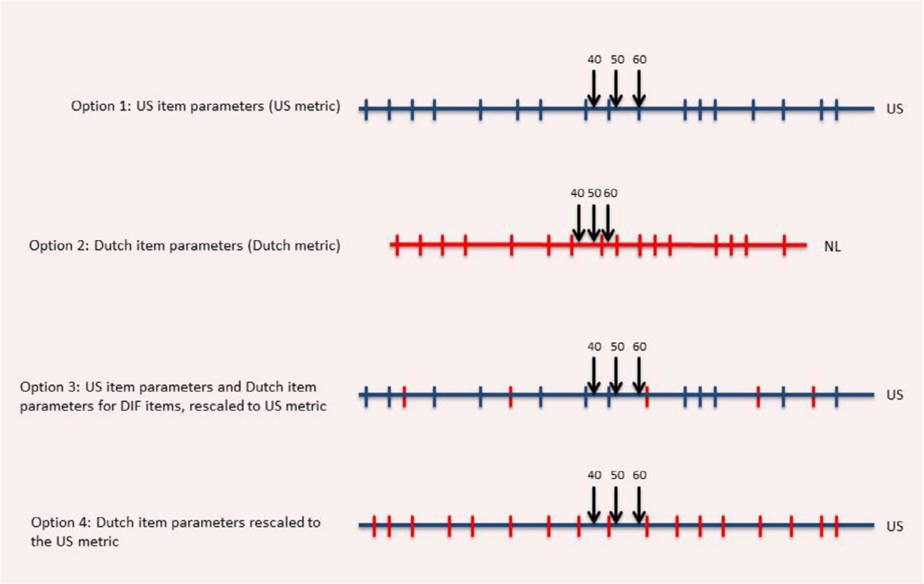
Option 1: US item parameters (US metric)
40 50 60
US
Option 2: Dutch item parameters (Dutch metric)
40 50 60
NL
Option 3: US item parameters and Dutch item parameters for DIF items, rescaled to US metric
40 50 60
US
Option 4: Dutch item parameters rescaled to the US metric
40 50 60
US

Fig. 1. Graphical representation of four Options for applying PROMIS CATs in non-English-speaking countries. Each horizontal line represents a metric (with its own mean and SD). Horizontal blue line represent the US metric. Horizontal red line represent the Dutch metric. Vertical blue lines represent example US item locations. Vertical red lines represent example Dutch item locations.

Model (GRM) was fit using MULTILOG [30]. For Option 2 (Dutch CAT T-scores), we estimated Dutch item parameters in the Dutch chronic pain sample using the GRM with the MIRT package [31] in R [32]. For Options 3 (hybrid CAT T-scores) and 4 (rescaled Dutch CAT T-scores), we used the Stocking-Lord method to rescale the Dutch item parameters for the DIF items only (Option 3) or for all items (Option 4; [33]). The Stocking-Lord method is a commonly used linking method to generate unique item parameters for some items in one group (in this case the Dutch sample), but keep a common metric (in this case the US metric) [25–27]. In the educational measurement, this technique is one of several used to incorporate newly developed test items to an existing item bank [27]; it has also been repeatedly applied to the calibration of PROMIS Spanish items [22,23]. Details of the Stocking-Lord method can be found in Appendix B.

We conducted simulation analyses with the Firestar (v1.3.2) program [34]. We selected CAT stopping rules consistent with current PROMIS recommendations: minimum 4 items, maximum 12 items, and a standard error of 0.3 (theta). For the remaining settings we followed Firestar defaults [34], which are also used in the PROMIS CAT software. CAT thetas were linearly transformed into T-scores per PROMIS convention. We compared mean (SD) CAT T-scores of the current default option (US CAT T-scores, Option 1) to each of the other options (Options 2, 3, and 4). We calculated Pearson correlations for the three comparisons. We plotted the US CAT T-scores (Option 1) against the differences with each of the other options. A Locally Weighted Scatterplot Smoothing (LOWESS) line was added to the figures to make apparent whether the difference varies across the scale. We calculated the mean (SD) differences in CAT T-scores between the US and the other options and the limits of agreement (mean difference ± 1.96*SD$_{\text{difference}}$). Finally, we calculated the percentage

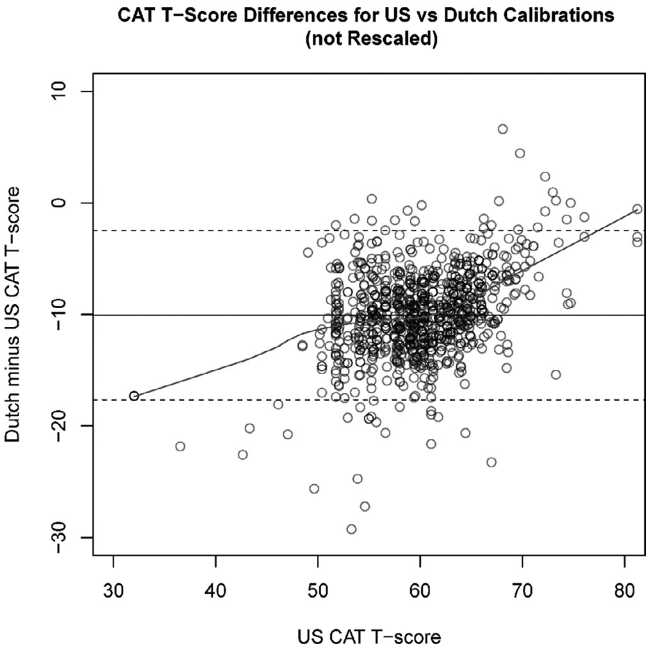

Fig. 2. US CAT T-scores (based on US item parameters [Option 1]) plotted against Dutch CAT T-scores (based on Dutch item parameters [Option 2]) minus US CAT T-scores, with LOWESS line.

Table 2. Comparison of CAT T-scores based on different sets of item parameters

Parameter sets used in comparison of CAT T-scores	Options compared	Pearson correlation	Mean T-score difference	SD of difference	Lower LOA	Upper LOA	% of patients with score difference >5 points
Dutch item parameters versus US item parameters[I]	2 vs 1	0.93	−10.06	3.87	−17.65	−2.47	92.32
US item parameters for non-DIF items and Dutch item parameters, rescaled to the US metric, for the DIF items versus US item parameters	3 vs 1	0.97	−0.23	1.49	−3.15	2.70	0.89
Dutch item parameters rescaled to the US metric versus US item parameters	4 vs 1	0.93	−0.41	2.42	−5.15	4.33	4.28

LOA = Limits of Agreement (mean difference ± 1.96 * SD difference).

[I] This comparison is provided for educational purposes only, to illustrate the large score differences that may occur when items are calibrated separately without any linking/rescaling strategy. When items sets are calibrated separately, the resulting item parameters reflect the latent mean and standard deviation of the two samples, such that the T-score units have a different meaning.

of patients for which the difference in T-score between the US and the comparison option was more than 5 points (both −5 and +5), as changes in PROMIS measures of 2–5 points have been suggested to be minimally important [35–37].

3. Results

Mean (SD) CAT T-scores of the patient population based on the four sets of item parameters are presented in Table 1. The mean T-scores were about 59 when scores were (rescaled) on the US metric and differences in means and SDs between Options 1, 3, and 4 were negligible. The Dutch pain sample represented a more restricted range of pain behavior scores relative to the representative subsample of the US general population as indicated by SDs of about 6 (instead of 10) on the US metric. The mean T-score was 49.5 (SD 9.1) on the Dutch metric (Option 2), as expected given the calibration defaults (theta mean of 0 and SD of 1 in the calibration sample).

We now move to compare the scores between US CAT T-scores and Dutch CAT T-scores directly, which is in some sense non-sensical: the scores are based on two different metrics, such that the T-scores (and the underlying IRT theta units) have an incommensurate meaning. But we do this to illustrate what would happen if international researchers separately calibrate PROMIS banks, and then do nothing further to place the parameters on a common metric. First, we note that the correlation between the US CAT T-scores (Option 1) and the Dutch CAT T-scores (Option 2) remains high, at 0.93 (Table 2), because the order of patients is similar with both options. That is, the T-scores in either case are based on the same set of item responses. However, the absolute T-scores and variation in T-scores are very different (Table 2). Fig. 2 shows the Dutch CAT T-scores (Option 2) minus US CAT T-scores plotted against the US CAT T-scores (Option 1).The mean difference in CAT T-scores was −10.1 points (SD 3.87; Table 2). If this strategy were in fact implemented, readers might wonder if the Dutch had a much higher pain tolerance than people in the US or a greater reluctance to exhibit pain behaviors. However, we note again that this comparison is improper: the magnitude of the difference is due to a comparison of two different metrics, originating from two different samples and languages—equivalent to comparing temperatures measured in centigrade to those measured in Fahrenheit.

The correlation between the US CAT T-scores (Option 1) and the hybrid CAT T-scores (Option 3: T-scores based on US item parameters for non-DIF items; Dutch item parameters, rescaled to the US metric, for the six DIF items) was 0.97. Fig. 3 shows the hybrid CAT T-scores minus US CAT T-scores plotted against the US CAT T-scores (Option 1). The Stocking-Lord constants for placing the Dutch DIF items onto the US metric were 0.5727 (A) and 1.008 (B). The mean difference in CAT T-scores between Options 1 and 3 was -0.23 (SD 1.49) and 0.9% of the differences was larger than 5 points (Table 2).

The correlation between the US CAT T-scores (Option 1) and the rescaled Dutch CAT T-scores (Option 4: T-scores based on Dutch item parameters, all rescaled to the US metric) was 0.93. Fig. 4 shows the rescaled Dutch CAT T-scores minus the US CAT T-scores plotted against the US CAT T-scores (Option 1). The mean difference in CAT T-scores was −0.41 but the SD was 2.42 and 4.3% of the differences was larger than 5 points (Table 2). In Option3 only the item parameters of the 6 DIF items are different than in Option 1. In Option 4, all item parameters are (slightly) different than in Option 1. Therefore, the differences in CAT T-scores between Option 4 and Option 1, were larger than the differences between Option 3 and Option 1.

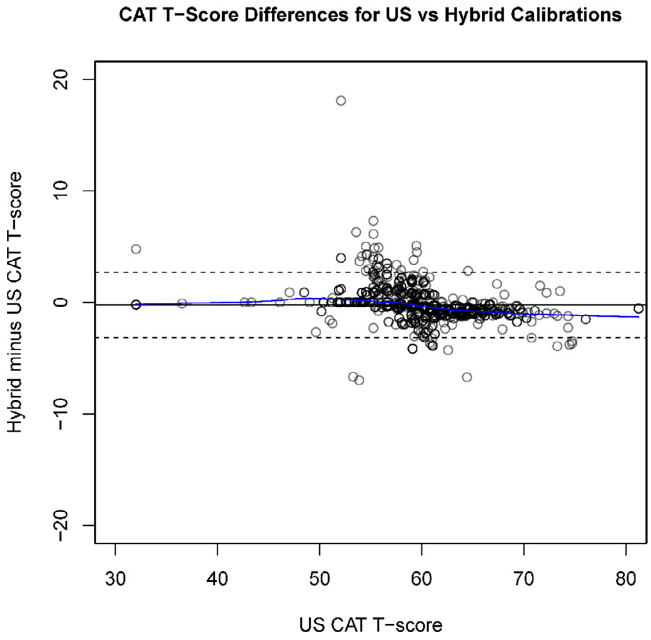

Fig. 3. US CAT T-scores (based on US item parameters [Option 1]) plotted against hybrid CAT T-scores (based on US item parameters for non-DIF items; Dutch item parameters, rescaled to the US metric, for the six DIF items [Option3]) minus US CAT T-scores, with LOWESS line (*the outlier reflects one patient who responded inconsistently to the items in the CAT*).

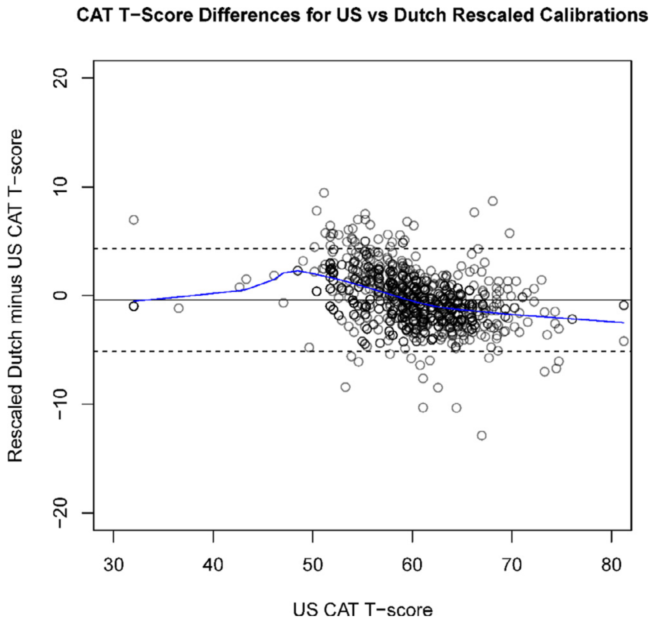

Fig. 4. US CAT T-scores (based on US item parameters [Option 1]) plotted against rescaled Dutch CAT T-scores (based on Dutch item parameters, rescaled to the US metric, for all items [Option 4]) minus US CAT T-scores with LOWESS line.

4. Discussion

This is the first empirical study showing the effects of using different sets of country-specific item parameters as compared to the default US parameters on PROMIS CAT T-scores. We used the PROMIS Pain Behavior item bank as an example. Naturally, we found high correlations between CAT T-scores across all options because identical item responses were entered in the CAT simulations. When country-specific item parameters were compared to US parameters, obviously large mean T-score differences were obtained because scores were not on a common metric. We provide this “result” as an illustrative and cautionary tale for international researchers who may re-calibrate PROMIS in samples outside of the US, not realizing that this changes the meaning and interpretation of the resulting T-score units. Only when samples are very closely matched on demographics and the domain of interest (in this case pain) can we expect similar item parameter estimates. Negligible mean differences were found when item parameters were rescaled to a common metric (Options 3 and 4). However, some substantial differences were found at the individual level. Between 0.9% (Option 3) and 4.3% (Option 4) of the T-score differences exceeded 5 points. This indicates that, while mean T-scores are little affected by the choice of item parameters as long as the scores are on a common metric, individual T-scores can vary by more than the amount generally considered minimally important (2–5 points). This should be considered when the measure is used in clinical or research settings.

In deciding whether or not to allow use of country-specific item parameters, tradeoffs must be made, and the measurement aim (who do we want to compare and for what specific purpose?) should be taken into account. If country-specific item parameters are estimated in a large and representative sample of the general population of the target country, they will have optimal validity for application in the local population (comparing patients within in country). Furthermore, the interpretation of a local metric will be easier because a score of 50 will represent the average of the local population. However, *without rescaling,* the result will be different metrics across countries, hampering comparison of patients across countries (Option 2). In many cases, data are used for multiple purposes (monitoring individual patients, comparing aggregated data within and between countries, and using the data for research) all of which require comparability of scores. Using different metrics may lead to confusion and therefore using country-specific item parameters without rescaling is currently not a PROMIS-sanctioned option.

Countries could strive for a comparable metric by using similar calibration samples for estimating item parameters. The US PROMIS Pain Behavior item bank was calibrated on a combined sample of people from the general population and chronic pain patients; subsequently the metric was centered on the general population sample [30]. Data from a Dutch general population sample was not available at the start of this study, therefore the Dutch item parameters used in this study were based on chronic pain patients only. The results of this study, therefore, should not be used to draw conclusions about the application of the PROMIS Pain Behavior item bank in the Netherlands. It would be interesting to repeat this study trying to match the PROMIS item calibration procedure in a similar mix of general population vs. patients. Using similar calibration samples, however, would not guarantee a similar metric because it would not account for actual differences in the distribution of the construct (pain behavior) in different countries.

Having a common metric is essential for multinational studies, and for benchmarking health across countries. However, using the same set of item parameters across countries may lead to validity problems if there is language DIF (Appendix A). A currently accepted option by PROMIS for dealing with language DIF is the hybrid approach, where US item parameters are used for non-DIF items and country-specific item parameters, rescaled to the US metric, for DIF items (Option 3). An alternative option would be to use country-specific item parameters, all rescaled to the US metric (Option 4). This option is currently not recommended by PROMIS but may be an attractive approach for PROMIS users outside of the US. Further study is required to identify conditions in which rescaled country-specific CAT T-scores may provide enhanced validity and precision relative to the currently-recommended hybrid approach. We also recommend studying the impact of using different item parameters on study designs (eg, the power to detect differences) and clinical decisions.

Accurate rescaling depends upon identifying which items have DIF and which items are relatively DIF-free. This is challenging because whether an item is flagged for DIF depends on the statistical procedure used, sample variation, sample size, score distribution, DIF criteria, and the characteristics investigated (eg, country, sex, age, diagnosis) [38–44]. Several authors have recommended applying multiple DIF methods to flag items that will likely have "moderate" or "large" DIF [38,45–47]. However, determining when and how much DIF matters depends on the cause of DIF and the intended application of the measure; thus, qualitative judgments may be required [41–43]. DIF with bias in different directions may cancel out across the whole bank, but perhaps not in CAT or short form administrations. When studying relations between variables within a country, translation DIF may not be a problem, while DIF caused by multidimensionality could be problematic [41]. Moreover, it has been argued that most available DIF methods can detect DIF but cannot identify the DIF items due to parameter identification issues [44]. We recommend more research on the development of robust DIF methods and criteria.

Limited data are available on mean T-scores in general populations outside of the US. Three studies found differences smaller than 3 T-score points (range 0.1–4.2) between general populations from Germany, France, UK, Spain, and the Netherlands, compared to the US [24,28,29]. One study found a mean T-score difference larger than 5 T-score points for PROMIS depression in a German general population sample [20]. A way to ensure that 50 is the average for all countries could be to develop a world metric, where the mean of the world population is set to 50 (Option 5, Table 1). Option 5 could not be tested in this study because a set of "world" item parameters does not yet exist. However, it is questionable whether this is achievable and whether a "world average" would be meaningful.

The PROMIS metric was centered in 2000 [48,49] and given the passage of time and changes in the US population, a score of 50 on the US metric may no longer represent the mean score of the non-US population. Perhaps we should move away from referring to 50 as the mean score of the US population. However, it may still be helpful for users to consider 50 as a reasonable center score, even if there is a drift in the original reference sample and even if the average might be slightly different across countries. In addition, other reference scores can facilitate interpretation of PROMIS scores. For example, country-specific means and percentiles could be reported (see examples in Appendix A). Also (country-specific) score ranges that correspond to within normal limits, mild, moderate, and severe pain behavior (and other domains) could be helpful. We recommend more research to facilitate score interpretations of PROMIS measures across populations.

Another suggestion could be to use country-specific item parameters for local applications and use US parameters (or rescale scores to the US metric) for international publications and studies that require international comparisons (Option 6, Table 1). However, this carries a risk of misinterpretation of scores from reports that may not carefully specify how the different types of PROMIS T-scores were obtained.

Although this study does not provide evidence to reject the current DIF-hybrid approach, we suggest there is value in continued study of alternate approaches to choosing item parameter sets for non-English languages, and for publication standards to ensure consistency in reporting PROMIS T-scores across populations and countries.

This study was carefully designed to compare different sets of item parameters in a simulation study based on real response data of a large sample of patients. However, our conclusions are based on only one simulation study, using one item bank, in one patient population, and applying one DIF method. Our study should be replicated with other item banks, other DIF methods and criteria, and in other countries.

We chose the Stocking-Lord method for linking, rather than other options, such as multiple-group calibration [50,51], which is currently used, for example, in the educational field in the Programme for International Student

Assessment (PISA), a cyclical international testing program comparing multiple countries on science, reading, and mathematics (Programme for International Student Assessment [52] 2020). An advantage of multi-group calibration over Stocking-Lord rescaling is that standard errors around the item parameters can be estimated. It also would have the advantage of using available data across multiple languages concurrently in estimating parameters for each group. This approach, however, is impractical at this stage for PROMIS, because of the complexities of the original design and calibration analysis centered on the US Census. Furthermore, PROMIS CATs and short forms are constantly being administered and reported clinically and in studies. Furthermore, PROMIS item parameters need to be maintained for more than 100 items banks. Therefore, PROMIS parameters (and the resulting scoring algorithm) cannot be adjusted or re-estimated regularly. The Stocking-Lord method has been used in several other PROMIS studies [22,23,53]. Both the Stocking-Lord method and multiple-group calibration are widely discussed and compared in the linking literature [25–27]. While each method has advantages and disadvantages, Lee and Lee showed that the Stocking-Lord method produces similar parameters as multiple group calibration in situations as in this study where there are only common items [27].

In conclusion, we found negligible mean differences between US CAT T-scores and hybrid or country-specific CAT-T-scores when scores were placed on the same metric, but the choice of item parameters can be consequential for individual patient scores. Use of US item parameters has both advantages and disadvantages as does the use of country-specific item parameters. Further study can help identify contexts in which alternative calibration and centering approaches can improve precision while ensuring consistency of reporting of PROMIS T-scores across populations and countries.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors contribution

Caroline Terwee: Conceptualization, Methodology, Writing - original draft, Visualization. Martine Crins: Conceptualization, Methodology, Investigation, Resources, Writing - review & editing. Leo Roorda: Conceptualization, Methodology, Writing - review & editing. Karon Cook: Methodology, Writing - review & editing. David Cella: Methodology, Writing - review & editing. Niels Smits: Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing. Benjamin Schalet: Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing, Visualization.

Acknowledgments

The empirical study of which the data was used in our simulation study, was approved by the local institutional review board (of Slotervaart and Reade, the Netherlands).

Appendix A. Options for item parameter use in PROMIS CATs outside the US

Option	Description	Metric	Validity	Interpretation	Example
1	Use US item parameters in the non-US country (this is the default for PROMIS).	All T-scores across language-versions will be presented on the same (US) metric. One set of item parameters needs to be maintained.	If there is language-DIF, the US item parameters may not be valid for non-US populations (scores will be biased) because US item parameters may not be accurate for non-US populations, or the assumed dimensionality of the item bank may not apply to the non-US population (scores may not represent the same construct).	US-based norm scores: A T-score of 50 represents the mean score of the US population. All item banks have the same interpretation in the US population ("50 is the average"). However, a T-score of 50 may have a different meaning on each item bank in non-US populations because the mean T-score may be different for each item bank.	Donald (American citizen) has a T-score of 50 on the US Pain Behavior metric. Mark (Dutch citizen) has a T-score of 45 on the US Pain Behavior Metric. Donald's T-score is equal to the mean T-score of the US population (50th US percentile). Mark's T-score is 0.5 SD lower than the mean T-score of the US population (45th US percentile). The average Pain Behavior T-score in the Netherlands is 45 on the US metric. Donald's T-score is 0.5 SD higher than the mean T-score of the Dutch population (55th Dutch percentile). Mark's T-score is equal to the mean T-score of the Dutch population (50th Dutch percentile). The T-scores of Donald and Mark are different on the US metric, but they represent the same percentile of their local population.
2	Use country-specific item parameters.	The correlation between T-scores based on US versus country-specific item parameters will be high because the item scores are identical. However, T-scores from different countries will not be on the same metric and cannot be interpreted in the same way because the metrics may have a different mean and SD (see Fig. 1). Under this option, pooling of multi-language study data, such as with international clinical trials, would be inappropriate. Different sets of item parameters need to be maintained. The metric used to express scores should be reported in publications.	Country-specific item parameters, when estimated in a large and representative sample of the general population of the target country have optimal accuracy for the local population.	Local population-based norm scores: If the item parameters are centered on a representative sample of the local population, a T-score of 50 will represent the average of the local population in each country. All item banks have the same interpretation across countries ("50 is the average"). All scores are expressed as deviations from the local mean.	Donald has a T-score of 50 on the US Pain Behavior metric (50th US percentile). Donald has a T-score of 55 on the Dutch Pain Behavior metric (55th Dutch percentile). Mark has a T-score of 45 on the US Pain Behavior Metric (45th US percentile) Mark has a T-score of 50 on the Dutch Pain Behavior metric (50th Dutch percentile). The T-scores of Donald and Mark are the same on the country-specific metric but different on the US metric.

(continued on next page)

Option	Description	Metric	Validity	Interpretation	Example
3	Hybrid approach: Use US item parameters for non-DIF items and use country-specific item parameters for items with language DIF. The country-specific item parameters of the DIF items are rescaled on the US metric. This is currently recommended for PROMIS measures in case of substantial DIF, and is used for some of the Spanish item banks [22,23].	All T-scores across language-versions will be presented on the same (US) metric. For DIF items, different sets of item parameters need to be maintained. The item parameters used to obtain scores should be reported in publications.	The item parameters for the DIF items will be more accurate for non-US populations. However, this option assumes that it is possible to unambiguously determine which items show DIF, which has been questioned. DIF identification depends on the statistical procedure used, sample variation, sample size, and DIF criteria (the number of DIF items may be underestimated). Determining when DIF matters and how much is highly contextual and may require qualitative judgment [41–44].	Similar to Option 1, a score of 50 may not represent the mean score of the non-US population.	Similar to Option 1.
4	Use country-specific item parameters for all items and rescale all parameters to the US metric.	This option preserves the relative ordering of parameters for all items from the non-US calibration. All T-scores across language-versions will be presented on the same (US) metric. Different sets of item parameters need to be maintained. The item parameters used to obtain scores should be reported in publications.	The item parameters are most accurate for the local population. It is assumed that all items have some degree of DIF. However, the rescaling methodology still depends on the identification of DIF and non-DIF items because a set of anchor items must be identified that are assumed to have no DIF [33]. (see also Option 3).	As with Options 1 and 3, a score of 50 may not represent the mean score of the non-US population.	Similar to Option 1.

(*continued on next page*)

Option	Description	Metric	Validity	Interpretation	Example
5	Use ‘world’ item parameters.	All T-scores across language-versions will be presented on the same (world) metric. Requires a substantial amount of data from multiple countries. One set of item parameters needs to be maintained.	Similar to Option 1. Data will likely not be representative for the whole world. There is a high probability of language DIF and differences in dimensionality.	‘World’-based norm scores: A T-score of 50 represents the mean score of the ‘world’ population. All item banks have the same interpretation (“50 is the ‘world’ average“). Questionable whether this ‘world’ average would be meaningful. A T-score of 50 may have a different meaning on each item bank in every country. Each item bank may have a different interpretation because the mean T-score is different for each item bank in each country.	Similar to Option 1. Similar T-scores on the ẃorld´metric will represent different percentiles of local populations.
6	Use country-specific item parameters for local applications and rescale scores to the US metric for international comparisons.	The metric will be different across countries. Different sets of item parameters need to be maintained. The item parameters used to obtain scores should be reported in publications and it should be reported if rescaling was applied.	There is a risk of misinterpretation of scores if item parameters and rescaling is not (correctly) reported.	Similar to Option 2 for local applications. Similar to Option 4 for international comparisons.	Similar to Option 2 for local applications. Similar to Option 4 for international comparisons.

Appendix B. Rescaling using the Stocking-Lord method

The Stocking-Lord (SL) rescaling method is adopted from the equating and linking literature [26,33,54]. The SL method allows researchers to put two sets of item calibrations that are derived separately (from two different samples) on the same scale. This is done by calculating multiplicative and additive constants that minimize the squared difference between two test characteristic curves (TCCs). In linking situations, this technique is often applied in common-item nonequivalent group designs [27]. In those designs, SL constants are derived from common items that have two sets of parameters obtained from separate samples. The SL constants are then applied to the unique items (administered to only one sample), so that these unique items will then be on the same scale.

In the present case, we only have common items, and no new items are actually linked. The separate calibrations we started with are: (1) the established operational parameters calibrated from the US sample, which we obtained from help@healthmeasures.net, and (2) the parameters we calibrated from the Dutch sample. The SL method was used to place parameters obtained from calibrations in a Dutch sample on the US metric. The procedure was applied as follows: DIF analysis was previously performed to identify which items show DIF [18]. The 25 DIF-free items were then used as anchor items to calculate linking coefficients (see below); these constants are then used to transform the Dutch item parameters for the six DIF items to the US metric (Option 3, hybrid approach). In the end, all items are scaled to the US metric: the 25 non-DIF items simply retain the US item parameters, while the six DIF items have parameters obtained from the Dutch sample that are rescaled to the US metric via SL constants. In the US, the US set of parameters are used in scoring the instrument; in the Netherlands, the parameter set from this hybrid approach are used for scoring. All scores are comparable because they are calculated on the same scale. For Option 4, the same SL constants were used as those calculated above with 25 DIF-free items. The difference is that these constants are then applied to all item parameters estimated from the Dutch sample, not only the DIF items. In this way, all item parameter estimates retain information specific to the Dutch population, yet rescaled to the US metric, to enable direct score comparisons.

We used the equate function in the Lordif package (version 0.3-3) in R to compute two linear transformation constants (A and B) that minimizes the sum over patients of the squared differences between the two TCCs [55]. The Stocking-Lord constants were calculated using the 25 out of 31 DIF-free items. These constants were then used to transform the Dutch discrimination (a) and location (b) parameters estimated for Option 2 into new item parameters (a_{new} and b_{new}) on the US metric, for the six DIF items in Option 3 and for all items in Option 4, according to these formulas:

$$a_{new} = \frac{a}{A}$$
$$b_{new} = (A*b) + B$$

References

[1] Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36:648–62.

[2] Chang CH. Patient-reported outcomes measurement and management with innovative methodologies and technologies. Qual Life Res 2007;16(Suppl 1):157–66.

[3] Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16(Suppl 1):5–18.

[4] Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7:23–35.

[5] Reeve BB, Burke LB, Chiang YP, Clauser SB, Colpe LJ, Elias JW, et al. Enhancing measurement in health outcomes research supported by Agencies within the US Department of Health and Human Services. Qual Life Res 2007;16(Suppl 1):175–86.

[6] Embretsen SE, Reise SP. Item Response Theory for psychologists. New York: Psychology Press; 2000.

[7] Thissen D, Wainer H. Test Scoring. New York: Routledge; 2001.

[8] Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? Health Serv Res 2005;40:1694–711.

[9] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self--reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63:1179–94.

[10] Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). Arthritis Care Res (Hoboken) 2011;63(Suppl 11):S486–90.

[11] Patel AA. Patient-Reported Outcome Measures: The Promise of PROMIS. J Am Acad Orthopaed Surg 2016;24:743.

[12] Witter JP. The promise of patient-reported outcomes measurement information system-turning theory into reality: a uniform approach to patient-reported outcomes across rheumatic diseases. Rheum Dis Clin North Am 2016;42:377–94.

[13] PROMIS Statistical Center Working Group. The Patient-Reported Outcomes Measurement Information System (PROMIS(R)) Perspective on: Universally-Relevant vs. Disease-Attributed Scales. http://wwwhealthmeasuresnet/images/PROMIS/Universally-Relevant_vs_Disease-Attributed_2014-2-12_final508pdf. 2014.

[14] xxx 2020 http://www.healthmeasures.net/score-and-interpret/interpret-scores/promis.

[15] Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res 2007;16(Suppl 1):33–42.

[16] xxx 2020 http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf.

[17] Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Chronic Pain. PLoS One 2015;10:e0134094.

[18] Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration of the Dutch-Flemish PROMIS Pain Behavior item bank in patients with chronic pain. Eur J Pain 2016;20:284–96.

[19] Crins MHP, Terwee CB, Klausch T, Smits N, de Vet HCW, Westhovens R, et al. The Dutch-Flemish PROMIS Physical Function

item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 2017;87:47–58.

[20] Fischer HF, Wahl I, Nolte S, Liegl G, Brahler E, Lowe B, et al. Language-related differential item functioning between English and German PROMIS Depression items is negligible. Int J Methods Psychiatr Res 2017;26:e1530.

[21] Hays RD, Calderon JL, Spritzer KL, Reise SP, Paz SH. Differential item functioning by language on the PROMIS(R) physical functioning items for children and adolescents. Qual Life Res 2018;27:235–47.

[22] Paz SH, Spritzer KL, Morales LS, Hays RD. Evaluation of the Patient-Reported Outcomes Information System (PROMIS(R)) Spanish-language physical functioning items. Qual Life Res 2013;22:1819–30.

[23] Paz SH, Spritzer KL, Reise SP, Hays RD. Differential item functioning of the patient-reported outcomes information system (PROMIS(R)) pain interference item bank by language (Spanish versus English). Qual Life Res 2017;26:1451–62.

[24] Fischer F, Gibbons C, Coste J, Valderas JM, Rose M, Leplege A. Measurement invariance and general population reference values of the PROMIS Profile 29 in the UK, France, and Germany. Qual Life Res 2018;27:999–1014.

[25] von Davier M, von Davier AA. Statistical models for test equating, scaling, and linking. New York: Springer; 2010.

[26] Kolen MJ, Brennan RL. Test equating, scaling, and linking: methods and practices. Springer Science and Business Media; 2014.

[27] Lee WC, Lee G. IRT linking and equating. The Wiley Handbook of Psychometric Testing: a multidisciplinary reference on survey. Scale and Test Development; 2018. p. 639–73.

[28] Terwee CB, Crins MHP, Boers M, de Vet HCW, Roorda LD. Validation of two PROMIS item banks for measuring social participation in the Dutch general population. Qual Life Res 2019;28:211–20.

[29] Vilagut G, Forero CG, Castro-Rodriguez JI, Olariu E, Barbaglia G, Astals M, et al. Measurement equivalence of PROMIS depression in Spain and the United States. Psychol Assess 2019;31:248–264.

[30] Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain 2009;146:158–69.

[31] Chalmers P. mirt: a multidimensional item response theory package for the R environment. J Statist Softw 2012;48:1–29.

[32] R Core Team A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

[33] Stocking M.L., Lord F.M. Developing a common metric in item response theory applied psychological measurement. 1983;7:201-10.

[34] Choi SW. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. Appl Psychol Measur 2009;33:644–5.

[35] Chen CX, Kroenke K, Stump TE, Kean J, Carpenter JS, Krebs EE, et al. Estimating minimally important differences for the PROMIS pain interference scales: results from 3 randomized clinical trials. Pain 2018;159:775–82.

[36] Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis 2015;74:104–7.

[37] Purvis TE, Neuman BJ, Riley LH, 3rd, Skolasky RL. Discriminant ability, concurrent validity, and responsiveness of PROMIS health domains among patients with lumbar degenerative disease undergoing decompression with or without arthrodesis. Spine 2018;43:1512–20.

[38] Sireci SG, Rios JA. Decisions that make a difference in detecting differential item functioning. Educ Res Eval 2013;19:170–87.

[39] Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. Med Care 2006;44:S39–49.

[40] Teresi J.A., Jones R.N. Methodological issues in examining measurement equivalence in patient reported outcomes measures: methods overview to the two-part series, "measurement equivalence of the patient reported outcomes measurement information system(R) (PROMIS(R)) Short Forms". Psychological test and assessment modeling. 2016;58:37-78.

[41] Borsboom D. When does measurement invariance matter? Med Care 2006;44:S176–81.

[42] Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med Care 2006;44:S115–23.

[43] Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res 2007;16(Suppl 1):69–84.

[44] Bechger TM, Maris G. A statistical test for diffetential item pair functioning. Psychometrika 2015;80:317–40.

[45] Gomez-Benito J, Sireci S, Padilla JL, Hidalgo MD, Benitez I. Differential item functioning: beyond validity evidence based on internal structure. Psicothema 2018;30:104–9.

[46] Hambleton RK. Good practices for identifying differential item functioning. Med Care 2006;44:S182–S1S8.

[47] Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is scored: an illustration with center for epidemiologic studies depression scale. Educ Psychol Measur 2003;63:65–74.

[48] Garratt A, Stavem K. Measurement properties and normative data for the Norwegian SF-36: results from a general population survey. Health Qual Life Outcomes 2017;15:51.

[49] Maglinte GA, Hays RD, Kaplan RM. US general population norms for telephone administration of the SF-36v2. J Clin Epidemiol 2012;65:497–502.

[50] Lee WC, Ban JC. A comparison of IRT linking procedures. Appl Measur Educ 2009;23:23–48.

[51] Hansen M, Cai L, Stucky BD, Tucker JS, Shadel WG, Edelen MO. Methodology for developing and evaluating the PROMIS smoking item banks. Nicotine Tobacco Res 2014;16(Suppl 3):S175–89.

[52] Programme for International Student Assessment (PISA). 2020. https://www.oecd.org/pisa/.

[53] Tulsky DS, Kisala PA, Victorson D, Choi SW, Gershon R, Heinemann AW, et al. Methodology for the development and calibration of the SCI-QOL item banks. J Spinal Cord Med 2015;38:270–87.

[54] Kang T, Petersen NS. Linking Item Parameters to a Base Scale. ACT Inc; 2009.

[55] Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39:1–30.