Journal of Physical Activity and Health, 2012, 9, 29-38

Response Conversion for Improving Comparability of International Physical Activity Data

Marijke Hopman-Rock, Elise Dusseldorp, Astrid Chorus, Gert Jacobusse, Alfred Ruetten, and Stef van Buuren

Background: Many questionnaires for measuring physical activity (PA) exist. This complicates the comparison of outcomes. ***Methods:*** In 8 European countries, PA was measured in random samples of 600 persons, using the IPAQ as a 'bridge' to historical sets of country-specific questions. We assume that a unidimensional scale of PA ability exists on which items and respondents can be placed, irrespective of country, culture, background factors, or measurement instrument. Response Conversion (RC) based on Item Response Theory (IRT) was used to estimate such a common PA scale, to compare PA levels between countries, and to create a conversion key. Comparisons were made with Eurobarometer (IPAQ) data. ***Results:*** Appropriateness of IRT was supported by the existence of a strong first dimension established by principal component analysis. The IRT analysis resulted in 1 common PA scale with a reasonable fit and face validity. However, evidence for cultural bias (Differential Item Functioning, DIF) was found in all IPAQ items. This result made actual comparison between countries difficult. ***Conclusions:*** Response Conversion can improve comparability in the field of PA. RC needs common items that are culturally unbiased. Wide-scale use of RC awaits measures that are more culturally invariant (such as international accelerometer data).

Keywords: item response theory, cultural bias, differential item functioning, IPAQ

Researchers in the field of physical activity (PA) and (public) health gather data on PA by a wide variety of measurement instruments. A few broad classes of such instruments are: questionnaires, accelerometers, and field tests. Even within these broad classes, the variability of instruments is great and likely to increase over time as new instruments are being developed. Each new generation of instruments attempts to remedy the deficiencies of older ones, ideally converging into tools that are free of the most obvious flaws. On the other hand, the actual situation is nowhere near this ideal. Different instruments express PA in different units (frequency, duration, intensity, energy expenditure), and there is no easy way to convert one measure into the other. For example, both the International Physical Activity Questionnaire (IPAQ)[1] and the Short Questionnaire to Assess Health-Enhancing Physical Activity (HEPA)[2] aspire to measure PA in community samples. Yet, one cannot simply convert or compare their scores. There is also no accepted validation paradigm (except research with the expensive doubly labeled water method) by which we judge the quality and comparability of the outcomes of an instrument. This hampers progress in the area of international comparability of PA levels of communities and individuals.

In this paper, we present a technique known as Response Conversion (RC).[3–5] This method is based on firm theoretical principles and detects and repairs comparability problems that arise out of differences in the formulation of survey questions and response categories. RC attempts to translate responses obtained on the same topic but with different questions into scores on a common underlying unidimensional scale. Scores on this scale are meant to be comparable, although they were derived from different questionnaires measured in different populations at different times.

RC is based on the assumption that instruments measure the same continuum (eg, PA), but do so in different ways. There is a clear analogy to physics, where the distance between 2 points may be measured by a ruler, by a difference between viewing angles, or by the time taken to reflect sound. As long as one knows how the resulting values (cm, degrees, seconds) can be expressed in terms of a common distance unit, it is possible to scale the outcome on the same continuum. The RC technique is a method to unveil such conversion rules using a linkage diagram with 'bridge' items (ie, common items) as a start to result in a routinely applicable conversion key. RC is a test linking technique based on Item Response Theory. (For a detailed description of test equating and test linking

Hopman-Rock, Dusseldorp, Chorus, and Jacobusse are with the Dept of Prevention and Healthcare, TNO Quality of Life, Leiden, Netherlands. Ruetten is with the Dept of Sport Science, Insitute of Sport Science and Sport, Friedrich-Alexander University, Erlangen, Germany. van Buuren is with the Dept of Methodology and Statistics, University of Utrecht, Utrecht, Netherlands.

in the field of physical activity, see Zhu[6,7]). RC has been applied successfully for dressing and walking disability.[8,9] RC is appropriate for linking questionnaire items (that are assumed to measure the same continuum) from several separate databases, where subjects completed at least 2 items, and where databases are linked by bridge items.

The current approach within the field of PA research is to express activity in MET-minutes[10,11] and then linking test data if possible. Use of METs-values requires extra analyses and computations and some questions are not suitable to transit in METs values. In addition, it is not possible to correct for cultural bias in the outcomes after the linking.

This paper explores the use of RC in the field of PA measurement. We will apply RC to an international European data set, the EUPASS data,[12] and compare the results to the Eurobarometer study.[13] RC has facilities to correct for cultural bias (technically known as Differential Item Functioning, DIF) This paper evaluates the potential of RC for current items on PA, examining cultural bias, and validates results in an example comparing PA in different age groups.

Methods

Study Design and Measures

The EUPASS study gathered data in the year 2000 using the IPAQ[1] in 8 different countries (Belgium, Finland, France, Germany, The Netherlands, United Kingdom, Italy, and Spain). In addition, the EUPASS study included existing country-specific PA questions. Within each of the 8 countries, about 600 adult respondents were randomly selected (total database N = 4976), and interviewed using a computer-aided telephone interview. Precise details can be found elsewhere.[12]

The IPAQ questions were asked in all countries. The IPAQ was explicitly designed to be cross-culturally equivalent.[1] In total, 9 IPAQ items and 40 national items were available for analysis. Conforming to IPAQ instructions,[10] we computed continuous compound variables for vigorous activity, moderate activity, and walking, using the items "hours a day," "minutes a day," and "days per week." We removed subjects with unlikely answers (ie, subjects who report more than 3 hours of vigorous activity, moderate activity, or walking a day). This resulted in a substantial reduction of the sample size to N = 3597 [lost for analyses is 27.7%; this value varied between countries from 11.7% (Italy) to 38.5% (Belgium)]. Subjects that remained in the sample had slightly worse health (16%, against dropout group 22%; Chisquare 39.5, df = 4, P = .00), were more female (58% against 52% in dropout group; Chisquare 15.2, df = 1, P = .00) and were slightly older (mean age 46 years against 44 years in dropout group; F = 14.5, P = .00). Continuous compound variables were expressed in total minutes per week. Two variables measuring sitting behavior either on a weekday or on a weekend day were merged together according to the IPAQ manual. Table 1 contains an overview of the 49 items in the form of a diagram that shows which items were administered in which countries and how the linkage was established.

Table 1 Linkage Diagram of the EUPASS Data; Overview of All Items and for Which Countries Each Item Is Applicable

			Country							
Questionnaire items	Ro[a]	Ru[a]	BE	FI	GE	IT	NL	UK	SP	FR
1. IPAQ At what pace usually walk*	3	3	X	X	X	X	X	X	X	X
2. IPAQ How much PA in place of work last 7 days	3	3	X	X	X	X	X	X	X	X
3. IPAQ How much PA for purpose of transportation last 7 days	3	3	X	X	X	X	X	X	X	X
4. IPAQ How much PA in and around home last 7 days*	3	3	X	X	X	X	X	X	X	X
5. IPAQ How much PA recreation, sport, leisure time	3	3	X	X	X	X	X	X	X	X
6. IPAQ how much time in usual week doing vigorous PA	C	7	X	X	X	X	X	X	X	X
7. IPAQ how much time in usual week doing moderate PA	C	7	X	X	X	X	X	X	X	X
8. IPAQ how much time in total you spend on walking	C	7	X	X	X	X	X	X	X	X
9. IPAQ sitting: sum in minutes for 1 day REVERSED[b]*	C	5	X	X	X	X	X	X	X	X
10. On how many days sweating at least 1 time per week	8	7	X							
11. Leisure time PA for at least half an hour (at least l sw)	7	7		X						
12. Minutes a day walking, running or riding a bicycle to/of work	6	5		X						
13. Demanding job physically	4	4		X						
14. How much exercise or PA in free time	4	4		X						

(continued)

Table 1 *(continued)*

Questionnaire items	Ro[a]	Ru[a]	Country							
			BE	FI	GE	IT	NL	UK	SP	FR
15. How often engaged in sports/ strenuous activities	5	5			X					
16. Get out of breath after climbing 3 floors REVERSED[b]	2	2			X					
17. How often do you participate in sports	5	5			X					
18. Time spend per day sleeping (Monday to Friday) REVERSED[b]	C	5			X					
19. Time spend per day sitting (M-F) REVERSED[b]	C	7			X					
20. Time spend per day light activities* (M-F)	C	8			X					
21. Time spend per day moderate activities (M-F)	C	8			X					
22. Time spend per day strenuous activities (M-F)	C	8			X					
23. Time spend per day sleeping (Weekend) REVERSED[b]	C	5			X					
24. Time spend per day sitting (Weekend) REVERSED[b]	C	7			X					
25. Time spend per day light activities* (Weekend)	C	8			X					
26. Time spend per day moderate activities (Weekend)	C	8			X					
27. Time spend per day strenuous activities (Weekend)	C	8			X					
28. How long are you engaged in sports / strenuous act.	4	4			X					
29. Regular sporting activities in free time	2	2				X				
30. Occasional sporting activities in free time	2	2				X				
31. How many month in total	C	4				X				
32. Consider all the sporting activities over past 12 months	6	6				X				
33. Any type of physical activity at least twice a year	4	4				X				
34. How many activities	5	5				X				
35. Sporting activities requiring payment	2	2				X				
36. Practice requiring payment (lessons)	2	2				X				
37. Annual (periodic) fee for sport club	2	2				X				
38. Number of times PA participation past 14 days	c	5					X			
39. How many times a day do you walk	c	5					X			
40. How many times sports or exercise	c	5					X			
41. Sum of minutes heavy PA yesterday	c	7					X			
42. Sum of minutes moderate PA yesterday	c	7					X			
43. Sum of minutes light PA yesterday*	c	7					X			
44. How many sports	5	3					X			
45. Did you sport yesterday	2	2					X			
46. Gardening, dig or building work done in the past 4 weeks*	2	2						X		
47. Any exercise or sport during the last 4 weeks	2	2						X		
48. Was the effort or activity usually makes you out of breath	2	2						X		
49. Walking of a quarter of a mile done locally or away from	2	2						X		

[a] Ro: Response scale of original items, the number of response categories is given, and 'c' is used for continuous items; Ru: number of response categories used in the Rasch model; BE, Belgium; FI, Finland; GE, Germany, IT, Italy; NL, The Netherlands; UK, United Kingdom; SP, Spain; FR, France.

[b] Items are coded such that a lower score reflects less physical activity (for example, more sitting).

* Removed from Rasch analysis.

Note. For overview of all questionnaires, see: http://www.public-health.tu-dresden.de/dotnetnuke3/Portals/5/Projects/EUPASS/appendix%20b.pdf.

All PA variables were coded such that a higher value reflected a higher activity level. For example, the coding of sitting items was reversed so that much sitting indicated low PA. Continuous PA variables were categorized for application of the Rasch model. For example, the IPAQ items expressed in total minutes of activity per week were categorized into average number of half hours per day (ie, divided by 210 and then rounded). The number of categories of some categorical variables was reduced, because of low frequencies in the extreme values. For example, categories 6 and 7 of the item "on how many days sweating at least 1 time per week" (item 10, Table 1) were merged to 1 category: 5 days or more. Thus, the number of categories for categorical measured variables varied between 5 and 8. The correlation between all categorized and corresponding original variables was always higher or equal to 0.90. Table 1 gives for each item the number of original response categories and the number of categories used in the IRT analysis.

On the basis of the IPAQ activity measures expressed in MET-minutes, the following 3 categories were computed by using the IPAQ manual: HEPA 1, 2, and 3. These categories reflect the percent of people at low (HEPA 1: sedentary/inactive), moderate (HEPA 2: not sufficient), and high levels (HEPA 3: sufficient) of PA (to accumulate to 100% of the population). The high level of PA is similar to the "sufficient total activity" level used in the Eurobarometer study.[13] This categorical representation of PA enables us to compare the results of the EUPASS data with the Eurobarometer data (please note that the EUPASS and the Eurobarometer study are separate studies, only similar in using the IPAQ in an European sample). In addition, we will compare it with the results from the RC analyses.

Statistical Analysis

We used the polytomous Rasch (IRT) model[14,15] to estimate the relative position (often interpreted as 'difficulty') of the items on the PA ability scale. The Rasch model describes the probability that a person responds into a category conditional on the location of the person on the continuum of PA (which can be interpreted as a scale that indicates how physically active a person is). The model has 1 or more difficulty parameters for each item (there are m difficulty parameters for an item with m+1 categories). For a dichotomous item (with categories 'no' and 'yes'), the difficulty parameter indicates how much ability a respondent needs to achieve a 50/50% chance of scoring 'yes.' An affirmative answer to a difficult question like 'do you go running for at least 3 hours a week' is generally associated with higher PA level than an affirmative response to an easier item like 'do you walk at least 1000 m a week.' For a polytomous item, the difficulty parameters can be considered as step difficulties associated with the transition from one category to the next.[14] A positive response, especially to a 'difficult' question, results in a higher respondent score on the PA scale. A negative response, especially to an easy question, results in a lower PA score. A better or more precise estimate of the ability of a person can be calculated from his or her responses to a series of items. We opted for the Rasch model since that model is the only one in which estimation of the difficulty parameters is independent of the distribution of PA in the reference population. Thus, the choice of the reference sample is not critical to parameter estimates.

An important assumption underlying the Rasch model is that items measure the same continuum (ie, that they are unidimensional). We checked this assumption by categorical principal components analysis using SPSS CATPCA[16] on the 9 IPAQ items, where we assumed ordered categories.

We used RUMM 2020[17] to estimate item difficulty parameters. The estimation method is based on the pairwise conditional approach, and has been described in detail by Andrich and Luo.[15] This approach generally works well with incomplete and sparse data.[15] Using the Bayes rule, the parameters can be used to calculate the RC key. The RC key is a simple table in which it is possible to transform the original category of a questionnaire into a place on the new PA scale. For additional information and examples of this procedure see Jacobusse et al.[18]

The reliability of the Rasch model was measured by the person separation index (also called the person separation reliability.[19] It range between 0 and 1, and the interpretation is similar to that of Cronbach's α. Furthermore, the item fit was measured by the fit residual statistic per item (given by RUMM). When the Rasch model is true, this measure follows a standard normal distribution (mean near 0, and standard deviation near 1); values higher than 2 indicate that unexpected deviations from the model occur.[20] Because of the large sample size of our study, we chose a more liberal criterion for misfit of items, that is, a standard residual greater than 3.5, and we excluded misfitting items from the analysis. Similarly as for items, respondents with a person-fit residual statistic greater than 3.5 were excluded from the final analysis. The IPAQ items acted as bridge items between the country samples (see Table 1). The assumption is that 9 bridge items measure PA in the same way in different countries (without cultural bias). If this assumption is false, the item has Differential Item Functioning (DIF),[21] and we cannot use it in its original form to equate items. We tested for DIF by ANOVA using a Type I error rate of $P < .001$ because of large sample size. If DIF was present, we refitted the data under the more relaxed model where deviating countries obtained separate difficulty parameter estimates, an action known as item splitting. DIF was tested again in the remaining items, until an acceptable solution was found without DIF. This procedure is known as the Stocking and Lord iterative procedure.[22]

Results

Table 2 summarizes the main characteristics of the selected sample (N = 3597) from the EUPASS data. The Netherlands and Belgium had older people than the other countries in the sample. Note that the responses to the

Table 2 Main Characteristics of the Respondents by Country (N = 3597)

	Country								
	Belgium	Finland	France	Germany	Italy	Netherl.	Spain	UK	Total
n respondents	376	379	390	384	530	467	500	571	3597
Mean age (SD)	51.5	46.4	40.3	42.8	44.3	51.3	46.1	43.8	45.8
	(18.1)	(15.3)	(16.7)	(15.2)	(15.9)	(18.8)	(18.4)	(17.0)	(17.4)
% Women	55	61	57	56	56	63	58	57	58
Mean BMI	24.5	25.0	23.0	24.3	24.2	24.6	24.4	24.3	24.3
Health (%)									
Very good/good	–	63	66	75	55	80	70	62	67
Fair/poor/bad	–	36	34	26	45	20	30	39	33
Occupation (%)									
Working	45	58	52	63	53	42	50	60	52
Retired	36	27	16	16	19	29	18	20	22
Other	19	15	16	22	29	29	33	20	27

Note. In Belgium, the general health question was not administered.

Abbreviations: Netherl., The Netherlands; UK, United Kingdom.

general health question also varied considerably across countries.

The first dimension based on categorical principal components analysis on the recoded IPAQ items explained 21.5% of the total variance. For comparison, a linear PCA with the same variables explained 20.7% of the variance. The percentage was 22.0% for the unrecoded continuous data, indicating that the loss by categorization is negligible. The first eigenvalue of the PCA solution was large compared with the second (1.9 vs. 1.1), and the eigenvalues other than the first were about the same size (between 1.1 and 0.7). These results pointed to a dominant first factor underlying the PA items.[6]

The infit statistics of the estimated Rasch model with all 49 items ranged from –6.6 to 5.9. After 4 item removal steps in which questions with a fit statistic over 3.5 were dropped, all items had a fit statistic lower than 3.5. The highest remaining infit statistic was 2.5. The following items were removed: 3 IPAQ items (items 1, 4, and 9; Table 1), 3 items referring to light activities (item 20, 25, and item 43), and 1 item measuring gardening, do-it-yourself, or building work (item 46). Most of these misfitting items had no or limited conceptual overlap with the other items. After removing these items, none of the persons had a person-fit statistic greater than 3.5. The overall goodness-of-fit index (the person-separation-index) of the Rasch model was reasonable: 0.68.[17] Figure 1 shows the distribution of the estimated common PA scale based on this model. The scale was transformed in such a way, that the mean score was set at 50 and a standard deviation of 10. Parameter estimates indicated that the easiest item (the item that most people were most likely to respond "yes" to) was "walking a quarter of a mile or more in the past 4 weeks, either locally or away from home" (no/yes; item 49, Table 1). The most difficult transition (this means only positively answered by respondents with relatively high ability levels) occurred for the item "sum of minutes heavy physical activity yesterday" (item 41): the transition from 0 to 10 minutes (category 1) to 10 to 40 minutes (category 2).

We investigated DIF of the remaining 6 IPAQ items. All 6 items had statistically significant DIF ($P < .001$), indicating that their interpretation differed across countries. Figure 2 shows how DIF between countries manifests itself in IPAQ question 'How much time in total you spend on walking" (item 8, Table 1). This item has 7 response categories (0 to 6), representing the average number of half hours walking per day. If there was no DIF, the response curves would be located closely to each other. For most countries, this is the case. However, it appears that persons from the Netherlands score consistently higher at the same level of PA (especially at the lower end of the scale). In other words, item 8 is more "easy" for the Dutch than for the other countries. The consequence is that we cannot use the responses on item 8 in the Netherlands in the same way as for the other countries. To correct for this, we estimated separate item parameters for the Netherlands. This item splitting procedure was performed for all items, until there was no significant DIF left.

To get more insight into the consequences of the item splitting procedure, we compared the results of the Rasch model without correcting for DIF (Figure 3, panel a) with those of the model while correcting for DIF (Figure 3, panel b). The black lines in the boxes in Figure 3 represent the country medians. Italy and UK have a median score below or equal to the overall median of 52.5 in both models. And additionally, The Netherlands and Belgium have a mean score below the overall mean of 50.0 in both models. So, the correction of DIF seems to have limited influence. The standardized difference in means is large (effect size $d = 0.7–0.8$) between Germany and United Kingdom (Figure 3, panel a), and between Spain and the United Kingdom (Figure 3, panels a and b). The

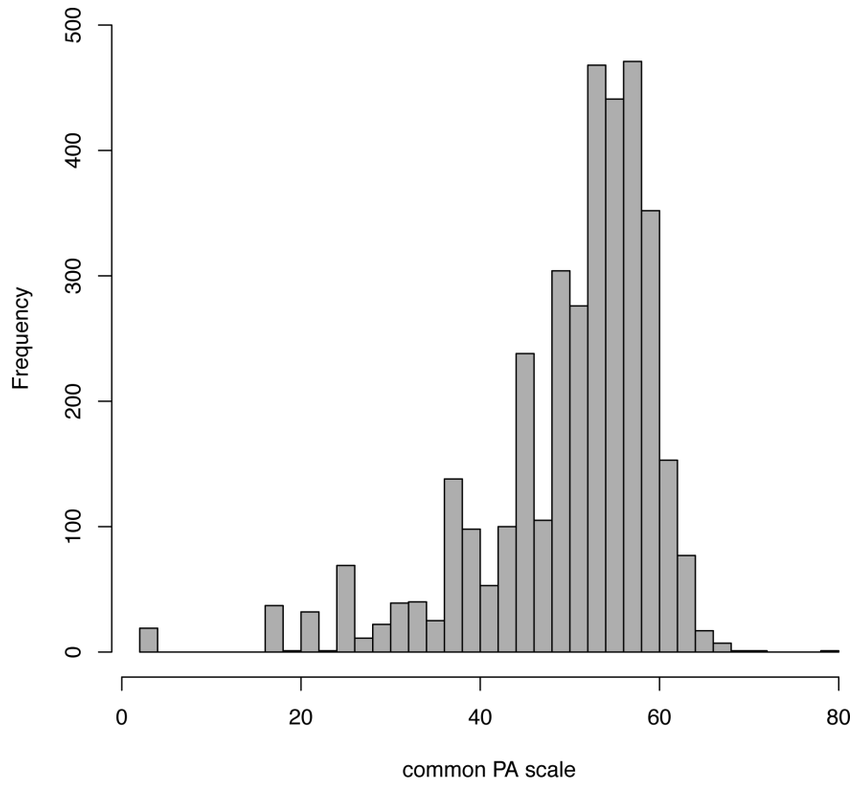

Figure 1 — Distribution of the individual ability scores on the common PA scale obtained by the Rasch analysis (N = 3579). Mean PA score is normalized as 50, and standard deviation is 10.

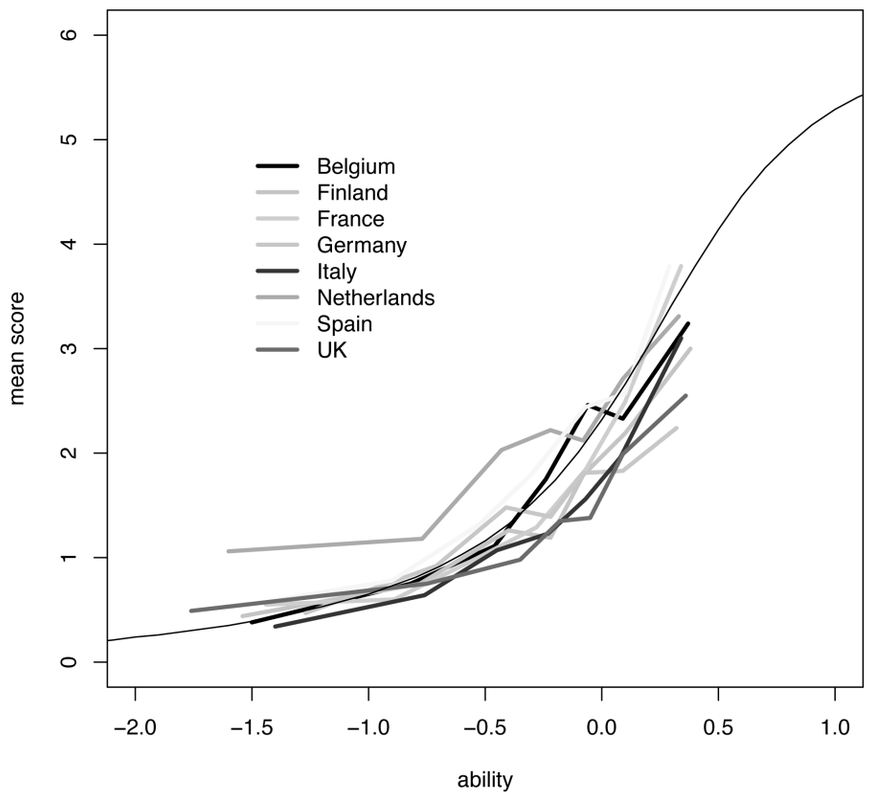

Figure 2 — Differential Item Functioning (DIF) indicating cultural differences in responding by country in IPAQ item: "How much time in total you spend on walking" (item 8, Table 1).

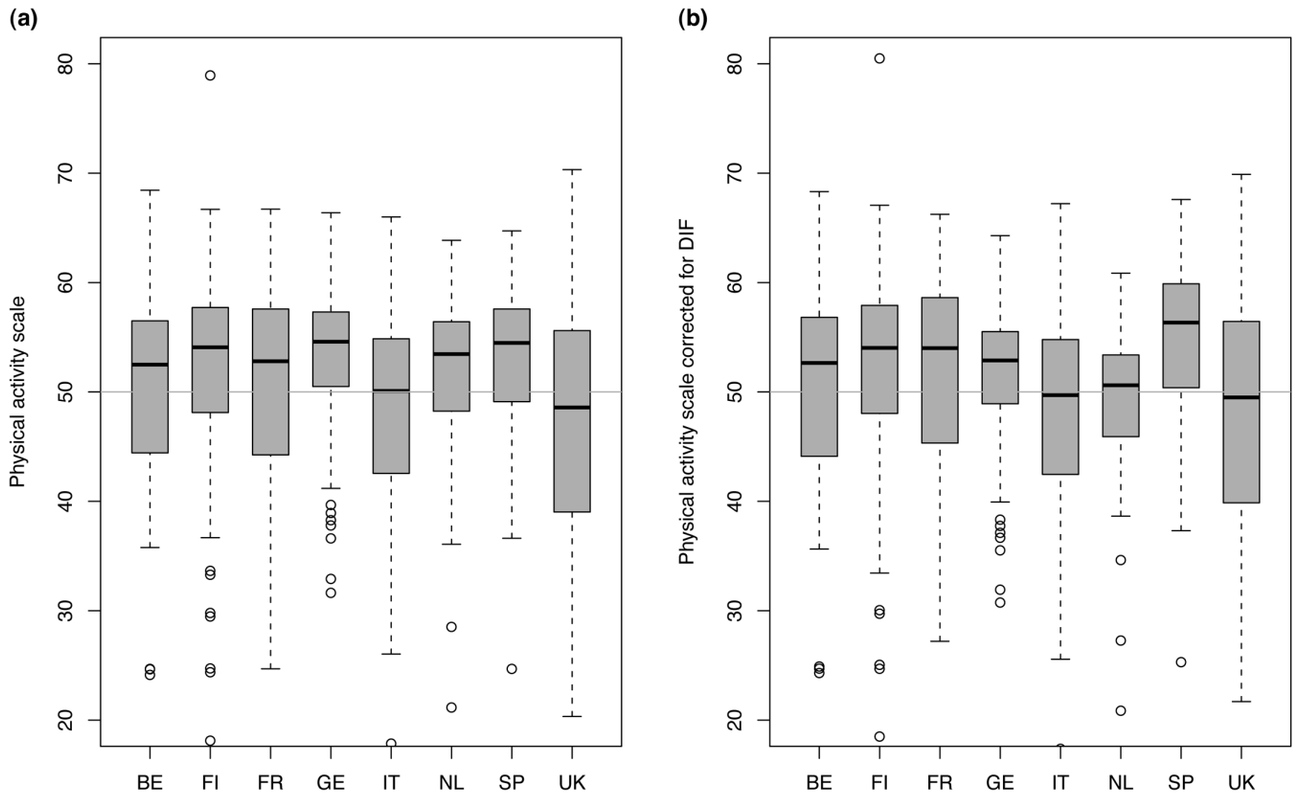

Figure 3 — Distribution of the common PA scale estimated by the Rasch model without correcting for DIF (panel a) and by the model with correcting for DIF (panel b). Both PA scales are normalized with mean 50, and standard deviation 10. A higher score means a higher level of physical activity.

other differences are moderate ($d = 0.5$–0.6) to negligible ($d = 0.1$). Most differences between country means are statistically significant, but not all. For example, the differences between the means of Finland and Germany and between those of the Netherlands and Italy are not significant for both models. The standard deviation of the scores (on both PA scales) was higher than 10 in the United Kingdom, Belgium, and the Netherlands.

Some countries were more responsible for DIF than others. The Netherlands showed DIF on 5 of the 6 IPAQ items, Italy on 4 items, and Spain, Finland, and Germany on 3 items each. The differences between the results of the model without DIF and with DIF were highest for Spain and the Netherlands. The mean PA score of Spain was lower in the model without DIF (Figure 3, panel a) than with DIF (Figure 3, panel b), and for the Netherlands the reverse was true.

We also compared the rank order of the countries to the prevalence of the HEPA categories estimated from the Eurobarometer study (collected in October to December 2002; see Table 3). Both Germany and Finland were in the top 3 of physically active countries, according to the continuous PA scale as well as according to the "sufficient activity" category (HEPA 3). This applies for both the Eurobarometer study and this study (Table 3). However, according to the common PA scale, Spain belongs also to this top 3 (Figure 3), whereas according to the HEPA 3, the Netherlands is one of the most physically active countries (Table 3). Spearman rank-order correlation between HEPA 3 of the EUPASS and HEPA 3 of the Eurobarometer was 0.55 ($P < .001$), indicating a moderate level of agreement.

With regard to the 3 least active countries, both representations of PA and both studies agree that Belgium and the United Kingdom fall in this category (Figure 3, panel a and Table 3, HEPA 1 columns). However, according to the common PA scale and HEPA 1 of this study, Italy is also one of the least active countries (Figure 3 and Table 3), whereas according to HEPA 1 of the Eurobarometer study, France is the least active country. Spearman rank-order correlation between HEPA 1 of the EUPASS and HEPA 1 of the Eurobarometer was 0.52, indicating a moderate level of agreement.

In general, we expect a decline in PA with age. With higher age, the mean PA score remains the same for France (Figure 4, panel a), shows a small decline for Spain, Finland, and Germany, and shows a large decline in the elderly (from 64 yrs old) for the Netherlands,

Table 3 Prevalence of 2 HEPA Categories of Physical Activity in the EUPASS Study and the Eurobarometer Study (Derived From Sjöström, 2006); Those Countries From the Eurobarometer Study Were Selected That Were Also Measured in This Study; Prevalences With Highest Rank Order (1) Are in Bold Face

	HEPA 1 (sedentary/inactive %) (rank)		HEPA 3 (sufficient PA %) (rank)	
	EUPASS	Eurobarometer	EUPASS	Eurobarometer
Belgium	30.1 (3)	39.8 (2)	31.9 (6)	25.0 (7)
Finland	18.2 (6)	23.8 (7)	45.1 (2)	32.5 (3)
France	25.6 (4)	**43.1 (1)**	37.9 (5)	24.1 (8)
Germany	15.1 (7)	24.1 (6)	**52.6 (1)**	40.2 (2)
Italy	**38.9 (1)**	35.3 (4)	18.1 (8)	25.8 (5)
Netherlands	22.9 (5)	19.3 (8)	44.3 (3)	**44.2 (1)**
Spain	10.8 (8)	31.2 (5)	39.0 (4)	25.2 (6)
United Kingdom	35.4 (2)	37.4 (3)	25.0 (7)	28.7 (4)

Note. EUPASS data uncorrected for DIF. Spearman correlation HEPA 3 for EUPASS and Eurobarometer = 0.55 (P = .00).

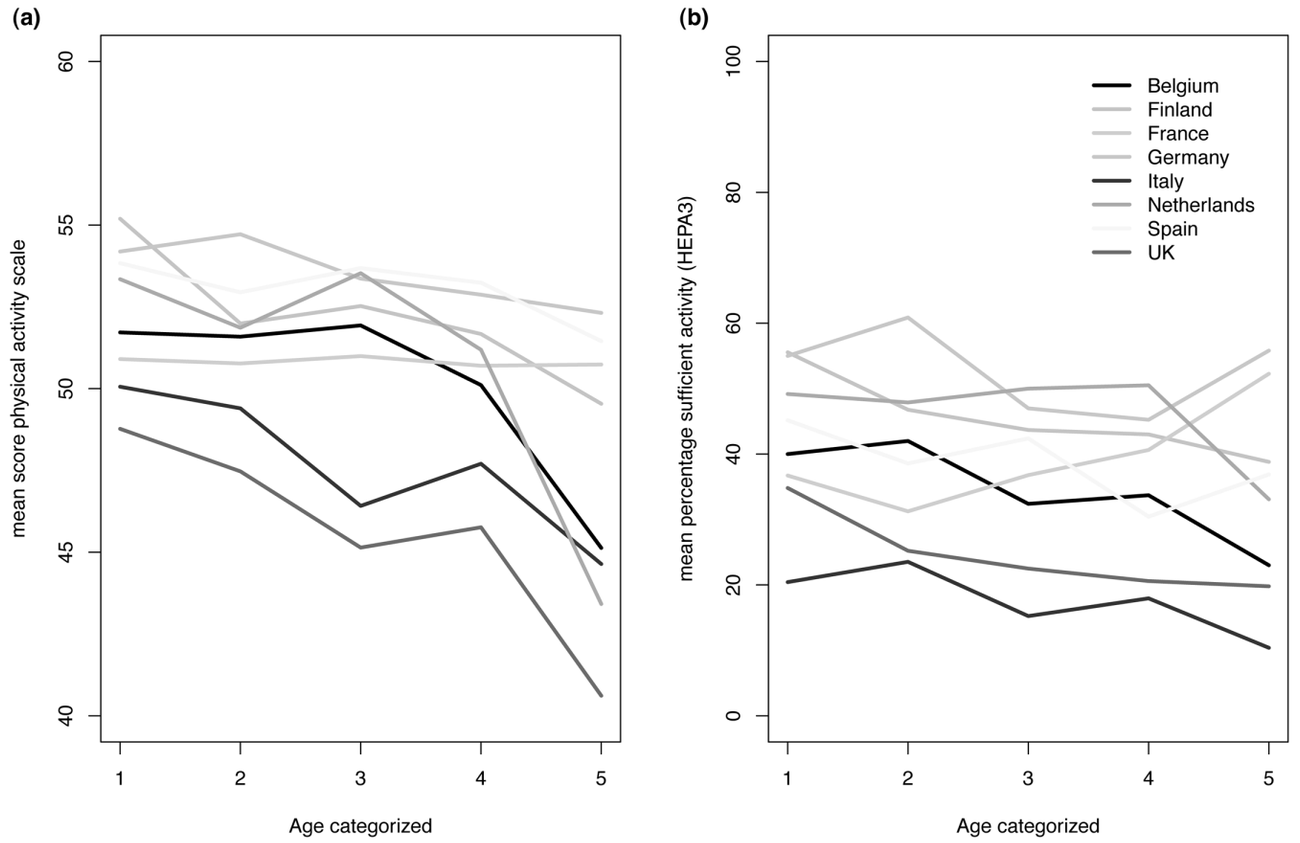

Figure 4 — Relationship between age and the PA scale (panel a) and between age and sufficient total activity (HEPA 3; panel b). Age is divided into the following categories: 1 = 15–29 yrs; 2 = 30–38 yrs; 3 = 39–49 yrs; 4 = 50–63 yrs; 5 = 64–93 yrs. Each category contains approximately 20% of the respondents.

Belgium, Italy, and the United Kingdom. Note that these trends are not present in the graph showing the relationship between age and HEPA 3 (Figure 4, panel b). These results suggest that the common PA scale could be a more sensitive measure of PA.

Discussion

This article introduced Response Conversion to improve international comparability of PA data. The PA data used in this study (the EUPASS data; 12) included both items from a relatively new instrument (IPAQ) as well as 'old' existing items from locally used questionnaires in 8 European countries.

Our results demonstrated that 1) the PA items included in the EUPASS study satisfied the assumption of unidimensionality, 2) a 'Physical Activity Scale' could be estimated by a Rasch model with reasonable fit, and 3) the Physical Activity Scale was more sensitive to the age-effect than the HEPA by showing a clear pattern of decreasing PA with age (thus showed face validity).

All IPAQ bridge items appeared to suffer from DIF. This means that items are not fully comparable between countries. For a similar conclusion see Bauman et al,[23] who included 20 worldwide countries using the IPAQ. Thus, the IPAQ might be less suitable for comparing populations from different countries because of cultural bias. This is disappointing because the IPAQ items were designed to be free of cultural bias. This suggests that it is difficult to develop items that are free of DIF. Another problem with the IPAQ was the relatively high loss of subjects that apparently had difficulty in understanding the questions. For example, 3.3% of the subjects indicated more than 7 hours of vigorous activity a day (up to 20 hours). We expect that respondents have interpreted this as activity per week. The difficulty of the questions was also observed by Heesch et al,[24] who performed a qualitative study in older people that completed the IPAQ. These authors emphasized that most items are very difficult to answer.

The ability to recognize and quantify DIF is a major methodological advance of RC. As the validity of the RC key relies on linked databases which are free from DIF, we have decided to withhold its publication until better quality data become available. One possibility is the GPAQ developed by the World Health Organization,[25] which will eventually be combined with international accelerometer data.

We refer to the common underlying scale resulting from RC as the "Physical Activity Scale." This scale is conceptually different from the energy expenditure scale, expressed in METs, that is often used to summarize IPAQ or other PA items. Physical activity is a somewhat more general concept than energy expenditure, relating to actual behaviors rather than the amount of energy required to perform these behaviors. An advantage of such a broader concept is that more PA indicators can be related to it (such as sitting behavior). This enables us to encompass a wider range of items, eventually resulting in increased measurement precision.

To check the validity of the data we made comparisons with data from the Eurobarometer study.[13] Country-specific percentages of activity categories as used by the Eurobarometer were in range with our findings. An unexpected finding was that compared with the other countries, the data from the Netherlands showed a low level of PA, especially in older people. In the Netherlands the year 2000 was—according to the national trend report—the year with the lowest level of PA in the older population (as measurement started in 2000). Since that time the national PA levels have been improved significantly for all age groups.[26]

We conclude that Response Conversion is a promising technique to improve comparability in the field of PA applying to both existing as well as new databases. The technique is able to integrate various components of PA into a common 'Physical Activity Scale.' This PA scale is a valuable addition to the concept of METs. However, wider application of the new technique requires better quality data with less cultural bias. We expect that new and improved measures will be developed in the future, thus making the benefits of RC available to the field of PA.

Acknowledgments

This study was supported by a grant (SI2.131854 /99CVF3-510) from the European Commission DG SANCO Health Monitoring Programme.

References

1. Craig CL, Marshall AL, Sjöström M, et al. International Physical Activity Questionnaire (IPAQ): 12-country reliability and validity. *Med Sci Sports Exerc*. 2003;35:1381–1395.
2. Wendel-Vos GC, Schuit AJ, Saris WH, Kromhout D. Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. *J Clin Epidemiol*. 2003;56(12):1163–1169.
3. Hopman-Rock M, Van Buuren S, de Kleijn-de Vrankrijker MW. Polytomous Rasch analysis as a tool in the revision of the severity of disability scale of the ICIDH. *Disabil Rehabil*. 2000;22:363–371.
4. van Buuren S, Eyres S, Tennant A, Hopman-Rock M. *Response conversion: a new technology for comparing existing health information*. TNO Report PG/VGZ/2001.097. Leiden: TNO Prevention and Health. ISBN 90-6743-813-8, 2001.
5. van Buuren S, Hopman-Rock M. Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory. *Stat Med*. 2001;20:1061–1076.
6. Zhu W. An empirical investigation of Rasch equating of motor function tasks. *Adap Phys Act Qu*. 2001;18:72–89.
7. Zhu W. Scaling, equating, and linking to make measures interpretable. In: Wood TM, Zhu W, eds. *Measurement theory and practice in kinesiology*. Champaign, IL: Human Kinetics; 2006:93–112.

8. van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Assessing comparability of dressing disability in different countries by response conversion. *Eur J Public Health*. 2003;13(3, Suppl 1):15–19.
9. van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Improving comparability of existing data by response conversion. *J Off Stat*. 2005;21(1):53–72.
10. IPAQ manual, 2005. http://www.ipaq.ki.se/scoring.htm
11. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc*. 2000;32(9, Suppl):S498–S504.
12. Rütten A, Ziemainz H, Schena F, et al. Using different physical activity measurements in eight european countries: results of the European Physical Activity Surveillance System (EUPASS) Time Series Survey. *Public Health Nutr.* 2003;6: 371–376. See also: http://www.public-health.tu-dresden.de/dotnetnuke3/eu/Projects/PastProjects/EUPASS/tabid/337/Default.aspx
13. Sjöström M, Oja P, Hagströmer M, Smith BJ, Bauman AE. Health-enhancing physical activity across European Union countries: the Eurobarometer study. *J Public Health.* 2006, 4:291–300. See also: http://www.euro.who.int/document/e89490.pdf
14. Embretson SE, Reise SP. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum; 2000.
15. Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *J Appl Meas*. 2003;4:205–221.
16. Meulman JJ, Heiser WJ. *SPSS. SPSS Categories 10.0*. Chicago: SPSS Inc.; 1999.
17. RUMM LABORATORIES. *Rumm 2020. Rasch Unidimensional Measurement Models*, 2003. www.rummlab.com.au
18. Jacobusse G, van Buuren S, Chorus AMJ, Hopman-Rock M. Physical activity. In: Buuren S van, Tennant A (ed). *Response conversion for the Health Monitoring Program*. Leiden: TNO Prevention and Health, Publ. nr. 04.145. ISBN 90-5986-082-9, 2004 (available at www.stefvanbuuren.nl).
19. Wright BD, Masters GN. *Rating scale analysis*. Chicago: MESA Press; 1982.
20. Smith RM. *Application of Rasch measurement*. Chicago: MESA Press; 1992.
21. Holland PW, Wainer H, eds. *Differential item functioning*. New York: Lawrence Erlbaum; 1993.
22. Stocking ML, Lord FM. Developing a common metric in item response theory. *Appl Psychol Meas*. 1983;7(2):201–210.
23. Bauman A, Bull F, Chey T, et al. The International Prevalence Study on Physical Activity: results from 20 countries. *Int J Behav Nutr Phys Act*. 2009;6(1):21.
24. Heesch KC, van Uffelen JG, Hill RL, Brown WJ . What do IPAQ questions mean to older adults? Lessons from cognitive interviews. *Int J Behav Nutr Phys Act.* 2010, 11;7:35.
25. Bull FC, Maslin T. *Final report on reliability and validity of the Global Physical Activity Questionnaire (GPAQ v1)*. Geneva: World Health Organization; 2006.
26. Hildebrandt VH, Ooijendijk WTM, Hopman-Rock M. *Trendrapport bewegen en gezondheid 2006/2007*. Leiden, The Netherlands: TNO; 2008. [Trend report Physical Activity and Health].