Journal of Official Statistics, Vol. 21, No. 1, 2005, pp. 53–72

Improving Comparability of Existing Data by Response Conversion

Stef van Buuren[1,2], *Sophie Eyres*[3], *Alan Tennant*[2], *and Marijke Hopman-Rock*[1,4]

Incomparability of information is the key problem in international comparisons. The usual way to improve comparability is to harmonise data collection efforts. However, harmonisation fails if the data have already been sampled, or if appropriate harmonisation cannot be achieved for whatever reason. This normally leaves no other option than to either make strong unwarranted assumptions about the data, or abandon any comparative work. This article proposes an approach that, under certain circumstances, might provide useful comparative analyses from existing incomparable data.

The method, termed Response Conversion (RC), addresses the problem of divergent formulations of survey questions. RC attempts to transform responses obtained on different items (questions) onto a common scale. Where this can be done, comparisons can be made using the common scale. The method consists of two steps. The first step is the construction of a conversion key by means of a statistical model, in our case the polytomous Rasch model. This can only be done if enough overlapping information is available between the different items, but when it is there, RC makes no assumptions about the distribution of the common scores across populations. A linkage map of studies by items provides an important tool to assess whether such overlapping information is available in the data at hand. The second step uses the conversion key to convert the information onto the common scale. This step is straightforward, and can be repeatedly done on a routine basis as new information arrives.

The properties of the Rasch model are well-known, but the model's application in this context introduces some methodological issues. These include the assessment of the model fit in sparse data situations (including the assessment of unidimensionality and the absence of differential item functioning), the robustness of the results regarding the choice of the prior distribution, and the uncertainty introduced if only one item is measured. We believe that all issues can be adequately addressed, and that RC is one of the very few principled approaches for analysing incomparable data.

The method was developed within the EC Health Monitoring Program, and is illustrated for estimating walking disability in different countries.

1. Introduction

Incomparability of information is the key problem in international comparisons (Van de Vijver and Leung 1997; Van Deth 1998). The usual way to improve comparability is to harmonise data collection. Internationally operating organisations like the World Health

[1] TNO Quality of Life, Leiden, The Netherlands. Email: S.vanBuuren@pg.tno.nl
[2] University of Utrecht, Department of Methodology and Statistics, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
[3] Psychometric Laboratory for Care Outcomes Measurement, University of Leeds, U.K.
[4] Body @ Work, Research Center for Physical Activity, Work and Health, TNO VU Medical Center, Amsterdam, The Netherlands. Email: m.hopman@pg.tno.nl or m.hopman@vumc.nl
Acknowledgements: We thank Dr. E. Odding, Dr. T.P.B.M. Suurmeijer and Dr. J.G. Hautvast for their valuable assistance in getting the data. The study was financially supported by the Health Monitoring Proramme of the European Commission under agreement number SI2.131854 (99CVF3-510). The views expressed in this article are the authors' and are not necessarily endorsed by the European Commission.

Organisation (WHO), the Organisation for Economic Co-operation and Development (OECD), the United Nations (UN) and the European Commission (EC) devote much time and effort in obtaining comparable data. However, harmonisation is limited as a method for achieving comparability. First, it works only for new data. Also, harmonisation typically slows down the uptake of scientific advances in measurement because logistic complexities like coordination and translation require time and resources. Finally, it is difficult because harmonisation tries to change established ways of working in environments with vested interests. If harmonisation cannot be done, then the options are restricted. One could either make strong unwarranted assumptions about the data, or not do any comparative work at all. This article proposes an approach that, under certain circumstances, might provide comparative analyses from existing incomparable data.

The Health Monitoring Programme (HMP) is an initiative of the European Commission (EC) to provide relevant and timely information about the health situation in each member state (European Commission 1998). To avoid unnecessary duplication, the monitoring system will have to be fed by existing data collected through health surveys performed by individual member states (MS). Though the content of these surveys is often quite similar, substantial variations in the actual measurement exist, e.g., in sampling procedures, in the coverage per topic, in the wording of questions and formulation of response categories. Thus, incomparability of information is a major problem in this context. Each MS has its own tradition in collecting and processing health related data, and changing established ways of working is not easy.

The type of comparability problem considered here occurs in the differences in formulation of survey questions and response categories. Suppose we want to get insight into the level of disability of the populations of different MS's. Many MS's conduct health surveys, but the precise way in which disability is measured could be quite different. For example, for walking disability, the UK health survey contains a question *How far can you walk without stopping/experiencing severe discomfort, on your own, with aid if normally used?* with response categories "can't walk," "a few steps only," "more than a few steps but less than 200 yards" and "200 yards or more." The Dutch health interview contains the question *Can you walk 400 metres without resting (with walking stick if necessary)?* with response categories "yes, no difficulty," "yes, minor difficulty," "yes, major difficulty" and "no." Both items obviously intend to measure the ability to walk, but it is far from clear how an answer on the UK item can be compared with one on the Dutch item.

The method, termed Response Conversion (RC), attempts to transform responses obtained on different items (questions) onto a common scale. Where this can be done, comparisons can be made using the common scale. The goal of this article is to demonstrate how RC can be applied in the context of harmonisation of data. The next section describes the method and its assumptions. In Section 3, the method is applied to harmonise walking disability in different MS's. A discussion of the relative merits and limitations of the method and the future implications concludes the article.

2. Method

2.1. General

RC is based on the idea that values measured by different instruments can be converted to a common unit. One could, for example, measure the distance between two points in many

ways: with a ruler, by the time taken to reflect sound (e.g., sonar), by a shift in the electromagnetic spectrum (as in astronomy), by a difference between viewing angles, and so on. The resulting values (cm, seconds, colours, and degrees) can be expressed in terms of a common distance unit if one knows how the observed data relate to the common unit.

The same idea can be applied to survey measurement. If we are presented with different questions measuring the same phenomenon, it is natural to ask if there is some way to place the responses on a common scale. This is what RC intends to bring about. Application of RC consists of two steps. The first step is the construction of a conversion key, which models the relation between the common scale and the observed data. Key construction is a relatively complex activity, but needs to be done only once. The second step consists of using the conversion key to convert the observed data into the common scale. This step is relatively simple, and can be repeatedly done on a routine basis as new information arrives. Once expressed in the common scale, information can be compared, for example, across countries that use different questionnaires.

2.2. *Model*

The conversion key is constructed by fitting a statistical model on appropriately linked data. This section illustrates the main issues in model fitting using a small data example involving just three survey questions and two studies.

Table 1 is an excerpt of data analysed previously by Hoparan Van Buuren, and De Kleijn-De Vrankrijker (2000) and Van Buuren and Hopman-Rock (2001). The rows contain survey questions that measure an aspect of walking disability (SI01, HAQ8, GAR9), and the columns represent two studies in which they were sampled (ERGOPLUS, EURIDISS). The ERGOPLUS study (Odding et al. 1995; Hopman-Rock et al. 1996) contains responses on the item SI01 from the ambulation scale of the Sickness Impact Profile. The EURIDISS study (European Research on Incapacitating Diseases and Social Support) contains responses on the item GAR9 with four response categories from the GARS questionnaire (Suurmeijer et al. 1994). Both SI01 and GAR9 measure the ability to walk, but with only these two items there is no way of comparing the amount of walking disability between ERGOPLUS and EURIDISS. Table 1 shows that both studies also administered the HAQ8 item, another walking disability item. The HAQ8 links SI01 to GAR9, and therefore HAQ8 is called a bridge item. Simple visual inspection of the category frequencies of HAQ8 tells us that the EURIDISS sample is more disabled than the ERGOPLUS sample. The more important observation, however, is that the link by HAQ8 allows us to relate the answers in SI01 and GAR9.

In order to construct the conversion key, the data are modelled by the polytomous Rasch model according to Masters (1982), also known as the Partial Credit Model. This model assumes the existence of a continuous latent trait θ that underlies all items. The term "latent" means that the true value of θ_i for person i is not known, and can only be observed through the manifest item responses on the items. In the example discussed above, the trait θ makes up a common scale for walking disability. In order to define the model, let item Y_j have $k_j + 1$ response categories. The polytomous Rasch model defines the probability $p(Y_j = c|\theta)$ of responding in Category $c = 0, \ldots, k_j$ as a function of the score on the latent trait θ by the following function:

Table 1. SI01 and GAR9 items linked by bridge item HAQ8

Item	Description	Response categories	Study	
			ERGOPLUS $n = 306$	EURIDISS $n = 292$
SI01	I walk shorter distances or often stop for a rest.	0 = No	276	
		1 = Yes	28	
HAQ8	Able to walk outdoors on flat ground?	0 = Without any difficulty	242	178
		1 = With some difficulty	43	68
		2 = With much difficulty	15	42
		3 = Unable to do so	0	2
GAR9	Can you, fully independently, walk outdoors (if necessary, with a cane)?	0 = Yes, no difficulty		145
		1 = Yes, with some difficulty		110
		2 = Yes, with much difficulty		29
		3 = No, only with help from others		8

$$p(Y_j = c|\theta) = \frac{\exp\sum_{k=0}^{c}(\theta - \delta_{jk})}{\sum_{r=0}^{k_j}\exp\sum_{k=0}^{r}(\theta - \delta_{jk})} \quad c = 0, 1, \ldots, k_j \tag{1}$$

where $\sum_{k=0}^{0}(\theta - \delta_{jk}) \equiv 0$ and $\sum_{k=0}^{r}(\theta - \delta_{jk}) \equiv \sum_{k=1}^{r}(\theta - \delta_{jk})$. When plotted against θ, the values of $p(Y_j = c|\theta)$ define the Category Probability Curves (CPC). Psychometric models other than the Rasch model are possible, but the Rasch model is special because of its specific objectivity (Rasch 1977). This implies that the model parameters can be separated from the sample. If the model fits, the ability level of the calibration sample does not affect the relative positions of the items, which is a desirable property. Furthermore, the Rasch model has few parameters, so it is relatively stable if the data are sparse. The parameter δ_{jk} is known as the threshold value. It can be interpreted as the point on the latent trait scale at which two consecutive CPC's intersect. Thus, for an item with $k_j + 1$ response categories, the k_j category intersection points define the relation between the latent trait and the observed item score. Knowledge of the thresholds is enough to reconstruct the curves.

Estimation of the model requires appropriate data. The parameters of the Rasch model (1) can only be estimated if the items are linked to each other. For example, if HAQ8 in Table 1 were not present, there would be no way of comparing the EURIDISS and ERGOPLUS samples, and construction of the conversion key would not be possible. More specifically, the Rasch model in (1) implies that bridge items j 1) measure the same characteristic as the target items; and 2) have identical relations between the latent trait and the observed data in the respective samples, i.e., for studies A and B:

$$\delta_{jk}^{A} = \delta_{jk}^{B} = \delta_{jk} \quad \text{for all } k = 1, \ldots, k_j \tag{2}$$

The second possibility for linking two items is a bridge study that contains information on both target items. In that case, Model (1) implies for target items a and b, and bridge study C that

$$\delta_{ak}^{A} = \delta_{ak}^{C} = \delta_{ak} \quad \text{for all } k = 1, \ldots, k_a \quad \text{and} \tag{3a}$$

$$\delta_{bk}^{B} = \delta_{bk}^{C} = \delta_{bk} \quad \text{for all } k = 1, \ldots, k_b \tag{3b}$$

Equations (2) and (3) are both implications of the Rasch model. They state that the linking information should be free of Differential Item Functioning (DIF), i.e., items are assumed to work in the same way across studies (Holland and Wainer 1993). Designs that adhere to these specifications are classified as nonequivalent linked grouped designs (Kolen and Brennan 1995). Verhelst and Kleintjes (1993) provide an application of a linked design within an educational context. Béguin (2000) reports related theoretical work on the appropriateness of the Rasch model for test equating. Note that no assumptions are required with respect to the level of disability in each study. Operationally, Equations (2) and (3) imply that the linking information can be coded into the same data column(s) across different studies.

RUMM 2010 (Andrich et al. 2001) was used to estimate the model. The estimation method is based on the pairwise conditional approach, and has been described in detail by

Andrich and Luo (2003). This approach generally works well with incomplete and sparse data (Andrich 1988). The method conditions on the latent ability, so the model estimates are not sensitive to the distribution of trait in the sample. Table 2 provides the threshold estimates obtained from the data in Table 1. The data fitted the Rasch model ($\chi^2 = 8.49$, $df = 10$, $P = 0.58$).

Figure 1 contains the CPC's of items SI01, HAQ8 and GAR9 as estimated by RUMM 2010. For low θ (e.g., no disability), the probability of answering in the most severe disability response categories is low. For example, a person without any walking restrictions is unlikely to respond in Category 1 ("Yes") of SI01, or in Category 3 of GAR9. On the other hand, persons with severe restrictions (i.e., with high values of θ) have high probabilities responding in those categories, and exhibit relatively low propensities to respond in the less disabled categories. In Figure 1, the horizontal axis orders walking disability from no disability (left) to high disability levels (right). The horizontal axes in the plots are identical. So, if we know the disability position θ_i of a person, then we can read off the response probabilities for every item. For example, someone with $\theta_i = -1$ has a probability of 0.62 of responding in, respectively, Category 0 of SI01, and a probability of 0.38 in Category 1. The same person has probabilities of 0.27, 0.50, 0.23, and 0.00 to respond in Categories 0, 1, 2, and 3 of HAQ8. The response probabilities for GAR9 are 0.11, 0.72, 0.16 and 0.01, respectively.

2.3. *Key construction*

Imagine that we have two *new* studies on different samples, where the first administers item SI01 (but not HAQ8) and the second administers GAR9 (but not HAQ8). Is it possible to compare the level of disability in the two new studies, even in the absence of bridge items? The answer is yes, provided that an appropriate conversion key is available. This section discusses ways to construct such a key.

Suppose we have observed data x_{ij} on a sample of persons $i = 1, \ldots, n$ for a given item j. The problem is to estimate the location θ_i of person i on the common scale θ from the data. This problem is known as "ability estimation" or "scoring," and several strategies can been pursued, such as maximum likelihood (ML), maximum a posteriori (MAP) and expected a posteriori (EAP) estimation. Embretson and Reise (2000) provide an overview of these methods. In the sequel, we use the EAP method (Bock and Mislevy 1982). The EAP estimator is a Bayesian method that is easy to calculate, gives finite trait level estimates for extreme response patterns, has minimum mean squared error if the prior is true, and is robust to misspecification errors (Wainer and Thissen 1987). As we will show,

Table 2. Threshold estimates for the data in Table 1

Item	Category transition		
	0/1	1/2	2/3
SI01	−0.802		
HAQ8	−1.413	−0.140	4.012
GAR9	−2.687	0.663	1.970

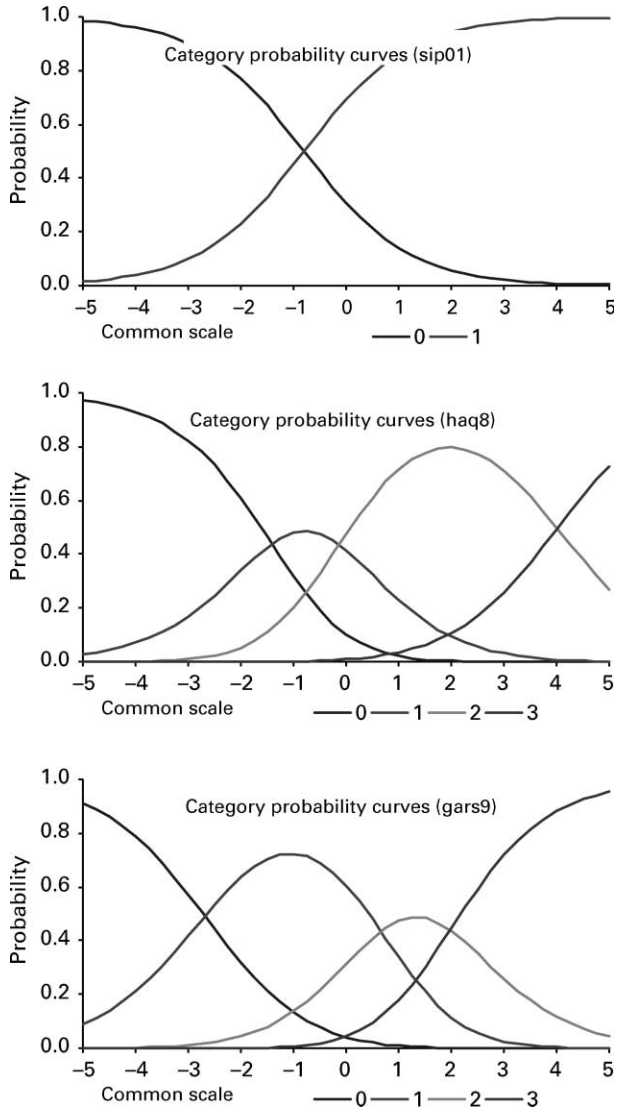

Fig. 1. Category Probability Curves: Probability of responding in each category for a given level of disability.

the approach provides a value on the common scale for each category of the item, but requires specification of a prior distribution. Response conversion replaces the category identification number by these values, and calculations can subsequently be made on the common scale.

The EAP estimator derives from Bayesian analysis. Let Y_j denote an item with $k_j + 1$ possible responses. According to Bayes theorem, the posterior distribution of θ for a given answer $Y_j = c$ can be written as

$$p(\theta|Y_j = c) = \frac{p(Y_j = c|\theta)p(\theta)}{\sum_\theta p(Y_j = c|\theta)p(\theta)} \quad \text{for } c = 0, 1 \ldots, k_j \tag{4}$$

It is easy to calculate $p(\theta|Y_j = c)$ on a grid of θ-values. A convenient choice for the grid is $\theta = \{-5, \ -4.75, \ \ldots, \ 4.75, \ 5\}$. The probability $p(Y_j = c|\theta)$ is given by Model (1). The expression $p(\theta)$ is a Bayesian prior and summarises all information that we have before the current data. Choosing an appropriate prior takes some care, and we will have back to it later. For the moment, let us assume a uniform distribution $p(\theta) \sim U(-5, 5)$.

Figure 2 illustrates the relevant calculations for item HAQ8. Figures 2a and 2b plot $p(Y_j = c|\theta)$ and $p(\theta)$, respectively. Figure 2c presents the results of calculating (4), i.e., the posterior distribution per category $p(\theta|Y_j = c)$. This distribution describes what is known about θ after an answer is observed. Note that each $p(\theta|Y_j = c)$ is scaled to unit area and can be interpreted as a density. Note also that $p(\theta|Y_j = c)$ is proportional to $p(Y_j = c|\theta)$, a consequence of the uniform prior. Figure 4d represents a mixture of the densities $p(\theta|Y_j = c)$, in this case with mixture weights 242, 43, 15, and 0, i.e., the observed counts on HAQ8 in the ERGOPLUS study. Thus, the posterior reflects how the ERGOPLUS sample is distributed on the common scale based on the HAQ8 data.

More formally, we may calculate the sample posterior density in Figure 4d as

$$p(\theta|Y_j) = \frac{\sum_{c=0}^{k_j} w_c p(\theta|Y_j = c)}{\sum_c w_c} \tag{5}$$

where w_c is the frequency of Category c in the sample of interest. It is not difficult to show that the sample EAP estimator is equal to

$$E[\theta|Y_j] = \frac{\sum_c w_c E[\theta|Y_j = c]}{\sum_c w_c} \tag{6}$$

where $E[.]$ is the expectation operator. Thus, we can calculate the sample EAP estimator as the weighted average of the mean of the category posteriors (4). This property gives us the possibility for the basic operation in RC: recode the category number identification by the mean category posterior, and aggregate these values over sample of interest to obtain the mean on the common scale for that sample. Consider the mean category posteriors of

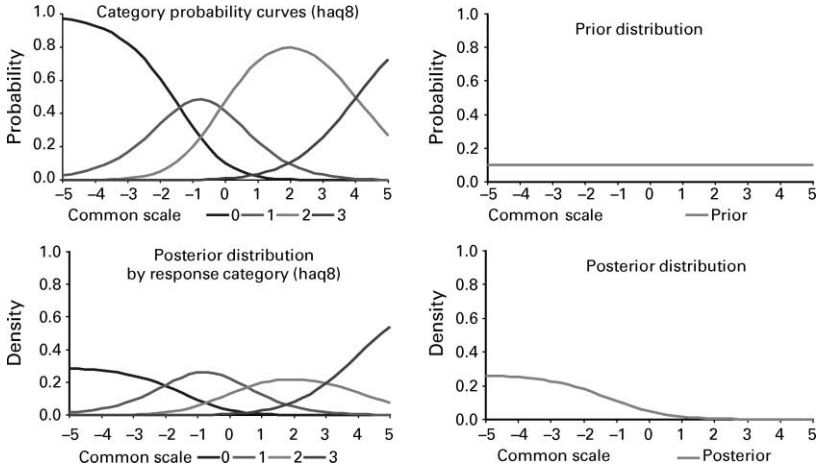

Fig. 2. HAQ8 item "Able to walk outdoors on flat ground?": Category Probability Curves (a), uniform prior distribution U(−5,5) (b), Category Posterior Curves (c) and Sample Posterior Distribution (d).

HAQ8 in Figure 2c, which are equal to −3.123, −0.803, 1.917, and 3.823. The following two SPSS commands convert the HAQ8 data into the common scale, and calculate the sample EAP estimate on the common scale for both samples:

```
RECODE haq8 (0 = −3.123) (1 = −0.803) (2 = 1.917) (3 = 3.823)
(ELSE = SYSMIS).
MEANS haq8 BY study.
```

The set of recode values makes up the conversion key. Table 3 contains the conversion key for the three items, as well as the result of the estimated mean disability level. Note that the estimated effects in terms of the common scale are in the expected direction. The frequency distributions of the HAQ8 item in Table 1 clearly indicate that the EURIDISS sample possesses more disabilities than the ERGOPLUS sample. Both EURIDISS estimates (−1.80 and −1.93) are larger than the ERGOPLUS estimates (−2.18 and −2.54).

The progress now made is that it is possible to compare the ERGOPLUS and EURIDISS samples without knowing any bridge items. For example, if we have measured only SI01 in ERGOPLUS and GAR9 in EURIDISS, then we can still calculate the difference between the samples in terms of the common scale as $(-2.18) - (-1.93) = -0.25$. Note that it is also possible to calculate various other combinations, of which the comparison HAQ8 - HAQ8 yields the largest difference, i.e., $(-2.54) - (-1.80) = -0.74$. These differences in effect estimates are not untypical, and are caused by a number of factors. First, note that the calibration sample is the same as the comparison sample, so part of the differences may be explained by overfitting. The model is essentially fitted on HAQ8, so it is not surprising that it optimises that difference. Another factor is regression to the mean of the common scale estimate, which especially occurs if the number of items is small (Wainer and Thissen 1987). Finally, some items measure the trait more precisely than others. For example, the dichotomous SI01 item provides less information than either HAQ8 or GAR9, which have more categories and cover more of the scale. All these are well-known statistical phenomena, and there are ways to circumvent them, e.g., by calculating appropriate confidence intervals, by applying "unshrinking" techniques and by obtaining denser data. These are topics for further work and beyond the scope of this article. The main progress made here is that the technique expresses such differences on a common scale, which is a prerequisite for doing any further quantitative work.

Table 3. Recode values (conversion key) under a uniform (−5,5) prior, and the mean disability levels for the ERGOPLUS and EURIDISS samples in Table 1 expressed on the common scale per item

Item	Conversion key				Mean disability on the common scale	
	Response category (c)				ERGOPLUS	EURIDISS
	0	1	2	3		
SI01	−2.598	1.903			−2.18	
HAQ8	−3.123	−0.803	1.917	3.823	−2.54	−1.80
GAR9	−3.449	−1.192	1.356	3.355		−1.93

2.4. Choice of the prior distribution

Statisticians can be divided into two camps: those who do not like prior distributions and never use them, and those who use them. The first group is much larger than the second, so it is natural to ask whether we need a prior distribution at all, and if so, what are the consequences of different choices?

The first question is easy to answer. The conventional Maximum Likelihood (ML) estimator is the optimal non-Bayesian choice. It yields unbiased ability estimates of the common scale, and is mathematically equivalent to the Bayesian estimator (5) with a uniform prior across the entire scale (Embretson and Reise 2000). The problem with the ML estimator is that its variance for extreme responses is infinite, so the common scale value for people with "all low" or "all high" scores cannot be determined. One solution is to eliminate the extreme persons from the estimation, which is O.K. if there are not many extremes. The present application, however, requires estimation of scale values from as little as one item. Eliminating the extremes will then lead to large losses of data. In the limiting case with a dichotomous item, there will be no data left because all persons end up being extreme. So conventional ML does not work here, and ways to make ML work do not work either. Alternatives to the ML estimator have been proposed, e.g., the estimator presented by Warm (1989). Such alternatives essentially weight down the extremes of the scale. The Bayesian estimator does the same thing, but it is simpler, provides the full posterior density, and makes the weighting process explicit.

The second question is how robust the inference on the common scale is under alternative priors. In general, the prior contracts scale estimation towards the highest prior densities. Alternatively one may use the prior to define gaps and end points of the scale by specifying zero mass. In order to get insight into the properties, we specified priors with very different shapes and properties, and studied the resulting estimates.

Figure 3 shows the mean estimates for the data in Table 1 under a variety of prior specifications. It will be immediately clear that estimates using different priors cannot be compared with each other. Thus, all comparisons between samples should use a common prior for them to be valid. Using a uniform prior $U(-10,10)$ instead of $U(-5,5)$ brings the method closer to the ML estimate. Note that the resulting mean estimates wander off to the left, thus indicating what happens if we use ML estimation, where the

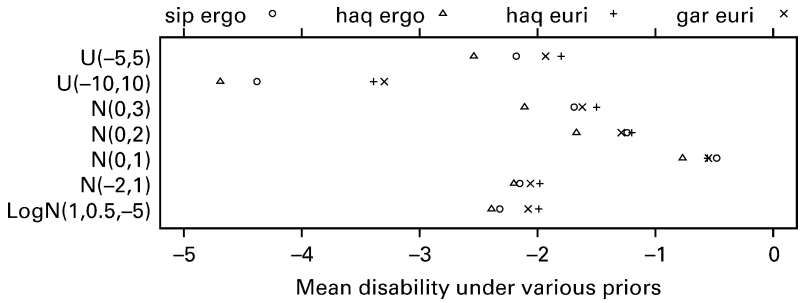

Fig. 3. Mean disability for the following sample-item combinations: SIP01 administered in ERGOPLUS, HAQ8 administered in ERGOPLUS, HAQ8 administered in EURIDISS, and GARS9 administered in EURIDISS under two uniform priors, four normal priors and one shifted lognormal prior.

prior is $U(-\infty, \infty)$. The normal priors around zero, $N(0,1)$, $N(0,2)$, and $N(0,3)$, all pull the estimates towards the origin. An advantage of using normal priors is that the resulting posteriors are also normal. Note that SIP01 is pulled more than HAQ8 or GAR9, eventually resulting in the odd finding that SIP01 moves beyond all other items under $N(0,1)$. Clearly, centring on zero is not a good idea. The two final priors, $N(-2,1)$ and the shifted lognormal prior $LogN(1, 0.5, -5)$ with logmean 1, variance 0.5 and a shift -5, exhibit less differential pooling. The lognormal prior is the only asymmetric one. We prefer it in this data because it is a left-skewed distribution that resembles the disability distribution in the general population. In the absence of any data, we would expect more people in the low disability levels. In addition, the lognormal yields estimates that are consistent in the sense that both items administered within the same study indicate approximately similar levels of disability. In the sequel, we will therefore base our analyses on the lognormal prior.

3. Walking Disability

Disability is one of the topics of the Health Monitoring Programme of the EC. This section uses the principles set forth above to harmonisation of measurements of walking disability, as defined according to the ICIDH-D Code 40 (WHO 1980) (During the study, the ICIDH was superseded by the ICF classification.)

Table 4 provides an overview of walking disability items that are currently being used in 10 European countries. All items measure the ability to walk, but do so in different ways. Variations occur in the formulations of both the question and the response categories. The most important differences relate to the concepts behind the item. For example, some items ask how difficult it is to walk a fixed distance (often 400 metres), others concentrate on how far you can walk, or how long or how fast you can walk without difficulty. Items can sometimes be traced back to a common "ancestor." For example, fixed distance items come from the OECD long-term disability questionnaire ("Can you walk 400 metres without resting?") Items that use "how far" in the question are variations on the WHO-Europe long-term disability questionnaire ("What is the furthest you can walk on your own without stopping and without severe discomfort?")

A total of 54 bridge items and 14 additional studies were identified that could potentially link items. The report by Van Buuren et al. (2001) contains a detailed description of all items. Figure 4 is the corresponding linkage diagram. The "*Y*"-symbol in a cell indicates that the specific study-item combination occurs. Cells with the "*Y*"-symbol are also coloured to make them easier to find. As far as possible, the items are sorted such that each item is located near the other item close in content. The existence of a link is a technical requirement for scaling different items on a common scale. Rather few items are directly linked. For example, FAR7 and MANAGE are linked by the bridge item AIDS1. Another example is the link between SI01 and GAR9 by HAQ8. However, many items are not directly linked, and so cannot be calibrated relative to each other without additional information. One solution is to get more data, with the risk of ending up with even more different versions of items for measuring walking disability. Our solution is to assume equivalent thresholds according to (2) where we feel that this can be done. For example, items FAR1 to FAR11 are all variations on the same basic formulation. Variations include additions like "on your

Table 4. Items for measuring walking disability in different European health surveys

State	Item	Description	Response categories
UK	far2	How far can you walk without stopping/experiencing severe discomfort, on your own, with aid if normally used?	0 200yds or more 1 More than a few steps but less than 200yds 2 A few steps only 3 Can't walk
CH	far11	How far can you walk without stopping/experiencing severe discomfort on your own?	0 200 m or more 1 More than a few steps but less than 200m 2 A few steps only 3 Cannot walk unaided
N	far4	How far can you walk without stopping/experiencing severe discomfort on your own?	0 200 m or more 1 More than a few steps but less than 200m 2 A few steps only
B	far1	How far can you walk without stopping/experiencing severe discomfort?	0 200 m or more 1 More than a few steps but less than 200m 2 A few steps only in metres
F	far3	How far can you walk without stopping/experiencing severe discomfort? (walk with/without aids/uses wheelchair etc.)	In meters
NL	w400c	Can you walk 400 metres without resting (with walking stick if necessary)?	0 Yes no difficulty 1 Yes minor difficulty 2 Yes major difficulty 3 No
DK	w400b	Can you walk 400 metres without resting?	0 Yes no difficulty 1 Yes minor difficulty 2 Yes major difficulty 3 No
FIN	w400d	Can you walk without any aids a distance of 400 metres without difficulty?	0 Yes no difficulty 1 Yes minor difficulty 2 Yes major difficulty 3 No
S	brisk	Can you take a short walk, say five minutes, at a fairly brisk pace?	0 Yes 1 No
A	dwelling	Walking up and down in the dwelling?	0 Yes possible w/o help, 1 Yes possible with help, 2 Not possible

own," "on a level ground" or "with a walking stick if needed" in the question, and the definition of one or two extra response categories. Another example comprises the "400 metres" items W400B, W400C and W400D that use questions that are somewhat different but have similar response categories. Assuming equivalence is part of the modelling process

and should not be done lightly. In general, we need to carefully identify items where comparability can be assumed across studies. This might even mean that we must break up links for items that are identical in their English translation, but for which comparability is suspect. An example of such an item is "How is your health in general?", for which widely different, and thus far unexplained, answers are obtained in different European countries. The appropriate treatment in RC is that such an item is considered as different items. As we will show below, it is possible to check on equivalence to some extent. Figure 4 codes the equivalence assumptions made in the current linkage structure using a different colour and with the symbol "I." Note that the Swedish item BRISK is not linked to any other item, and therefore cannot be calibrated.

We collected existing data from statistical offices and investigators of potential bridge studies. Data were provided in the form of individual microdata, or as multidimensional contingency tables, and included age and sex. These data were organised into a database according to the structure of the linkage matrix. The statistical analysis of the linked data consisted of several model fitting phases with RUMM 2010, including an assessment of model fit and a check of equivalence assumptions. This involved removing items that could not be linked because we could not get the appropriate data that had fit residuals over three, and that showed DIF across studies. The worst-fitting item was HAQ8. Removing HAQ8 would result in two data sets that are essentially unconnected. Because HAQ8 is so vital to the structure and behaved well in the example in Section 2, we decided to leave it in. Other items that did not fit well were HOUSE and QMILELEV. For items that were assumed to be equivalent, we calculated a tension coefficient to assess the validity of the assumption. For the set of items assumed to be equivalent, we fitted two copies of the same item in one analysis, one copy as a separate item and one copy as part of the common item. So if there are s items assumed to be equivalent, the data matrix consists of $s + 1$ columns, one for each item and one for the common item. The tension coefficient is calculated as

$$\mathrm{T}_j = \frac{1}{k_j} \sum_{k=1}^{k_j} \left(\delta_{jk}^C - \delta_{jk} \right)^2 \tag{7}$$

where δ_{jk}^C are the threshold estimates of the common item and where δ_{jk} are the estimates of the item j when treated as a separate item. The tension coefficient expresses badness of fit to the equivalence assumption. For our data, T_j varied between 0.0 and 3.6, with the majority being lower than 1.0. The distribution of T_j is unknown. Seventeen out of 25 items had low tension (<0.5). High-tension items appeared in blocks K (item SIP12) and M (SIP11 and AIDS1). Breaking up these blocks is somewhat dangerous because, as Figure 4 shows, both are central components in the linkage structure. We reran the analysis where blocks K and M were broken up into separate items, but as this led to nonsensical threshold estimates, we thought it would be better to leave the blocks unbroken. These modelling issues illustrate that we pushed to the limit what could be achieved with these data. In general, the structure of the data imposes a great responsibility on the quality and fit of the linkage items. The analysis resulted in a conversion key of 36 out of the 54 items for measuring walking disability.

Mean disability levels per category on the common scale were calculated from the conversion key under the lognormal prior. Walking disability estimates on a national level were then estimated in two steps. First, each observed score was replaced by the

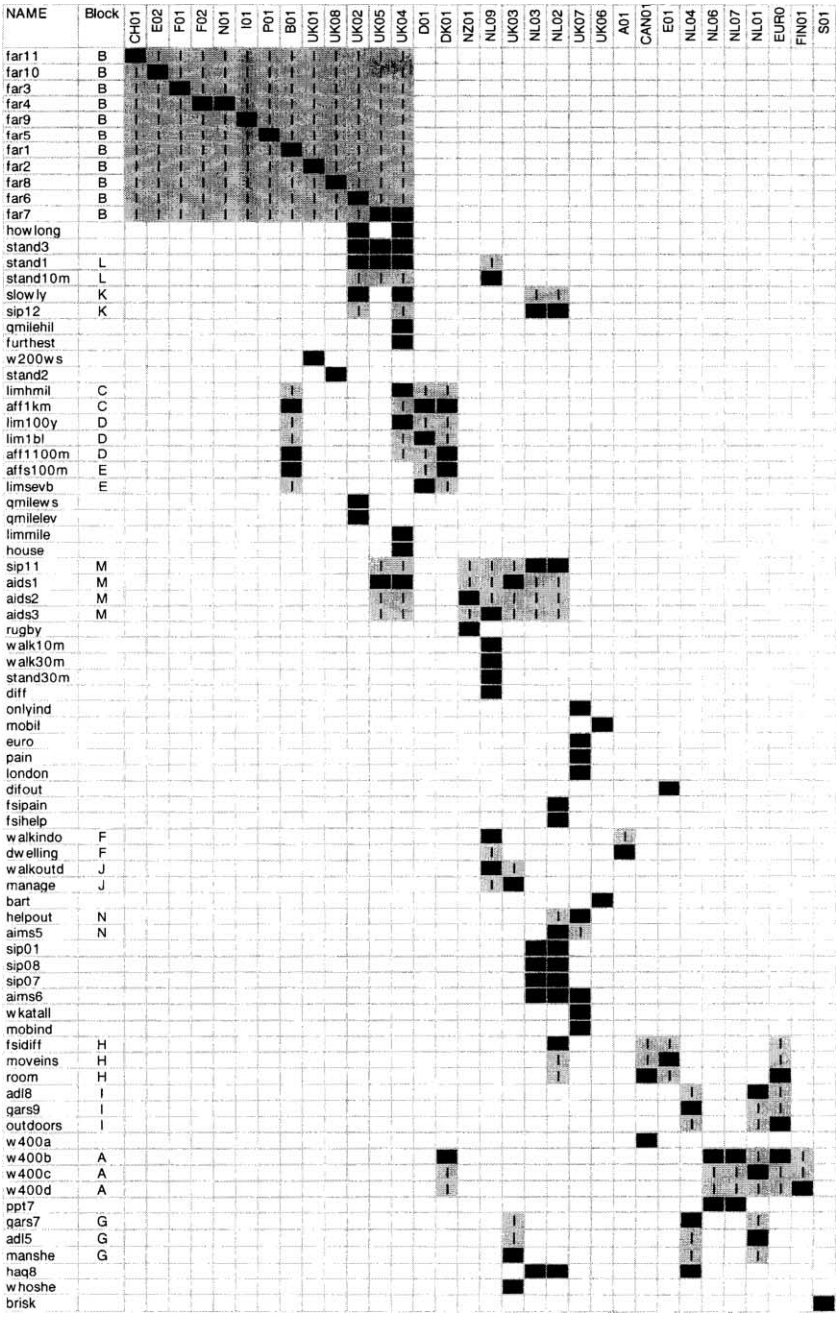

Fig. 4. Linkage diagram of 54 items for measuring walking disability.

corresponding mean disability estimate. Second, we computed the mean of the sample posterior by Equation (6), i.e., as the average over groups of interest such as age and sex. We did not have appropriate survey weights available, but if we had had them, it would have been easy to incorporate them in the calculations.

Figure 5 displays the results for women between 40 and 90 years after the responses were converted to the common scale by response conversion. As not all items were sampled at all ages, not all age groups are present in all plots. As expected, disability generally increases with age. In general, the relation between age and walking disability in Belgium, Denmark, Italy, The Netherlands, and the UK follows a similar pattern. The Austrian line is peculiar in the sense that it is high and shows little age trend. An explanation for this is that the conversion key does not account for the Austrian item very well. Swiss walking disability is extremely low across all ages, which is also already apparent from the raw data (not shown) that show suspiciously few scores in the categories indicating more disabilities. The Danish and Dutch estimates appear somewhat larger than average. As both studies are based on a "400 metres" item, this could raise suspicion about any systematic bias in the conversion key with respect to such an item. Observe however that the low position of the SENECA study (Van 't Hof et al. 1991), which posed a "400 metres" question to a mix of European countries, does not support this. The results for men were similar, but because of smaller samples in the upper ages show slightly more variability.

4. Discussion

Incomparability of data is a key problem in international comparisons. Response conversion is a new method for "making incomparable data comparable." RC consists of two steps: construction of a conversion key, and application of the key to data. The method systematically exploits any overlap between existing data sources. The statistical methodology is built on well-established psychometric theory, and has been applied before on linked health data (cf. Van Buuren et al. 2003). The most important asset of the methodology is that it can work with existing, incomparable data. This aids in setting up a health monitoring system without the need to drastically change established ways of working in the participating countries.

The role of the linkage structure is critical in establishing conversion keys. RC only works for those items that are linked. Unlinked items cannot be placed on a common scale. Items can be linked by data (i.e., by bridge studies and bridge items), by assuming equivalence, or by a combination of both. It will be clear that linkage by data is preferable to the use of equivalence assumptions, as the latter are, by definition, not backed up by data. But even then there can be differences in the quality of the link. Some items appeal to an internal frame of reference. This means that the choice among alternatives relies on internal and personal criteria. An example is the HAQ8-item, which contains "with difficulty" categories. For such items, it is difficult to determine whether answers are comparable across different persons, cultures or countries. Comparing the responses regarding such items is only possible under explicit equivalence assumptions. Bridge items with an external frame of reference (e.g., can walk 50 metres) and not showing DIF are therefore preferable.

There are various ways to check for DIF (Holland and Wainer, 1993). The general idea is that equivalence holds if the response probability within different groups (or studies) is similar across different values of the latent trait θ. This method requires an estimate of θ to form groups of equal disability in different studies. The traditional methods for assessing

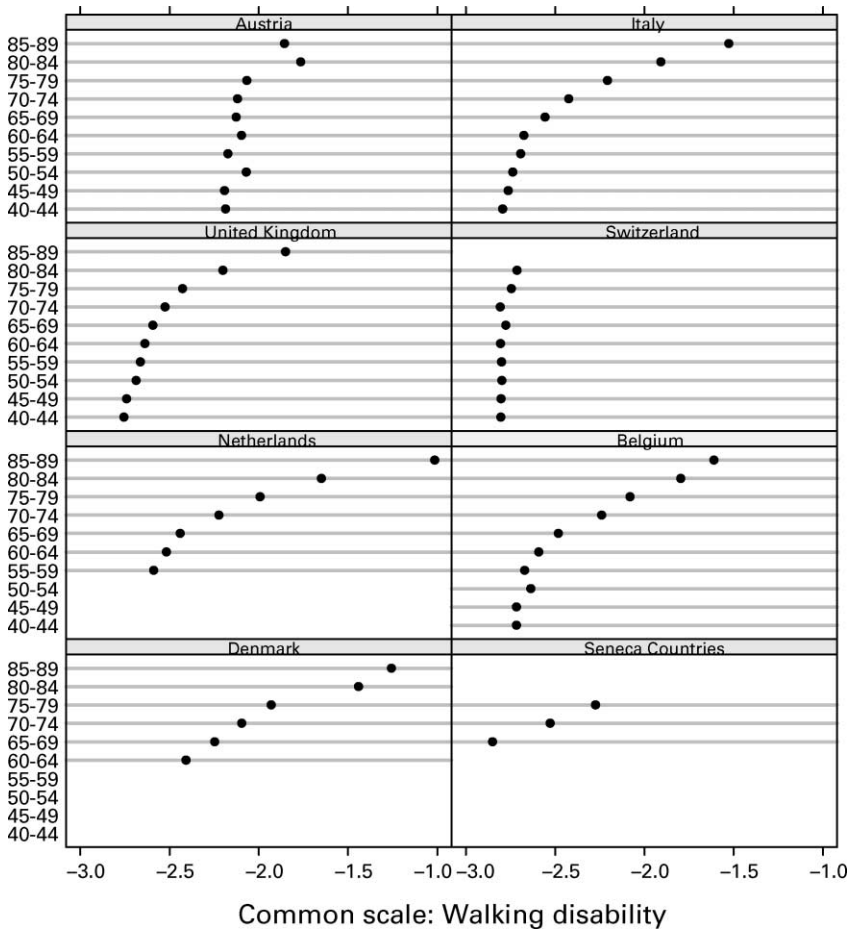

Fig. 5. Mean walking disability in several European countries, expressed on the common scale, for women, split according to age. Data were collected in national health surveys 1995–1998. The countries in the SENECA studies are Belgium, Denmark, France, Greece, Hungary, Italy, The Netherlands, Norway, Portugal, Spain, Switzerland and Poland. The SENECA study collected data during 1988–1989. The figure gives the average over the SENECA countries.

DIF require that at least one second item, known to be free of DIF, has been administered in the studies. In this article, we used an alternative method, based on tension coefficients, for the situation where there are no such bridge items. More work is needed to investigate the properties of this method.

Another factor critical to the success of the method is the requirement that items measure the same latent trait, a property known as unidimensionality. Unidimensionality can be defined in various ways, and there are several approaches to actually check unidimensionality in a given set of data (Hattie 1985). For the domain of walking disability, it is often fairly easy to infer whether an item measures walking disability, but this may not be the case with data in general. Though the model-fitting strategy we employed is likely to lead to a more unidimensional set of items, a more complete methodology to verify unidimensionality for this type of linked data would be useful.

The application of the Rasch model as done here differs from applications in psychological testing and education. The objective in psychological testing and education is to measure the ability of the individual. The number of items typically varies between 5 and 50, and the standard error of the ability estimate is hopefully small. In the present application, the objective is to assess differences between countries. In order to get stable estimates at the group level, we trade in precision at the individual level for a larger sample size. In other words, it is possible to estimate disability on the group level by as little as one item provided that the sample size is substantial, and the relation between the measurement and the common scale is known. The first condition is usually not a problem in health surveys. This article provides a method to meet the second condition by means of a Rasch model.

RC builds on the Rasch model, which is firmly rooted in psychometrics but differs from the traditional approaches because the number of items is extremely low. Having just one or two items for a person introduces new complexities in modelling and estimation. For example, treatment of the extremes becomes critical. Also, common scale values will depend on the measurement properties of the individual items, and are subject to regression to the mean. We have tried to deal with those complexities in the best possible way, but clearly there is room for improvement.

Though substantial effort went into creating the linked data set, the walking disability data were quite sparse and contained relatively few links. This is just a reflection of what you get when you are restricted to combining existing data. As demonstrated, the sparseness of the data prohibited some improvements needed in the model. The modelling task would have been easier if data had been better. RC suggests new ways to collect data. It is of central importance that studies that collect data contain both country-specific and country-independent common indicators. The planning of such studies makes the construction of the linkage matrix more controllable, as its entries will not depend any more on what items happen to be available from previous studies. Such dedicated bridge studies need not be very large or costly, and will lead to more compact and workable linkage matrices. An example is the EUPASS project (Rütten et al. 2003), where new physical activity indicators were collected next to the current country-specific indicator for the same persons. Another one is the KIDSCREEN project (Ravens-Sieberer et al. 2001), where data were collected on both (common) new and (country-specific) old instruments. Data from such studies are in principle more suitable to our approach than the walking disability data. We are currently developing conversion keys for these data.

A discussion of some important practical questions follows:

If new items become available, should the entire conversion key change? It is possible to fix the values of the old items and estimate the conversion key for items that are new to the conversion. Thus, the key can grow while preserving the old elements. However, at a given point it might be desirable to develop a new version of the key. Preferably, the whole process of key construction should be fed and endorsed by appropriate scientific and field authorities.

Can the method be used for sum scores of two or more items? Yes, the psychometric theory underlying response conversion is actually very well suited to estimate ability on two or more items. In practice, it requires some modifications to the procedure to scale items on the common scale. The same threshold estimates could be used as a basis.

Is it possible to quantify the variability of values on the common scale, for example by a 95% confidence interval? There is more than one way of doing this. The simplest approach is to draw imputed values for θ from the category posterior densities. It might be necessary to do so within a multiple imputation framework (Rubin 1987). Rubin's theory leads to interval estimates with appropriate coverage, but simulation work would actually be needed to confirm this for a particular application. One further issue is whether or not to include the variability of the threshold-estimates themselves, as would be required in a full Bayesian analysis. Incorporating the extra threshold uncertainty will flatten out the posterior distributions. In some sense, this is fair because it keeps track of all uncertainty within the system. On the other hand, if the key is updated with more elaborate future data, the measurement properties of the item change in a predictable way, i.e., the item will lead to uniformly sharper inferences. In our view, the size of the calibration sample and the measurement properties should be independent of each other. We thus lean towards the position that all inferences on the converted data should be made conditional on the key used.

Can RC be useful if comparable data are already available? There is increasing evidence that "getting comparable data" is not as easy as it sounds, even when the questions are the same (Harkness et al. 2002). Cross-cultural validity now goes beyond translation and cultural equivalence, and requires scale equivalence through more formal methods like the one discussed in this article.

Finally, let us be clear that we do not advocate our technology as a panacea for solving all harmonisation problems. It is of course always better to get comparable data in the first place. Getting comparable data is easiest for new topics in new samples. It is already more difficult to attain comparability on old topics in new samples, as harmonisation might compromise historic comparability. But with conventional methods it is impossible to get comparable data from past samples. It would be wasteful to discard historical data with the argument that they are not comparable. Likewise, it will be inefficient if new community-wide surveys duplicated data collection efforts of the individual member states. Our method is particularly useful to deal with these kinds of inefficiencies.

5. References

Andrich, D. (1988). Rasch Models for Measurement. Newbury Park: Sage.

Andrich, D., Sheridan, B.S., and Luo, G. (2001). RUMM2010: Rasch Unidimensional Models for Measurement. Perth: RUMM Laboratory.

Andrich, D. and Luo, G. (2003). Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories Using Principal Components. Journal of Applied Measurement, 4, 205–221.

Béguin, A.A. (2000). Robustness of Equating High-Stakes Tests. Dissertation, Universiteit Twente.

Bock, R.D. and Mislevy, R.J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. Applied Psychological Measurement, 6, 431–444.

Embretson, S.E. and Reise, S.P. (2000). Item Response Theory for Psychologists. London: Lawrence Erlbaum.

European Commission (1998). Programme of Community Action on Health Monitoring. Work Programme 1998–1999. Art. 5.2.b of Decision 1400/97/EC. Luxembourg: European Commission.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. Applied Psychological Measurement, 9, 139–164.

Holland, P.W., and Wainer, H. (eds). (1993). Differential Item Functioning. New York: Lawrence Erlbaum.

Hopman-Rock, M., Van Buuren, S., and De Kleijn-De Vrankrijker, M. (2000). Polytomous Rasch Analysis as a Tool in the Revision of the Severity of Disability Scale of the ICIDH. Disability and Rehabilitation, 22, 363–371.

Hopman-Rock, M., Odding, E., Hofman, A., Kraaimaat, F.W., and Bijlsma, J.W.J. (1996). Physical and Psychodisability in Elderly Subjects in Relation to Pain in the Hip or Knee. Journal of Rheumatology, 23, 1037–1044.

Harkness, J.A., van de Vijver, F.J.R., and P.Ph. Mohler (eds). (2002). Cross-cultural Survey Methods. New York: Wiley.

Kolen, M.J. and Brennan, R.L. (1995). Test Equating: Methods and Practices. New York: Springer.

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. Psychometrika, 47, 149–174.

Odding, E., Valkenburg, H.A., Algra, D., Vandenouweland, F.A., Grobbee, D.E., and Hofman, A. (1995). Association of Locomotor Complaints and Disability in the Rotterdam Study. Annals of Rheumatic Diseases, 54, 721–725.

Rasch, G. (1977). On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. In M. Glegvad (ed.), The Danish Yearbook of Philosophy. Copenhagen: Munksgaard, 58–94.

Ravens-Sieberer, U., Gosch, A., Abel, T., Auquier, P., Bellach, B.-M., Bruil, J., Dür, W., Power, M., Rajmil, L. and the European KIDSCREEN Group (2001). Quality of Life in Children and Adolescents: A European Public Health Perspective. Social and Preventive Medicine, 46, 297–302.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rütten, A., Ziemainz, H., Schena, F., Stahl, T., Stiggelbout, M., Auweele, Y.V., Vuillemin, A., and Welshman, J. (2003). Using Different Physical Activity Measurements in Eight European Countries: Results of the European Physical Activity Surveillance System (EUPASS) Time Series Survey. Public Health Nutrition, 6, 371–376.

Suurmeijer, T.P.B.M., Doeglas, D.M., Moum, T., Briancon, S., Krol, B., Sanderman, R., Guillemin, F., Bjelle, A., and Van den Heuvel, W.J. (1994). The Groningen Activity Restriction Scale for Measuring Disability: Its Utility in International Comparisons. American Journal of Public Health, 84, 1270–1273.

Van Buuren, S. and Hopman-Rock, M. (2001). Revision of the ICIDH Severity of Disabilities Scale by Data Linking and Item Response Theory. Statistics in Medicine, 20, 1061–1076.

Van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2001). Response Conversion: A New Technology for Comparing Existing Health Information. TNO report 2001.097, TNO Prevention and Health, Leiden (available from http://europa.eu.int/comm/health/ph_projects/1999/monitoring/fp_monitoring_1999_frep_05_en.pdf).

Van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2003). Assessing Comparability of Dressing Disability in Different Countries by Response Conversion. European Journal of Public Health, 13(Suppl 3), 15–19.

Van Deth, J.W. (1998). Equivalence in Comparative Political Research. In J.W. van Deth (ed.), Comparative Politics: The Problem of Equivalence. London: Routledge, 1–19.

Van de Vijver, F.J.R. and Leung, K. (1997). Methods and Data Analysis for Cross-Cultural Research. Thousand Oaks: Sage.

Van't Hof, M.A., Hautvast, J.G., Schroll, M., and Vlachonikolis, I.G. (1991). Design, Methods and Participation. Euronut SENECA investigators. European Journal of Clinical Nutrition, 45(Suppl 3), 5–22.

Verhelst, N.D. and Kleintjes, F.G.M. (1993). Toepassingen van Item Response Theorie. In T.J.H.M. Eggen en and P.F. Sanders (eds). Psychometrie in de Praktijk. [In Dutch], 285–310 Arnhem, The Netherlands: Cito Holding. [In Dutch].

Wainer, H. and Thissen, D. (1987). Estimating Ability with the Wrong Model. Journal of Educational Statistics, 12, 339–368.

Warm, T.A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika, 54, 427–450.

World Health Organisation (WHO) (1980). International Classification of Impairments, Disabilities and Handicaps. Geneva: World Health Organisation.

Received May 2002
Revised December 2004