TNO Prevention and Health

Physical Activity and Health
Wassenaarseweg 56
P.O. Box 2215
2301 CE Leiden
The Netherlands

TNO report

PG/B&G 2004.145 Final report

Response Conversion for the Health Monitoring Program

www.tno.nl

T +31 71 518 18 18
F +31 71 518 19 15
info-B&G@pg.tno.nl

Date	June 2004
Author(s)	S van Buuren, A Tennant (Eds.)
No. of copies	100
Number of pages	85
ISBN	90-5986-082-9
Sponsor	DG-Sanco, European Commission
Project name	International Rasch Conversion Centre for Community indicators (Phase 2)
Project number	011.41178

Summary

Lack of comparability is a major factor impeding the development of a co-ordinated European health information system. Response Conversion (RC) attempts to improve comparability in existing data.

RC addresses the situation where data are obtained using different questions and different response scales. By systematically exploiting the overlap of data, RC attempts to convert data into a common European scale. Where this can be done, comparisons can be made using the common scale.

RC consists of two steps. The first step involves the construction of a conversion key. This is a relatively complex activity, but needs to be done only once. The second step is the actual data transformation. This is relatively simple, and can be repeatedly done on a routine basis as new information arrives.

The current project aims to disseminate and apply RC within the Health Monitoring Programme. The project addressed the following tasks:

- Evaluation of the suitability of RC for data within the HMP;
- Construction of four new conversion keys;
- Development of a web site to support the conversion to indicators; Integration of RC into the IDA-HIEMS system.

We have taken the ECHI-2 draft short list, and assessed which topics from the short list would be amenable for Response Conversion. This resulted in a list of 26 indicator for which RC could be potentially useful.

Using a wide variety of data sources, new keys were produced for the following topics: physical activity, communication disabilities and sensory functioning, personal care, and physical well being. The report illustrates the use of these keys by comparing the position of Member States on the common scale.

A conversion key calculator was developed that allows users to generate conversion keys under different priors. During the project, development of the IDA-HIEMS system was discontinued, while no successor was yet planned. We therefore provided an alternative facility to generate syntax files with SPSS commands, thus allowing individual researchers to recode their data into a common scale. These facilities can be assessed on the internet at http://www.tno.nl/responseconversion.

All assumptions in RC are explicit. The conversion process takes small steps, is fully repeatable, and leads to verifiable quantitative results. Application of the method helps to evade some common pitfalls when dealing with cross-cultural comparability.

Many technical advances were made during the course of the project. Topics for further work were identified. We believe that further developments along these lines will strengthen the information system needed to advance European health policy.

Contents

1 Introduction

Stef van Buuren

1.1 Background

The goal of the Health Monitoring Program (HMP) of the European Commission (EC) is to provide relevant and timely information about the health situation in each member state (European Commission, 1998). To avoid unnecessary duplication, the health information system will have to be fed by a mix of existing and new data collected through health surveys performed in different Member States (MS). Although the content of surveys in different MS is often quite similar, substantial variations in the actual measurement exist, for example in sampling procedures, in the coverage of a given topic, in the wording of questions and response category formats. Thus inconsistency of information is a major problem. Such inconsistencies forbid straightforward comparability statements and call for comparability to be established in each and every case. In addition, each MS has its own tradition in collecting and processing health related data, and changing established ways of working usually takes time. In addition it may not only be complicated to change instruments or methods in practical terms; changes may also conflict with existing theoretical or institutional needs.

The current situation can be characterised as follows:

- MS are reporting data to a number of international bodies which implies multiple reporting;
- There is unnecessary duplication of effort;
- Data and information are often of limited comparability between countries and sometimes of medium or poor quality;
- There are significant gaps in the data available on a number of important diseases;
- The accession of 10 new MS in May 2004 with quite dissimilar statistical traditions poses new challenges for the comparability of information.

Incomparability may occur at different levels, for example:

- Appropriate data may not be collected at all in some MS;
- Some MS collect appropriate data for specific samples, or with special designs;
- The definition of diseases may differ between MS, through, for example, using different classifications;
- The wording of the question can differ;
- The formulation of the response categories can differ;
- Close translation of questions and response scales introduces differences in meaning.

Each of there problems can seriously affect comparability, and so each of these needs to be adequately addressed before a meaningful comparison between MS can be made.

Comparability problems may go unrecognised or where recognised, the complexity of the issues involved may seem daunting. While problems for some data can be resolved -- income distributions according to EU-Statistics on Income and Living Conditions (EU-SILC) are regarded as comparable across MS--the situation is less favourable for health data. At the EU level, it is thus no surprise that harmonisation efforts for health

data have been of rather limited success. Montserrat and Sicard (2004) say that "…there is a huge potential for improving the international comparability of health interview survey data".

Incomparability of data is *the key problem* in international comparisons. Recently, a Joint UNECE/WHO/Eurostat Meeting on Health Statistics was held in Geneva (UNECE, 2004). Nearly all contributors emphasised the need for comparable data. As comparability can occur at many levels, there are many approaches possible for improving comparability.

This report deals with a statistical technique, called Response Conversion, for improving comparability in the case that the wording of the question and/or the formulation of the response categories differ. This is a very common problem. In addition, the method can identify areas where comparability remains suspect, even after harmonisation. The method contributes to the correct data analyses of such data, and can assist in converting existing ('old') items into new scales.

1.2 Harmonisation methodologies

Several strategies have been developed to deal with incomparability. These can be broadly distinguished as (Grais, 1998; Günther, 2003) as

- Ex-ante input harmonisation;
- Ex-ante output harmonisation;
- Ex-post output harmonisation.

Input harmonisation is the strategy to attain maximal comparability across MS at the time of the survey design. It is always ex-ante harmonisation because it is done before data are sampled. For output harmonisation, both ex-ante and ex-post strategies are possible. Ex-ante harmonisation refers to the strategy in which procedures for achieving comparability are developed at the stage of survey design. Ex-post harmonisation is the strategy that could be tried if no ex-ante strategies were applied, or if these were not successful.

It will be clear that ex-ante harmonisation is generally preferable to ex-post harmonisation as it yields better guarantees for high quality data. However, the ex-ante strategy will not always be feasible. First, ex-ante harmonisation only works for new data. It cannot be applied if the data have already been sampled. Ex-ante harmonisation may slow down the uptake of scientific advances in measurement because logistic complexities like co-ordination and translation require time and resources. In addition, ex-ante harmonisation will not always work. The intuitively appealing and frequently applied 'Ask-the-Same-Question'(ASQ) model is not without pitfalls, so more principled methods are needed (Harkness, 2003; Smith, 2003). Finally, there is often a need to change established ways of working in environments with vested interests. There are limits to what Eurostat can achieve. The Director General of Eurostat wrote in 1998: "How a Member State chooses to organise its statistical service will depend on its own traditions and the structure of its civil service." (Franchet, 1998).

Incomparability is a measurement problem; it may arise from differences in constructs, in operational procedures or in instruments. However, not all differences lead to

incomparable data. Conversely, setting out to "keep everything the same everywhere" is no guarantee for comparability of data. In order to be comparable, data must either be shown to be unbiased or the bias most be known and controllable, i.e., be uniform bias. Full score or scalar equivalence is the highest level of equivalence allowing for item by item comparisons across instruments and the most sophisticated forms of analysis. Good methodological references are Van de Vijver and Leung (1997), Van Deth (1998) and Harkness *et al*. (2003).

This report focuses on ex-post harmonisation. Adequate techniques are still largely lacking, and where they exist, are not systematically explored. We like to emphasise that ex-post harmonisation alone, though helpful, is not enough. In practice, a combination of both ex-ante and ex-post strategies works best.

1.3 Response Conversion: General approach

The type of comparability problem considered in this report occurs because of differences in the formulation of survey questions and response categories. Suppose we want to get insight into the level of disability of the populations of different MS. Many MS conduct health surveys, but the precise way in which disability is measured could be quite different.

An example is walking disability. The U.K. 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 U.K.-item can be compared with one on the Dutch item.

Response Conversion (RC) attempts to transform responses obtained on the same topic but with different questions onto a common scale. Where this can be done, comparisons can be made using the common scale. The technique consists of two steps. The first step involves the construction of a conversion key. This is a relatively complex activity, but needs to be done only once. The second step is the actual data transformation. This is simple, and can be repeatedly done on a routine basis as new information arrives. Construction of the key is only possible if enough overlapping information can be found. References of the technique include Van Buuren *et al* (2001, 2003, 2004).

1.4 Project tasks

The project aims to disseminate and apply RC within the Health Monitoring Programme. The project addressed the following tasks:

1. Evaluation of the suitability of RC for data within the HMP;
2. Construction of four new conversion keys;
3. Development of a web site to support the conversion to indicators; Integration of RC into the IDA-HIEMS system.

Objective 1: The ECHI-1 (ECHI, 2001; Kramers, 2003) and ECHI2-projects (ECHI, 2004a, 2004b) have developed into integrative projects covering the Health Monitoring Programme. We have taken the ECHI-2 draft short list, and assessed which topics from the short list would be amenable for Response Conversion. Chapter 3 summarizes the results.

Objective 2: In addition to the existing keys for walking and dressing disability, we developed four new keys. Keys were produced for physical activity, communication disabilities and sensory functioning, personal care, and physical well being. Chapters 4-7 describe how these keys were developed.

Objective 3: A conversion key calculator can be assessed through the web site http://www.tno.nl/responseconversion. This allows users to generate conversion keys under different priors. During the project, development of the IDA-HIEMS system was discontinued, while no successor was yet planned. We therefore provided an alternative facility to generate syntax files with SPSS commands, thus allowing individual researchers to recode their data into a common scale. This facility can be assessed from the same web site. Chapter 2 provides more details on the methodology used, as well as the available tools.

2 Methodology

Stef van Buuren, Alan Tennant

2.1 General

Response Conversion (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: by 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, colors, 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 main 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 2.1 is an excerpt of data taken from 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. Likewise, 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.

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

Table 2.1 Small example: SI01 and GAR9 items linked by bridge item HAQ8.

Table 2.1 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 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:

$$p(Y_j = c \mid \theta) = \frac{\exp \sum_{k=0}^{c} (\theta - \delta_{jk})}{\sum_{r=0}^{k_j} \exp \sum_{k=0}^{r} (\theta - \delta_{jk})}, \qquad c = 0,1,\ldots, k_j \tag{2.1}$$

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 $p(Y_j = c \mid \theta)$ 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 (2.1) can only be estimated if the items are linked to each other. For example, if HAQ8 in Table 2.1 would not be present, there is 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 (2.1) implies that bridge items 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:

$$\delta_{jk}^{A} = \delta_{jk}^{B} = \delta_{jk} \qquad \text{for all } k = 1,\ldots,k_j \tag{2.2}$$

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

$$\delta_{ak}^{A} = \delta_{ak}^{C} = \delta_{ak} \qquad \text{for all } k = 1,\ldots,k_a, \text{ and} \tag{3a}$$
$$\delta_{bk}^{B} = \delta_{bk}^{C} = \delta_{bk} \qquad \text{for all } k = 1,\ldots,k_b. \tag{3b}$$

Equations (2.2) and (2.3) 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 non-equivalent linked grouped designs (Kolen and Brennan, 1995).

Note that no assumptions are required with respect to the level of disability in each study. Operationally, equations (2.2) and (2.3) imply that the linking information can be coded into the same data column(s) across different studies.

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

Table 2.2 Threshold estimates for the data in Table 2.1.

RUMM 2010 (RUMM Laboratories, 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 2.2 provides the threshold estimates obtained from the data in Table 2.1. The data fitted the Rasch model ($\chi^2 = 8.49$, df = 10, P=0.58).

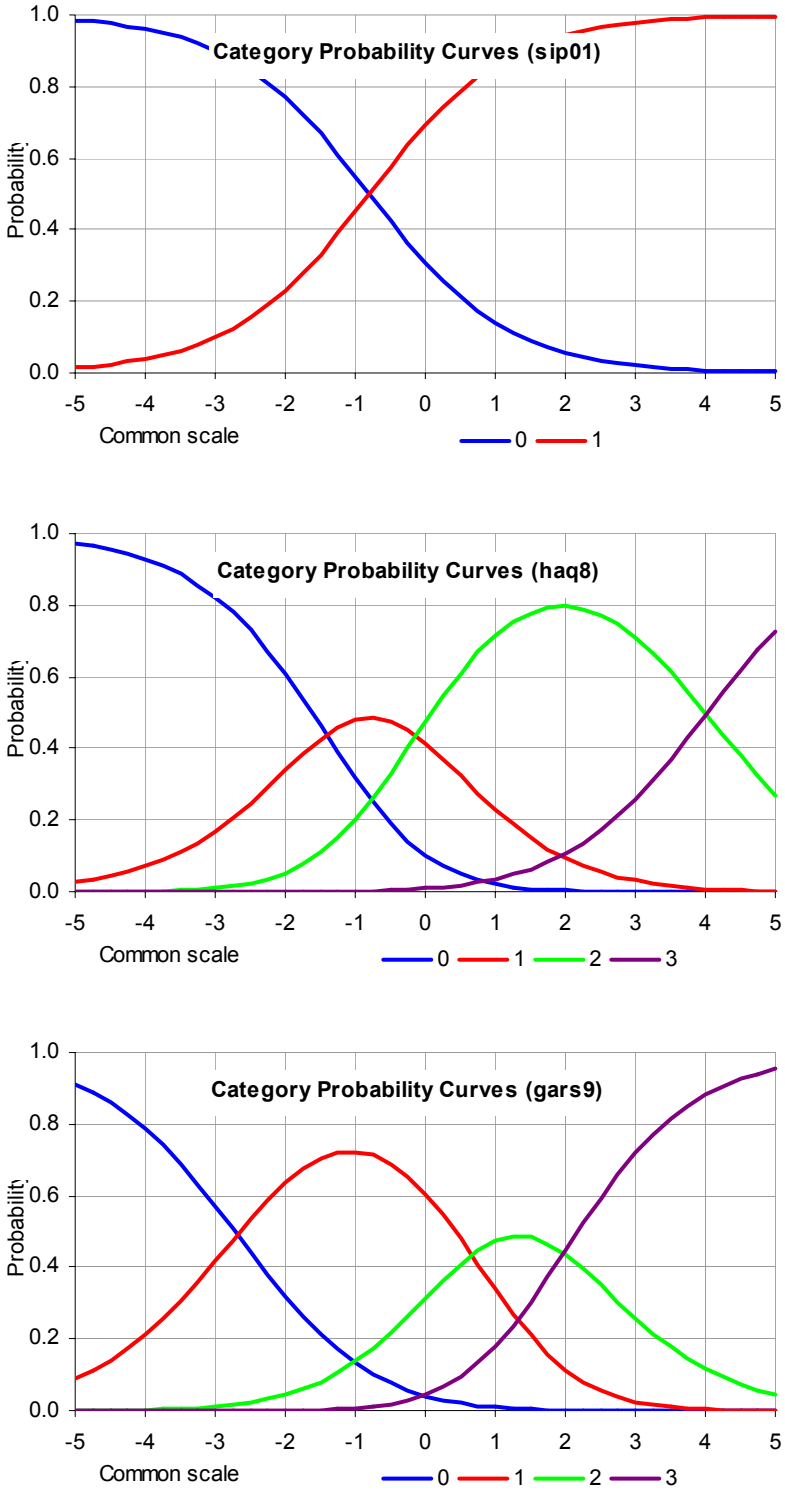

Figure 2.1 Probability of responding in each category as a function of walking disability. Based on the solution in Table 2.2.

Figure 2.1 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 to respond in those categories, and exhibit relatively low propensity to respond in the less disabled categories. In Figure 2.1, 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 θ_i = -1 has a probability of 0.62 of responding in category 0 of SI01, and a probability of 0.38 of answering category 1. The same person has probabilities of 0.27, 0.50, 0.23 and 0.00 to respond in respectively categories 0, 1, 2 and 3 of HAQ8. The response probabilities for GAR9 are respectively 0.11, 0.72, 0.16 and 0.01.

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 a 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, 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

$$p(\theta \mid Y_j = c) = \frac{p(Y_j = c \mid \theta)\, p(\theta)}{\sum_c p(Y_j = c \mid \theta)\, p(\theta)} \qquad \text{for } c = 0, \ldots, k_j \qquad (2.4)$$

It is easy to calculate $p(\theta \mid Y_j = c)$ on a grid of θ-values. A convenient choice for the grid is θ ={-5, -4.75, …, 4.75, 5}. The probability $p(Y_j = c \mid \theta)$ is given by model (2.1). The expression $p(\theta)$ is a Bayesian prior and summarises all information that we know before the current data. Choosing an appropriate prior takes some care, and we will come back to it later. For the moment, let us assume a uniform distribution $p(\theta) \sim$ U(-5,5).

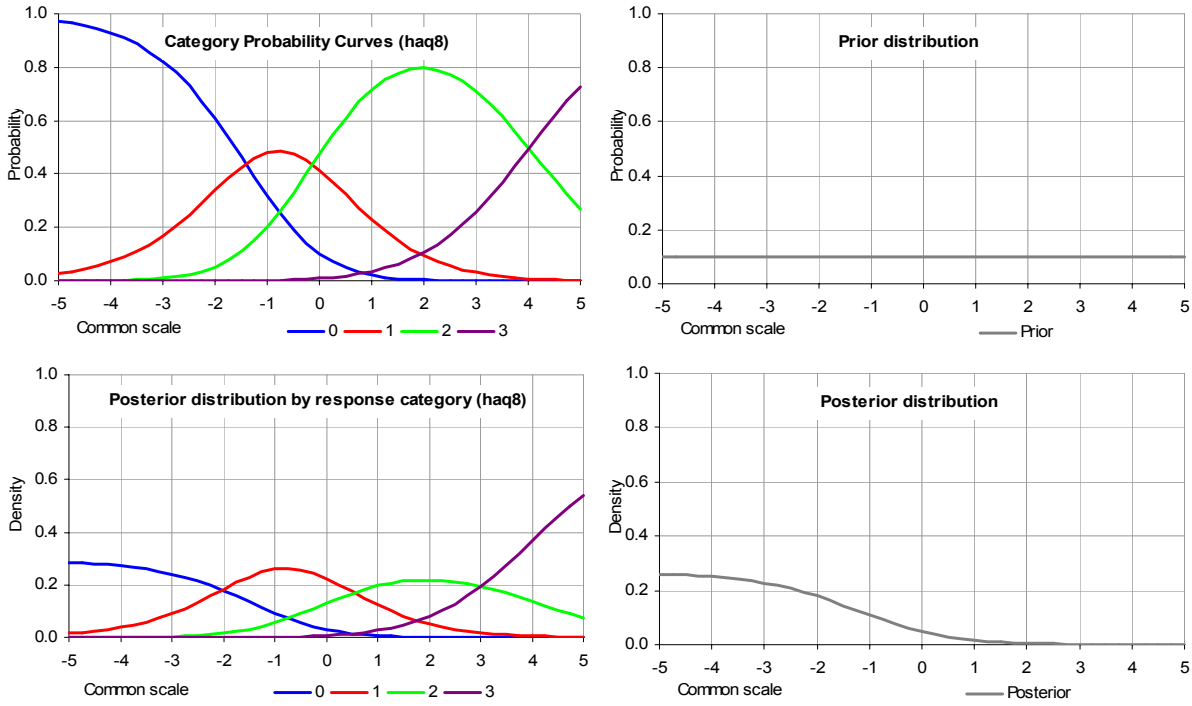

Figure 2.2 HAQ8 item "Able to walk outdoors on flat ground?": Category Probability Curves (a), uniform prior distribution U(-5,5) (b), Category Posterior Curves (c) and Sample Posterior

Figure 2.2 illustrates the relevant calculations for item HAQ8. Figure 2.2a and Figure 2.2b plot $p(Y_j = c \mid \theta)$ and $p(\theta)$, respectively. Figure 2c is the result of calculating (2.4), i.e., the posterior distribution per category $p(\theta \mid Y_j = c)$. This distribution describes what is known about θ after an answer is observed. Note that each $p(\theta \mid Y_j = c)$ is scaled to unit area, and can be interpreted as a density. Note also that $p(\theta \mid Y_j = c)$ is proportional to $p(Y_j = c \mid \theta)$, a consequence of the uniform prior. Figure 2.2d represents a mixture of the densities $p(\theta \mid Y_j = c)$, 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

$$p(\theta \mid Y_j) = \frac{\sum_{c=0}^{k_j} w_c \, p(\theta \mid Y_j = c)}{\sum_c w_c}, \qquad (2.5)$$

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

$$\mathrm{E}[\theta \mid Y_j] = \frac{\sum_c w_c \mathrm{E}[\theta \mid Y_j = c]}{\sum_c w_c}, \tag{2.6}$$

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 (2.4).

2.4 Conversion into the common scale

Equation (2.6) 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 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 2.3 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 2.1 clearly indicate that the EURIDISS sample possesses more disabilities than the ERGOPLUS sample. Both EURIDISS estimates (-1.80 and -1.93) are higher than the ERGOPLUS estimates (-2.18 and -2.54).

	Conversion key				Mean disability on the common scale	
Item	Response Category				ERGOPLS	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

Table 2.3 Recode values (conversion key) under a uniform (-5,5) prior, and the mean disability levels for the ERGOPLUS and EURIDISS samples in Table 2.1 expressed on the common scale per item.

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 would have measured only SI01 in ERGOPLUS and GAR9 in EURIDISS, then we still can 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 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 the model 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, by obtaining denser data, and so on. These are topics for further work and beyond the scope of this report. 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.

2.5 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 estimator is the optimal non-Bayesian choice. It yields unbiased ability estimates of the common scale, and is mathematically equivalent to the Bayesian estimator (2.5) 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 few 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 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. Vice versa, 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.

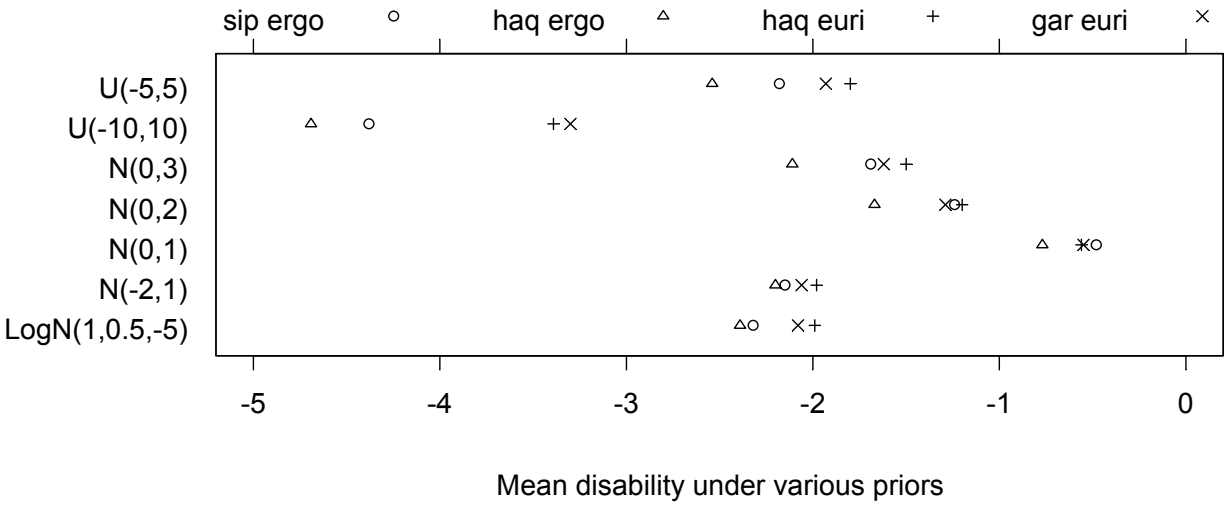

Figure 2.3 Mean disability for the following sample-item combinations: SIP01 administered in ERGOPLUS, HAQ8 administered in ERGOPLUS, HAQ8 administered in EURIDISS, and GARS9 administered in EURIDISS under two uniform priors, four normal priors and one shifted lognormal prior.

Figure 2.3 shows the mean estimates for the data in Table 2.1 under a variety of prior specifications. It will be immediately clear that estimates using different priors are incomparable to each other. Thus, all comparisons between samples should use a common prior for any comparison 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 would use ML estimation, where the prior is U(-∞, ∞). 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 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, centering 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 that disability distribution in the general population. In the absence of any data, we would expect more people in the low disability levels. The best guess of the ability of corresponds to a random draw from the population distribution. 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, and thus improve consistency among estimates of different items. For these reasons, we will choose priors that resemble the (pooled) population distribution.

The choice of the precise prior density will become irrelevant if ability is estimated from many items simultaneously. As long as the prior is not too informative, the data will quickly outrun the information provided by the prior, and produces similar ability estimates that will not depend on the starting prior.

2.6 Model fitting issues

Several model fitting issues should be addressed in the statistical analysis. This section discusses the different types of model fit issues, and the strategies used to address them.

2.6.1 Overall model fit

Several statistical tests exist for assessing overall model fit. A common problem of these statistics is that their significance levels are highly dependent on sample size. Since conversion keys are typically based on the analysis of data with thousands of records, the overall tests of fit are nearly always significant. Another technical problem is that the behaviour of the fit statistics is unknown under the type of incomplete data structures as in our applications. Therefore, relatively little attention will be paid to overall model statistics, and instead the analysis will focus on item fit.

2.6.2 Item fit

The assessment of item fit may reveal items that do not fit the common scale. Depending of the type of item misfit, various options are open to deal with a misfitting item:

- remove the item;
- collapse the categories into a smaller number;
- split the item over subgroups.

Various diagnostic measures are used to diagnose item fit. These include:

- RUMM residual fit;
- Outfit statistic;
- Threshold reversal;
- DIF effect estimate.

The RUMM residual fit statistic (RUMM Laboratories, 2003) is a summary statistic that measures the difference between the observed and expected counts calculated within classes of individuals based on the common scale estimate. The number of classes usually varies between 6 and 10. The statistic is standardised with zero mean and unit standard deviation. A value of zero means that the item acts exactly according to the model. Positive values beyond a cut-off value of, say, +3.0 indicate that the observed data are related less strongly than expected to the common scale. Negative values indicate a relation with the common scale that is stronger than expected, and are usually not considered to be a problem. A high RUMM residual may be lowered by splitting the item or by collapsing its categories, but these options may not always work. Items that still misfit after splitting or collapsing will be removed, and thus cannot be part of the conversion key. The RUMM residual statistic depends on sample size, which makes a choice of a cut point somewhat arbitrary.

The outfit statistic (Wright & Masters, 1982) can be calculated from the RUMM solution. The statistic is less sensitive to sample size. Though both statistics are theoretically identical, we found that they may lead to different conclusions. Chapter 7 contains a short comparison of both.

Threshold reversal occurs if the estimated thresholds have a different ordering than the item categories. Threshold reversal is considered to be a type of misfit, but it is not yet

clear whether it is a problem related to fit or a problem related to low frequencies. The usual response is to collapse categories until the reversal disappears.

Another type of misfit is Differential Item Functioning (DIF). DIF occurs of the relation between the common scale and the response probabilities depends on the group, for example, country. Occurrence of DIF implies that the transformation of data values into the common scale should depend on the group in order to be comparable. The conversion key can account for if the items are "split" over the group. In that case, country specific parameters conversion values can be derived.

Diagnosing DIF can be done in several ways. The simplest and most convincing method is to graph the relation between the common scale and the data scale separately for each group. The item contains no DIF if all mean curves coincide. RUMM 2020 has good graphing facilities that make this type of analyses easy. One could also look at the variation of the mean curve across the common scale. If this variation is large (say 1.0 logits), then there is DIF. DIF can also be established more formally through ANOVA by group. This has the same problem as before: For large samples, the test will always be significant and thus indicate DIF. An alternative is to estimate effect size, either from the ANOVA analysis or from special DIF oriented analyses outside RUMM. Various possibilities are used throughout Chapters 4 through 7. Where applicable, items with DIF will be split or eliminated.

2.6.3 Unidimensionality

A requirement of the Rasch model is that all items measure the same underlying trait. If so, the item set is said to be unidimensional. In practice, it is not so easy to establish unidimensionality of a set of items. Many different types of criteria have been put forward for assessing unidimensionality. See Hattie (1985) for an overview.

For the type of constructs studied in this report, it is usually fairly obvious to see that items share a common trait. Nevertheless, we performed additional analyses addressing unidimensionality where things appear less obvious. We applied two-dimensional homogeneity analysis/multiple correspondence analysis (Gifi, 1990) on the Physical Activity data in Chapter 4. Homogeneity analysis is a non-linear form of principal components analysis. The presence of a horse shoe in the person scores indicates that a one-dimensional solution would have sufficed. A strong horse shoe pattern emerged from the bridge items, convincingly indicating that the items measure the same trait. A more widespread application of this technique, i.e., by extending the item set to include country-specific items, was however hampered by the fact that the solution turned out to be severely affected by the missing data structure.

Another way of assessing unidimensionality is to apply principal components analysis on the residuals from the final Partial Credit model. If no dominant dimensions appear among the components, one can be fairly confident that the items are unidimensional. In practice, unidimensionality and good item fit often go together. Building on that observation, we concentrated on getting good item fits, as described in section 2.6.2. Removing badly fitting items will tend to improve unidimensionality. Therefore, no separate tests for unidimensionality are being applied.

2.7 Tools

Some tools make it easier to derive and apply actual conversions. Some users may wants to derive new keys on their data using the methodology described in this report. For these users, the *Quick Calculator* at http:/www.tno.nl/responseconversion allows to calculate conversion values under a prior of choice from the input threshold estimates from the Rasch model. The same can be done for a batch of threshold values for different items using the *File Converter*. In addition, the user can automatically generate SPSS-recode files that incorporate the appropriate recode values, thus saving the user from a lot of typing trouble. The SPSS syntax file can be applied to the user's data file to perform actual conversion. SPSS syntax files for all published conversion keys can be downloaded.

3 Response Conversion for indicators in the ECHI list

Astrid M.J. Chorus, Gert Jacobusse

3.1 Introduction

A first task in the project was to evaluate where Response Conversion could be applied within the Health Monitoring Program (HMP). To select indicators for assessment we based ourselves on the indicator list produced by the European Community Health Indicators projects ECHI-1 (ECHI 2001, Kramers, 2003) and ECHI-2 (ECHI, 2004a, 2004b) performed under the HMP. These projects developed a comprehensive list of indicators, in close co-operation with the other projects in the HMP.

The ECHI-2 shortlist (ECHI, 2004b) identifies in total 80 indicators, classified into 10 groups. An indicator is defined as a simple entity, most often numerical, which gives a quick insight into an important aspect of the field. For example, occupational class is an important indicator for socio-economic differences in health.

3.2 Method

Applicability of RC to ECHI-indicators was assessed with respect to a number of evaluation criteria reflecting technical conditions required for RC. Two members of the project team independently assessed the criteria for each indicator. Outcomes were compared and where differences occurred, assessment were discussed to reach consensus.

A set of five evaluation criteria was constructed for assessing the applicability of RC on indicators. Indicators should meet all these criteria. Table 3.1 lists the evaluation criteria.

Evaluation criteria for indicators
1. primary unit of measurement is an individual
2. the underlying scale is latent
3. the indicator comprises of a cut-off point on a continuous scale
4. no problems with differential item functioning (DIF) are to expected
5. data are available, or can be made available fairly easily

Table 3.1. Evaluation criteria for assessing applicability of RC

3.3 Results

Of the 80 indicators in the ECHI-2 short list, there was immediate agreement on 23 indicators to be included and on 38 to be excluded. Most of the excluded indicators did not meet the first criterion, usually because the primary unit of measurement was at the country level. Infant mortality is an example of an indicator at the population level.

Of the remaining 19 indicators, nine were questioned on meeting either the criterion 1 or 2. These were body mass index, perinatal conditions, blood pressure, serum

cholesterol, nutritional status, osteoporosis, breastfeeding, induced abortions and disability free life expectancy. The evaluation team decided that none of these met criterion 1 and 2, and thus these were removed from the list for which RC could be applied.

Ten indicators remained that were examined on criteria 3 and 4. These indicators were: short term activity restrictions, absenteeism from work, traffic behaviour, life events, waiting lists, perceived health, coping ability, sense of mastery, optimism, knowledge/attitudes. After some discussion, the evaluators judged that short term activity restrictions, absenteeism from work, and traffic behaviour did meet criteria 3 and 4, while the other did not. The total number of indicators to which RC was considered useful is thus 26.

	ECHI area	Indicator
		Demography and socio-economic situation
1	1.1.1	Urbanisation level
2	1.2.3	Educational attainment
3	1.2.5	Income level
		Health status
4	2.4.3	Functional limitations
5	2.4.3	Limitations in seeing*, Hearing*, mobility, speaking*, biting, agility
6	2.4.4	Activity limitations
7	2.4.4	Limitations of usual activities, past 6 months, health-related
8	2.4.4	Short term activity restrictions
9	2.4.3	General mental health
10	2.4.6	Psychological distress
11	2.4.7	General quality of life*
		Determinants of health
12	3.2.1	Smoking, alcohol use, (il)licit drug use
13	3.2.1	Regular smokers
14	3.2.1	Alcohol: % of heavy drinkers, frequency of heavy drinking
15	3.2.1	Use of illicit drugs (including children)
16	3.2.2	Energy indicators, consumption of macro/micro nutrients, contaminants
17	3.2.2	Intake of fruit excluding juice
18	3.2.2	Intake of vegetables excl. potatoes and juice
19	3.2.3	Physical activity, traffic behaviour
20	3.2.3	Physical activity (time spent, energy expenditure)*
21	3.3.1	Noise etc.
22	3.3.1	Environmental health indicator
23	3.3.3	Social support/ isolation, parental support for children, violence
24	3.3.3	Social and/or workplace indicator
		Health systems
25	4.3.4	Medicine use (total/specific)
26	4.3.4	Medicine use, selected items

*RC key developed within the present project.

Table 3.2 Indicators from the ECHI-2 short list for which Response Conversion could be useful.

3.4 Conclusion

We found that Response Conversion could be potentially applied to about one third of the ECHI-2 short list. This part of the ECHI-2 list covers many topics for which no traditional indicators yet exist. Of course RC might not actually be needed in all cases, because alternative ways of establishing comparability are also used on these indicators. An example is the 18 new HIS items currently being planned as part of a cross-national survey. On the other hand, it is likely that gaps will remain for which the RC technique is the most viable option.

The next four chapters introduce conversion keys for parts of the list in Table 3.2.

4 Physical activity

Gert Jacobusse, Stef van Buuren, Astrid M.J. Chorus, Marijke Hopman-Rock

4.1 Introduction

A wide variety of instruments for measuring physical activity is being used in different EU member states. Differences between these instruments make it difficult to compare physical activity outcomes between member states. In the ECHI-list, physical activity is classified as a determinant of health under code 3.2.3.

The European Physical Activity Surveillance System (EUPASS, 2001) project explored and advanced the International Physical Activity Questionnaire (IPAQ), a cross-nationally applicable set of indicators for measuring physical activity. Eight member states (Belgium, Finland, France, Germany, Netherlands, United Kingdom, Italy and Spain) took part in a study on this new set of indicators, together with existing indicators per member state. Thus, EUPASS is a study that provides multiple bridges on the item level. The common IPAQ items serve as bridge items to provide linkages between different indicators, as subjects from all member states responded to these items.

4.2 Method

4.2.1 International dataset

The international dataset that we use was collected by the EUPASS project (EUPASS, 2001; Rütten et al, 2003a/b). The goal of the project was to investigate the properties of a new measure, called International Physical Activity Questionnaire (IPAQ), a cross-nationally applicable set of indicators for measuring physical activity. The IPAQ was administered in eight countries, in combination with existing local instruments for measuring physical activity. The dataset contains almost 5000 cases, around 600 from each of the eight participating member states. Respondents were randomly selected on a nation-wide basis, and contacted using a computer-aided telephone interview. Although major efforts were made to standardize sampling procedures and fieldwork in the participating member states, response rates were not so high and varied between 25.5 % (UK) and 54.5% (Finland).

We preprocessed the EUPASS data by

- changing all missing codes to empty cells, to prevent them from being confused with real data;
- recoding variables in such a way that the lowest value indicates the highest physical activity;
- deleting old items that had been combined into new variables, to avoid double information in the data.

There were 10 common items administered in all countries and there were in total 41 national items available for analysis. The corresponding linkage diagram, showing

which items are administered in which study, is given in Table 4.1. See the EUPASS documentation for more detail (EUPASS, 2001).

Table 4.1 Linkage structure of the EUPASS data (EUPASS, 2001).

Item	Label	Cat	Country**					
			Be	Fi	Ge	It	NL	UK
eup1	At what pace usually walk	3	Y	Y	Y	Y	Y	Y
eup2	How much pa in place of work last 7 days	3	Y	Y	Y	Y	Y	Y
eup3	How much pa for purpose of transportation last 7 days	3	Y	Y	Y	Y	Y	Y
eup4	How much pa in and around home last 7 days	3	Y	Y	Y	Y	Y	Y
eup5	How much pa recreation, sport, leisure time	3	Y	Y	Y	Y	Y	Y
eup12	how much time in usual week doing vigorous pa	C	Y	Y	Y	Y	Y	Y
eup13	how much time in usual week doing moderate pa	C	Y	Y	Y	Y	Y	Y
eup14	how much time in total you spend on walking in a usual week	C	Y	Y	Y	Y	Y	Y
eup16	sitting weekday: sum in minutes for 1 day	C	Y	Y	Y	Y	Y	Y
eup18	sitting weekend: sum in minutes for 1 day	C	Y	Y	Y	Y	Y	Y
d01_b	Belgium: sweating at least 1 time per week	2	Y					
f01_b	Belgium: on how many days per week	7	Y					
b01_fin	Finland: Leisure time pa for at least half an hour (at least l sw)	7		Y				
c01_fin	minutes a day walking, running or riding a bicycle to/f work:HC	6		Y				
d02_fin	demanding job physically (recoded)	4		Y				
e01_fin	how much exercise or pa in free time (recoded)	4		Y				
b02_d	Germany: How often engaged in sports/ strenous activities	5			Y			
d03_d	Germany: Get out of breath after climbing 3 floors	2			Y			
e02_d	Germany: How often do you participate in sports?	5			Y			
e03_d	Germany: Time spend per day sleeping (M-F)	C			Y			
e04_d	Germany: Time spend per day sitting (M-F)	C			Y			
e05_d	Germany: Time spend per day light activities (M-F)	C			Y			
e06_d	Germany: Time spend per day moderate activities (M-F)	C			Y			
e07_d	Germany: Time spend per day strenous activities (M-F)	C			Y			
e08_d	Germany: Time spend per day sleeping (Weekend)	C			Y			
e09_d	Germany: Time spend per day sitting (Weekend)	C			Y			
e10_d	Germany: Time spend per day light activities (Weekend)	C			Y			
e11_d	Germany: Time spend per day moderate activities (Weekend)	C			Y			
e12_d	Germany: Time spend per day strenous activities (Weekend)	C			Y			
f02_d	how often are you engaged in sports or other streneous: HC	4			Y			
b03_i	regular sporting activities in free time: (recoded)	2				Y		
b04_i	occasional sporting activities in free time (recoded)	2				Y		
b05_i	Italy: How many month in total	C				Y		
b06_i	consider all the sporting activities over past 12 mo: HC	6				Y		
b07_i	any type of physical activity at least twice a year: HC	4				Y		
l01_i	Italy: How many activities	5				Y		
l02_i	Italy: Sporting activities requiring payment	2				Y		

l03_i	Italy: Practice requiring payment (lessons)	2				Y		
l04_i	Italy: Annual (periodic) fee for sport club	2				Y		
b08_nl	NL: sum a3	c					Y	
b09_nl	NL: how many times a day do you walk	c					Y	
c02_nl	NL: sum a4, how many times sports or exercise	c					Y	
f03_nl_	NL: sum of minutes heavy PA yesterday	c					Y	
f04_nl_	NL: sum of minutes moderate PA yesterday	c					Y	
f05_nl_	NL: sum of minutes light PA yesterday	c					Y	
l01b_nl	how many sports, computed from A02	5					Y	
l15_nl	NL: did you sport yesterday?	2					Y	
a06_uk	gardening, diy or building work done in the past 4 weeks: HC	2						Y
b10_uk	any exercise or sport during the last 4 weeks: gehercodeerd	2						Y
d04_uk	was the effort or activity usually makes you out of breath	2						Y
g01_uk	walking of a quarter of a mile done locally or away from: HC	2						Y

* number of categories, 'c' for continuous item

** in France and Spain, only the common first 10 items were collected

4.2.2 *Equivalence assumptions*

The conversion methodology assumes that, for the common items, the relation between the trait and the response probability is equivalent across countries. We call this the equivalence assumption. If this assumption does not hold, we say that there is differential item functioning (DIF) by member state (Holland and Wainer, 1982). It is possible to test the DIF assumption. If we find significant DIF for some item, we cannot assume that the data it produces are comparable across MS. In such cases, we will treat the item as if it were different items administered in different countries. Thus, the conversion methodology will only use information from items that are cross-nationally equivalent.

Some of the physical activity variables in Table 4.1 have continuous responses (like 'number of minutes'). In order to fit a Rasch model, such responses are recoded into categories. This can be done in several ways. We used two different approaches towards this categorization: one aimed at an optimal fit to our model (usually coding data into a small number of categories), the other at a minimal loss of information (essentially coding responses to a large number of categories). These approaches lead to two different conversion keys.

4.2.3 *Recoding strategy 1: Optimal model fit*

Under the Rasch model, the probability to respond in one of the categories of an item is a function of a person's location on the underlying trait, see Figure 4.1.

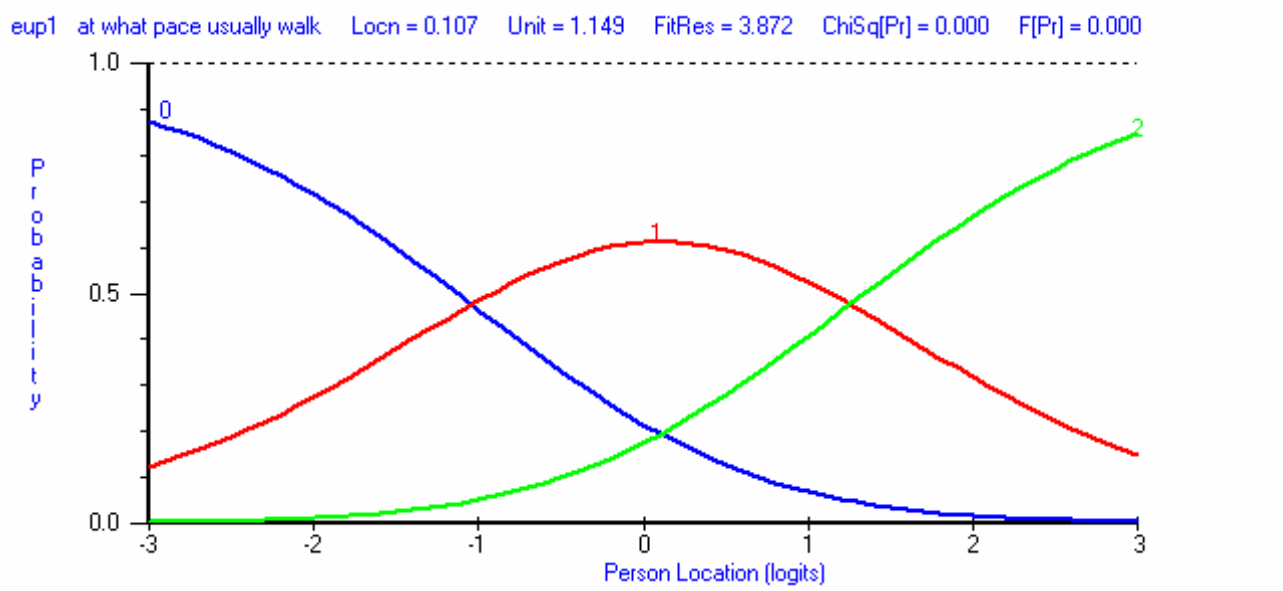

Figure 4.1 Category probability curve of item EUPASS1, with ordered thresholds

The threshold between two adjacent categories is the location for which the chance to respond in one or the other category is equal. This corresponds to the location where the curves of adjacent categories cross. The first threshold (from 0 to 1) in Figure 4.1 lies just below -1, the second just above 1. The threshold estimates are not made up by the investigator, but are estimated from the data at hand. Ideally, the response categories within each item are ordered, so one criterion for evaluation of the quality of the model fit is the ordering of thresholds (Andrich *et al.*, 1997).

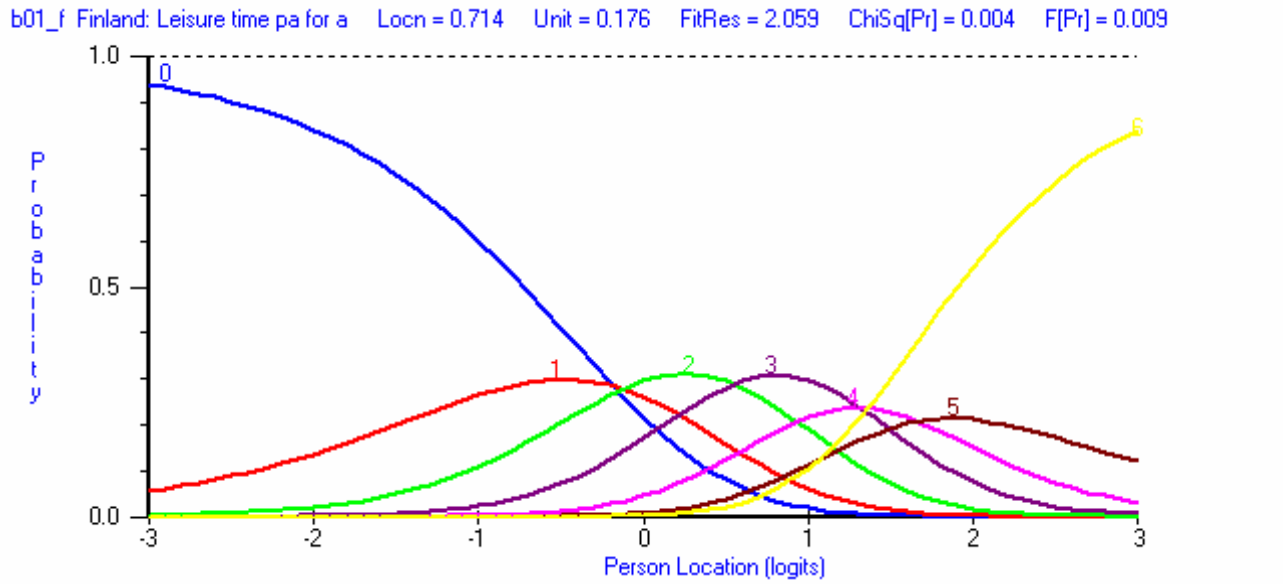

Figure 4.2 Category probability curves of item B01_fin with reversed thresholds.

Figure 4.2 gives an example were thresholds are reversed. The location of the sixth threshold (from category 5 to 6) lies left of the fifth threshold (from 4 to 5). The chance on obtaining reversed thresholds increases if the item has more categories. In principle, more categories can hold more information, so that slight deviations from the model are more likely to result in a discrepancy between the data and the model. Thus, in the

optimal model fitting strategy, we categorize continuous variables into a relatively small number of categories. Based on an exploratory analysis, we decided to categorize the 22 continuous items into 4 categories. It turned out that some 4-category items (still) had reversed thresholds, and these were recoded into 3 categories. We always tried to identify conceptually sensible intervals for the categorization.

4.2.4 Recoding strategy 2: Minimal information loss

The disadvantage of crude categorization is the potential loss of information that was present in the original continuous variables. In order to make a precise and powerful conversion key, we should restrict this loss of information to a minimum. In our second approach, we want to preserve as much information as possible; therefore we categorize variables using two criteria. First, we categorize variables based on equally spaced intervals, so that each interval represents the same range width of the continuous item scale. As a consequence, the distribution of subjects over category frequencies resembles the distribution of subjects over the original continuous scale. Second, we categorize each variable into a maximal number of categories. Both approaches may lead to sparse data, i.e. categories may contain very few or even zero observations, which may lead to estimation problems. We carefully aggregated categories until categories were sufficiently filled to be able to fit a model. Up to fifty-one categories per item were distinguished.

4.2.5 Item discrimination

The ability to distinguish persons that have different positions on the underlying trait, physical activity, varies between items. Items that discriminate very well at a certain value of the trait are generally beneficial in the sense that these items resemble the rest of the scale. On the other hand, items that discriminate poorly have less in common with the underlying trait. We eliminate such items from the conversion key as they weaken the basis of the common scale. The two types can be discriminated on the basis of their sign of the residual fit statistic. A negative residual is associated with a strongly discriminating item (c.f. Figure 4.3), where a positive residual indicates weak discrimination (c.f. Figure 4.4). Thus, items with large positive residuals (e.g. > 3.5) do not fit the model and are candidates for removal.

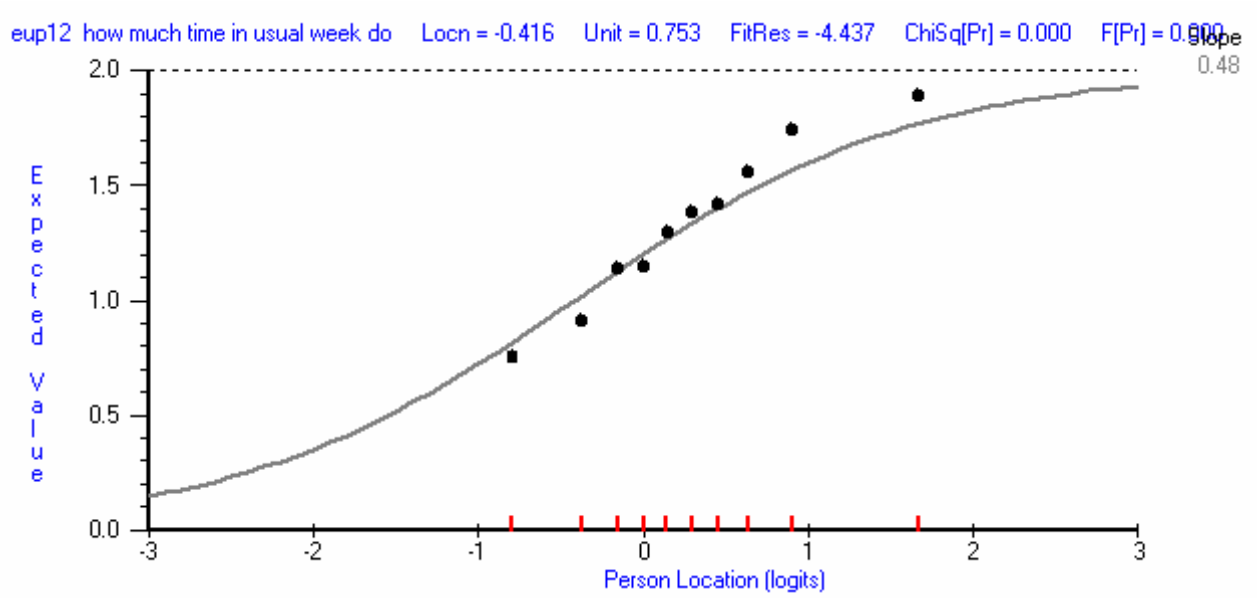

Figure 4.3 Item characteristic curve of eup12, with a negative item fit residual.

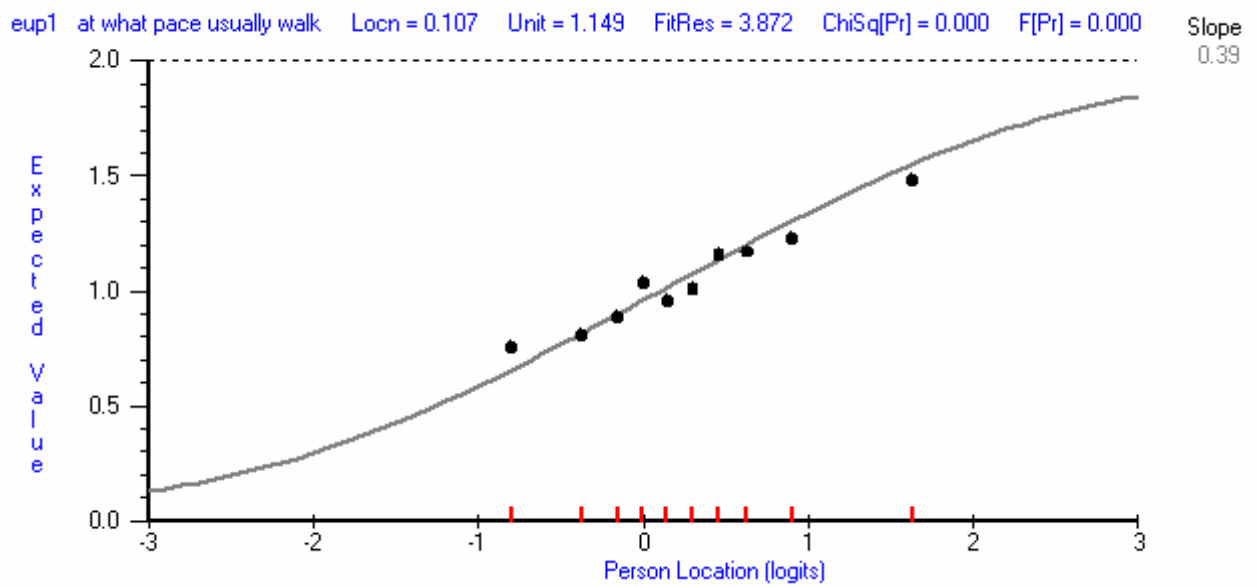

Figure 4.4 Item characteristic curve of eup1, with a positive item fit residual

4.2.6 Person fit residuals

Parallel to the item fit residuals, one can also compute the person fit residuals. A similar interpretation holds for the persons fit residuals. Persons with a high fit residual show a very atypical pattern of responses, often caused by responding at the same end of an item scale, even though items are reversed. This inconsistency in responses could indicate that persons do not take the questions seriously, or at least, not in the way intended by the test developer.

Persons with a high fit residual don't influence the item fit residuals dramatically if the sample size is big enough, but they do reduce variation among the estimated item and threshold locations (Curtis, 2001).

4.2.7 Steps to optimize the measurement properties of the common scale

In order to increase the quality of the conversion key, we will delete items and persons with high positive fit residuals - indicating that they don't conform to the common measurement scale for physical activity. We distinguish between three 'optimization steps', which are applied to both datasets created under the different categorization strategies:

1. Delete items with a positive residual higher than 3.5, run a new analysis, and delete items with a residual higher than 3.5 again. Repeat this until all fit residuals are smaller than 3.5.
2. Delete persons with a positive residual over 3.5.
3. Repeat step 1.

According to Curtis (2001), deleting persons with a high residual in step 2 will not change the item fit residuals that much, so step 3 should not lead to the deletion of many items.

4.3 Results

4.3.1 Recoding strategy 1: optimal model fit

After 22 continuous variables were categorized into 4 categories (six items) or 3 categories (sixteen items), there remained 16 items with reversed thresholds, of which only 6 were former continuous variables. The other 10 reversed thresholds occurred in items that already had been collapsed. We did not try to re-categorize these items, as this would disturb the comparison between our two methods to categorize continuous variables. Thresholds of items with ordered thresholds are visualized in Figure 4.5, the threshold map. In this initial solution, the item fit residuals range from -4.4 to 5.5, with mean 0.466.

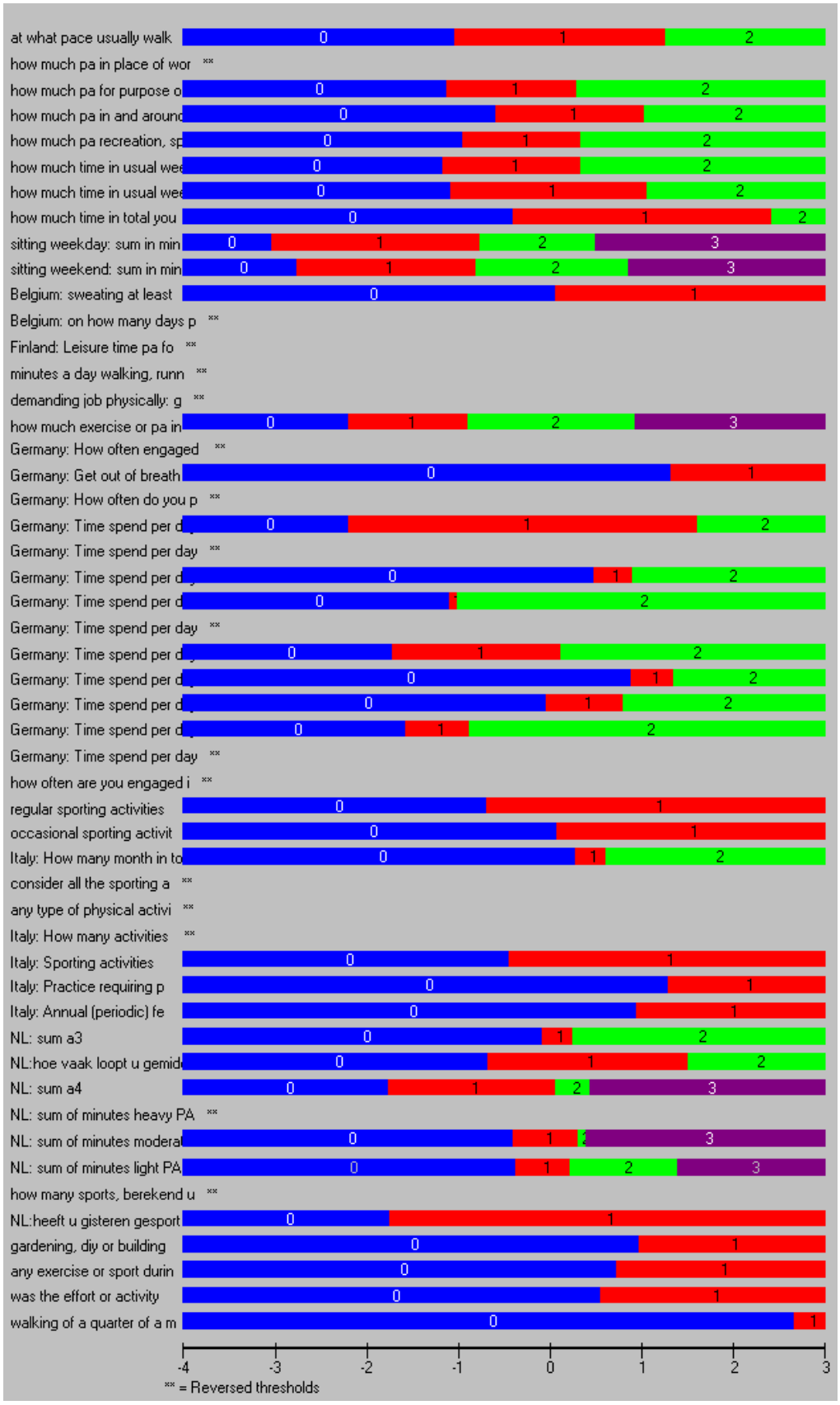

Figure 4.5 Threshold map of the initial solution in the dataset under the optimal model fit strategy.

After a sequence of four analyses in which items with a fit residual over 3.5 were deleted (optimization step 1), all items now have a fit residual lower than 3.5, actually the highest fit residual is now 2.561. Seven items were deleted. These are listed in Table 4.2.

Step	Item	Residual	Item label
1	f05_nl	5.508	Sum of minutes light PA yesterday
	eup4	4.745	How much PA in and around home last 7 days
	a06_uk	3.902	Gardening, diy or building work done in the last 4 weeks
	eup1	3.872	At what pace usually walk
2	eup18	3.773	Sitting weekend: sum in minutes for one day
3	eup16	6.343	Sitting weekday: sum in minutes for one day
4	l02_i	3.514	Sporting activities requiring payment

Table 4.2 Maximal fit strategy: Deleted items with high positive residuals statistics

For most items, it is not difficult to understand why they were deleted. Light physical activity and walking indicate only a slight degree of physical activity, probably not sufficient to discriminate enough between persons with a different position on the physical activity scale. Sitting items (eup16 and eup18) are indicators of the opposite of physical activity. Note that after eup18 was deleted, the scale changed so much that the residual of eup16, the other sitting item, rose over 6. This shows how a lack of conceptual overlap with other items immediately translates itself into a high item fit residual.

Optimization step 2 leads to 61 persons (1.2 %) with a person fit residual over 3.5. As expected, deleting these persons had almost no effect on the item fit residuals, and no more items had to be deleted in optimization step 3.

Using this solution, we investigated differential item functioning (DIF) of the common items. ANOVA showed that all bridge items have statistically significant DIF ($P < 0.001$) between member states. Note however that at a sample size of n=5000 even small differences between member states become statistically significant. The ANOVA test is therefore very stringent, and probably too strict for our purposes.

Figure 4.6 is a graphical representation of the DIF between member states in IPAQ item EUPASS3 ‘PA for purpose of transportation’; it shows the relation between the common scale and the mean item score for each member state.

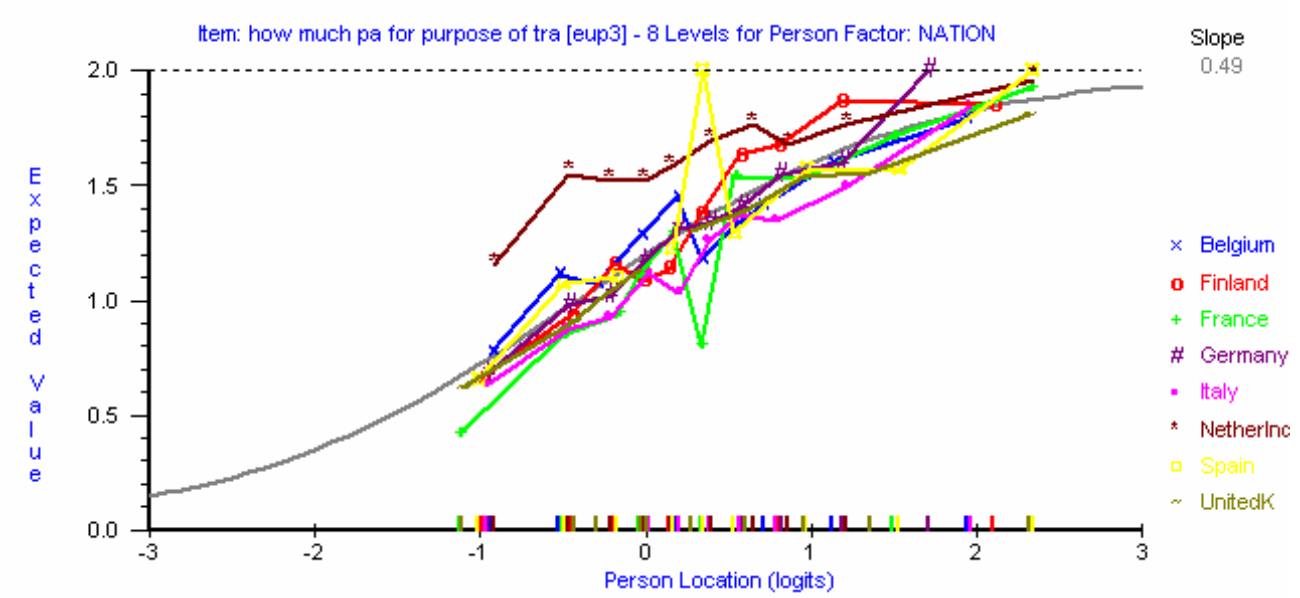

Figure 4.6 DIF by member state in item EUPASS3: PA for purpose of transportation.

Ideally, the curves for all countries should coincide. It appears however that persons from the Netherlands with much physical activity (on the left of the scale) have a higher average score on item EUPASS3 than people from other MS. Thus, it seems that EUPASS3 is behaving differently in The Netherlands. A similar observation holds for Finland. The Finnish sample has somewhat higher mean scores on EUPASS3 at the upper end of the scale. The other countries are, by and large, in agreement with each other. The occasional peaks for Spain and France are caused by small samples at these points, and can be safely ignored. For example, the peak in the Spanish line is based on just one person who has that location on the underlying physical activity trait, and who scored category 2 of item EUPASS3.

The DIF analysis shows us that we could use EUPASS3 to link up six of the eight MS. We will regard EUPASS3 as a different item in The Netherlands and in Finland, and regard the answers on the EUPASS3 item from these countries as incomparable to those of the other MS.

DIF for all 10 bridge items was investigated in the same way, splitting up items until there was no significant ($p<0.001$) DIF left. This resulted in the item splits as in Table 4.3. While some items show DIF for some MS, the remaining linkages across MS are quite strong and free of DIF.

	Be	Fi	Ge	It	NL	UK	Fr	Sp
eup2	A	A	A	B	A	A	A	C
eup3	A	B	A	A	C	A	A	A
eup5	A	A	A	B	C	A	A	A
eup12	A	B	B	C	A	A	A	D
eup13	A	A	B	A	C	A	D	E
eup14	B	A	A	A	C	D	E	F

Table 4.3 Item split table under recoding strategy 1. Comparable items have identical letters on the relevant row.

4.3.2 *Recoding strategy 2: minimal loss of information*

Categorizing continuous variables based on equally spaced intervals leads to items with very skewed distributions. Some persons reported extreme amounts of physical activity (like up to 98 hours a week vigorous physical activity) that lead to empty or almost empty categories. Figure 4.7 shows the distribution after categorization of IPAQ item eup12.

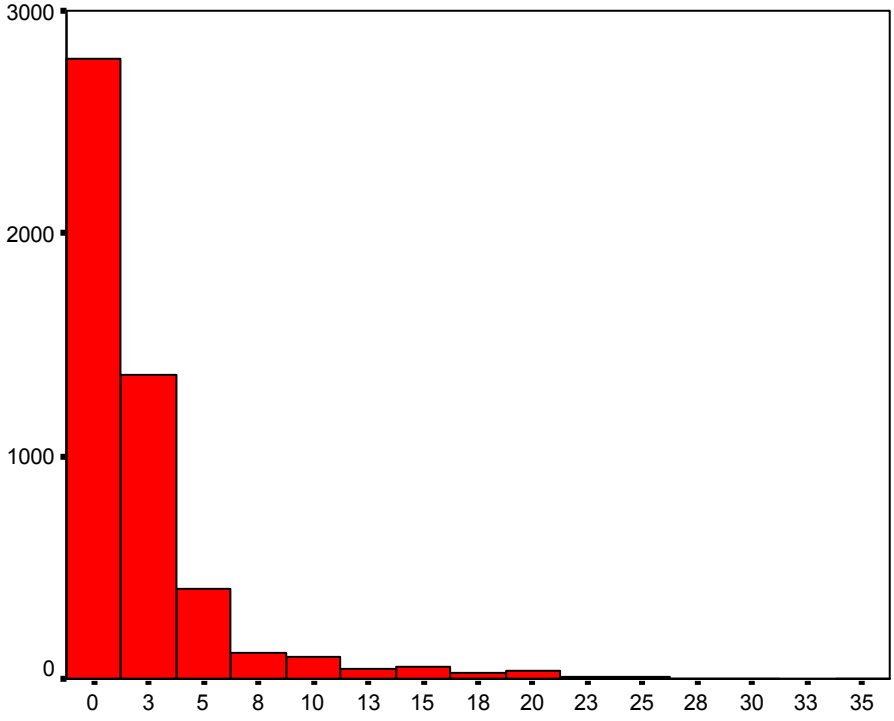

Figure 4.7 Distribution of eup12: How much time in usual week doing vigorous PA.

In a preliminary analysis of all 51 items, the item fit residuals ranged from -8.6 to +5.8, with mean 0.278. In optimization step 1, a total of 13 items was deleted in a sequence of 6 analyses. After analysis 6 there were no items with a residual higher than 3.5 left. The items that were deleted during this step are shown in Table 4.4.

Step	Item	Residual	Item label
1	eup1	5.838	At what pace usually walk
	d04_uk	3.913	Was the effort/ activity enough to make you out of breath
2	d01_b	3.837	Sweating at least one time per week
	e05_d	3.610	Time spend per day light activities (M-F)
	a06_uk	3.584	Gardening, diy or building work done in the last 4 weeks
	e10_d	3.574	Time spend per day light activities (weekend)
3	eup4	3.777	How much PA in and around home last 7 days
	b04_i	3.766	Occasional sporting activities in free time
	e02_d	3.589	How often do you participate in sports
4	b10_uk	3.790	Any exercise or sport during the last 4 weeks
	eup5	3.751	How much pa recreation, sport, leisure time

5	c01_fin	3.536	Minutes a day walking, running, riding to/from work
6	b08_nl	4.096	Sum A4, how many times sports or exercise

Table 4.4 Minimal loss strategy: Deleted items with high positive residuals statistics.

For most items in Table 4.4 it can be understood why they don't discriminate well with respect to physical activity. Walking and light activities indicate only a slight degree of physical activity. The high residuals for these items suggest that perhaps very light activities are intrinsically different from the other activities. Sweating or getting out of breath can indicate physical activity, but may also be a sign of a bad physical condition. PA around home or in free time and PA for work may also be somewhat different types of activities.

In step 2, we found 56 persons (1.1 %) with a person fit residual over 3.5. Deleting these persons had a surprisingly big impact on some of the item fit residuals. The item fit residual of item f03_nl, for example, changed from -0.723 to 3.594. In step 3, this item and item l02_i, with an item fit residual of 3.557, were deleted.

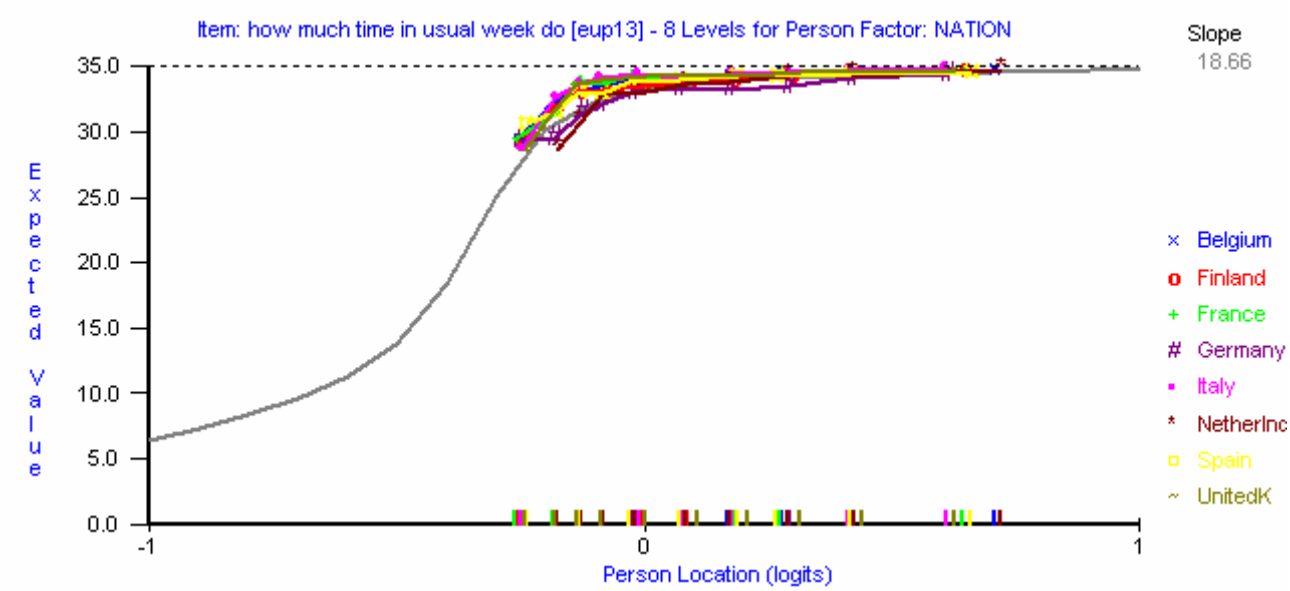

Figure 4.8 DIF by member state in item eup13 time per week moderate PA

As before, all bridge items show significant DIF ($P < 0.001$). The ICC as a graphical representation of DIF (Figure 4.8) demonstrated that a new problem emerges. The person locations only cover a very limited range of the physical activity trait. Most persons are located within the interval [-1,1]. We tried to increase precision at the item level by using many categories. The side effect of this turned out to be a compression of the whole score range towards the centre. Thus, in effect we find a solution that gives us less information about the location of persons on the underlying physical activity trait.

The lines that represent the mean observed scores only cover a part of the curve in Figure 4.8. Only very few people had a score between 0 and 25 (vertical axis). The small coverage of the lines makes it difficult to judge the amount of DIF, though generally the lines seem to follow the curve.

4.3.3 Recoding strategies: Conclusions

Two strategies were used to categorize continuous variables. The first method aimed at an optimal model fit, using the criterion of a minimal number of reversed thresholds, usually resulting in items with three or four categories. The second aimed at a minimal loss of information during categorization, using the criterion of a maximal number of categories, based on equally spaced intervals, resulting in up to 51 categories.

The minimal loss strategy led to a higher number of items of misfitting items (compare Table 4.2 with Table 4.4). This was to be expected as the potential discrepancy between the data and the model is likely to be larger in data with more information. We also observed that the item fit residuals are far less stable under minimal loss of information. An explanation for this instability could be the presence of items with empty or almost empty categories. Sparse data generally lead to less stable solutions. The variance among the estimated person locations turned out to be much smaller under the minimal loss strategy. This is somewhat counter-intuitive at first sight. It may seem attractive to use a large number of categories for continuous variables. The downside of this is that many categories will not be filled properly, but still take up space on the trait. Thus, in a sense, the scale may be too long for the sample at hand under the minimal loss strategy.

Based on these observations, we conclude that the best way to categorize skewed continuous variables into a small number of categories, even though this leads to loss of precision between adjacent categories. Within the context of the Rasch Model, this strategy seems best for three reasons: (1) a smaller number of original items has to be deleted (2) the solution is more stable and (3) the variance among estimated person locations is larger, thus resulting, somewhat unexpectedly, in a more precise measurement. Note that this conclusion pertains to skewed variables, but we expect similar phenomena will hold for symmetric data.

4.4 Conversion key for Physical Activity

The conversion key is calculated on the analysis of optimal model fit strategy, where some bridge items were split because of DIF as in Table 4.3. It appears that there was only one item with a fit residual > 3.5, eup14c, a split group item for Netherlands. It would be impractical to exclude a bridge item for one member state and not for others, so we stick to the complete set of items that we selected during the steps for optimization. The distribution of item fit details are given in Figure 4.9.

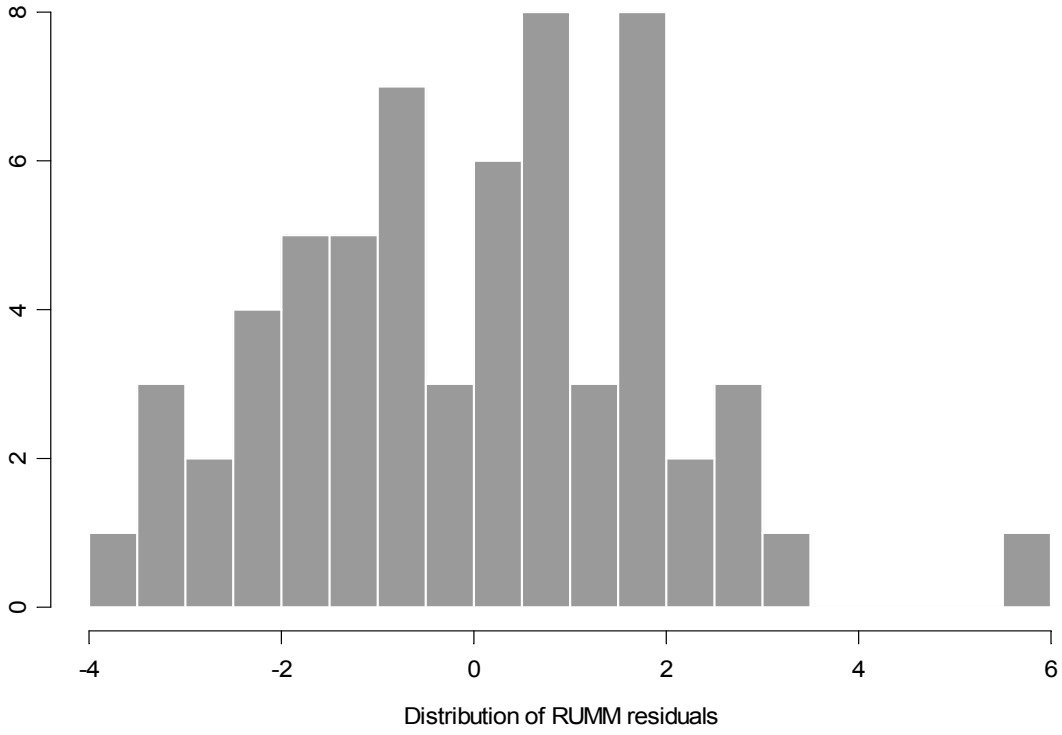

Figure 4.9 Frequency distribution of the RUMM fit residuals for the Physical Activity conversion key.

The splitting of some items implicates that those items will have member state-specific recode values. The recode values of the physical activity items (the conversion key) are given in Table 4.5. These have been calculated using a normal prior (c.f. Chapter 2).

Item	Member States	Category						
		0	1	2	3	4	5	6
eup2_i	It	-2.220	-1.430	-0.714				
eup2_s	Sp	-2.237	-1.458	-0.753				
eup2_rem	Be,Fi,Ge,NL,UK,Fr	-2.259	-1.505	-0.824				
eup3_f	Fi	-2.403	-1.637	-0.899				
eup3_n	NL	-2.382	-1.653	-0.970				
eup3_rem	Be,Ge,It,UK,Sp	-2.313	-1.536	-0.807				
eup5_i	It	-2.313	-1.526	-0.780				
eup5_n	NL	-2.205	-1.383	-0.621				
eup5	Be,Fi,Ge,UK,Fr,Sp	-2.332	-1.569	-0.855				
eup12_fg	Fi,Ge	-2.282	-1.489	-0.744				
eup12_i	It	-2.586	-1.839	-1.097				
eup12_s	Sp	-2.410	-1.673	-0.974				
eup12	Be,NL,Fr,UK	-2.370	-1.604	-0.874				
eup13_fr	Fr	-2.337	-1.564	-0.834				
eup13_g	Ge	-2.239	-1.425	-0.658				
eup13_n	NL	-2.286	-1.473	-0.691				
eup13_s	Sp	-2.391	-1.617	-0.871				
eup13	Be,Fi,It,NL,UK	-2.319	-1.526	-0.769				
eup14_b	Be	-2.154	-1.298	-0.495				
eup14_fr	Fr	-2.151	-1.292	-0.487				
eup14_n	NL	-2.116	-1.238	-0.418				
eup14_s	Sp	-2.155	-1.289	-0.459				
eup14_u	UK	-2.297	-1.473	-0.661				
eup14	Fi,Ge,It	-2.220	-1.386	-0.590				
d01_b	Be	-2.122	-1.238					
f01_b	Be	-2.411	-1.871	-1.480	-1.176	-0.908	-0.632	-0.310
b01_f	Fi	-2.260	-1.533	-0.927	-0.416	0.030	0.438	0.827
c01_f	Fi	-2.509	-1.876	-1.361	-0.938	-0.563	-0.190	
d02_f	Fi	-2.473	-1.788	-1.183	-0.630			
e01_f	Fi	-2.653	-1.970	-1.312	-0.659			
b02_d	Ge	-2.306	-1.569	-0.923	-0.355	0.162		
d03_d	Ge	-2.048	-1.112					
e02_d	Ge	-2.254	-1.532	-0.943	-0.460	-0.033		
e03_d	Ge	-2.515	-1.715	-0.906				
e04_d	Ge	-2.113	-1.267	-0.518				
e05_d	Ge	-2.127	-1.276	-0.508				
e06_d	Ge	-2.431	-1.728	-1.064				
e07_d	Ge	-2.404	-1.732	-1.105				
e08_d	Ge	-2.500	-1.750	-1.017				
e09_d	Ge	-2.088	-1.205	-0.401				
e10_d	Ge	-2.185	-1.366	-0.619				
e11_d	Ge	-2.503	-1.799	-1.119				
e12_d	Ge	-2.485	-1.832	-1.205				
f02_d	Ge	-2.049	-1.155	-0.379	0.262			

b03_i	It	-2.199	-1.346				
b04_i	It	-2.101	-1.205				
b05_i	It	-2.147	-1.313	-0.561			
b06_i	It	-2.220	-1.460	-0.830	-0.319	0.115	0.516
b07_i	It	-2.133	-1.276	-0.500	0.186		
l01_i	It	-2.453	-1.854	-1.364	-0.934	-0.510	
l03_i	It	-2.054	-1.123				
l04_i	It	-2.071	-1.154				
b08_n	NL	-2.184	-1.372	-0.637			
b09_n	NL	-2.232	-1.413	-0.642			
c02_n	NL	-2.467	-1.739	-1.063	-0.426		
f03_n	NL	-2.388	-1.753	-1.217	-0.722		
f04_n	NL	-2.228	-1.457	-0.786	-0.193		
l01b_	NL	-2.910	-2.333	-1.775	-1.208	-0.605	
l15_n	NL	-2.361	-1.538				
b10_u	UK	-2.066	-1.147				
d04_u	UK	-2.077	-1.164				
g01_u	UK	-2.006	-1.023				

Table 4.5 Conversion key for physical activity based on the EUPASS data set. Use the values to recode item into the common scale under the N(-2,1) prior.

4.5 Differences between Member States

The conversion key can be used to recode the original data into the common scale. The values on the common scale can be used to estimate the amount of physical activity, and compared across different MS.

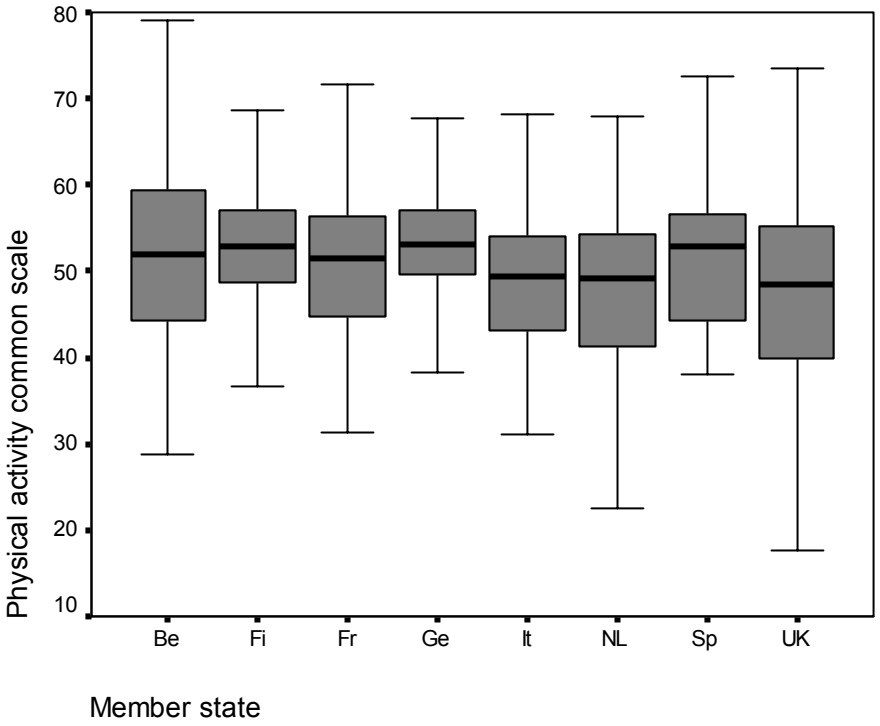

Figure 4.10 Distribution of physical activity estimated from all available items in the EUPASS study per MS. The common scale is standardised with mean 50 and s.d. 10. A higher score means a higher level of physical activity.

Figure 4.10 displays the distribution of physical activity scores on the common scale. RUMM ability estimates θ at the individual level were standardised with mean 50 and s.d. 10 by the linear function -10.478*θ + 53.49, and subsequently aggregated over member states. Note that the direction of the common scale in Figure 4.10 has been reversed, with a higher score indicating a higher ability level. Member States with relatively high physical activity levels include Finland, Germany and Spain. Comparatively low levels are found in Italy, The Netherlands and United Kingdom. In addition, differences in spread occur, e.g. F and Ge versus Be and UK.

4.6 Conclusion

In this chapter we developed a conversion key for measuring physical activity. The key contains both item from a new instrument (IPAQ) as well as 'old' items. Bridge items were all checked for DIF, and suspect items treated as country-specific, i.e., as not comparable. We started with 51 items, of which 10 were common to all countries, and 41 were existing items. Seven items were deleted (4 common items). In the remaining set, all six common items were split to improve comparability.

Though the conversion methodology is clearly a step forward, we should remain careful in drawing conclusions with respect to differences in physical activity between member states. First, response rates during the EUPASS project were not so high, which could have lead to non-random selection and thus to possible unrealistic differences in samples between member states. Second, there could be some overall tendency to respond more optimistic or more pessimistic in some of the member states. The DIF

that we adjusted for when we split items, is DIF that was present in some items (or member states), but not in others. If there is DIF that occurs systematically in all items, e.g. a common response shift, then our method may fail to pick that up, and consequently will not correct for such biases.

Taking these considerations into account, the conversion key for physical activity allows us to make comparisons between member states that could not have been done before. We are now able to compare physical activity between member states, not only in the EUPASS dataset, but also by the traditional indicator in each individual country. Thus we could even convert older datasets based on the member-state-bound indicators onto the common scale, and make longitudinal comparisons between member states.

As a side lesson, we note that it is difficult to develop a questionnaire that is free of DIF. All IPAQ items had some form of DIF. The ability to pick this up is a methodological advance that will also be useful beyond the limits of this project.

5 Personal Care

Mike Horton, Gemma Lawton, Alan Tennant

5.1 Trait to be measured

To define Personal Care (PC) for the purposes of this project, the ICIDH-2 description of Self Care Activities was studied as a guideline for item selection. Of the eight sub categories set out, six were chosen for the definition, these were:

- Activities of washing and drying oneself;
- Activities of caring for body parts;
- Activities related to toileting;
- Dressing activities;
- Activities of eating;
- Activities of drinking.

The sub-categories of 'Activities of looking after one's health' and 'Activities related to menstruation' were excluded. In the ECHI-list, the topic is classified as health status as a limitation of usual activities, code 2.4.4.

5.2 Method

5.2.1 Data sources

The main source of information on PC items being collected in the EC was the HIS-HES database at https://www.iph.fgov.be/hishes/. As well as consulting previous work done in this area, a comprehensive search made of this database and items selected according to the above criteria. The types of PC questions asked in these surveys can be summarised into six sub-categories of questions that reflect the ICIDH-2 Self Care definition: dressing, eating, toileting, washing, cutting toenails, and general/combined. Within each of these categories two main types of questions emerged, these were: question regarding difficulty in performing activity and questions about help needed with the activity.

Once the initial data matrix, containing all of the potentially relevant items, had been formulated, all of the corresponding institutions were contacted to obtain the relevant microdata. Contact addresses were all either found on the HIS-HES database or on the websites of the relevant institutions. A table of all 18 potential data sources that were contacted can be found in Table 5.1.

Code	Country	Name of study	Year	Organisation
A02	Austria	Disabled Persons	1995	Statistik Austria
B01	Belgium	Health Interview Survey 1997	1997	Institut Scientifique de la Santé Publique / ISP
CH01	Switzerland	Swiss Health Survey	1997	Bundesamt für Statistik
D02	Germany	Survey on living conditions, health and environment	1999	Bundesinstitut für Bevolkerungsforschung / BiB
D05	Germany	German National Health Examination and Interview	1998	Robert Koch Institut / RKI

		Survey		
DK01	Denmark	Danish Health and Morbidity Survey	1994	National Institute of Public Health / NIPH
E04	Spain	Impairments Disabilities and Health Status Survey	1999	Instituto Nacional de Estadistica
F01	France	Health and Care Interview Survey	1996	INSEE Inst Nat de la Stat et des Etudes Economiques
F02	France	Handicaps, Disabilities and Dependency Survey	1999	INSEE Inst Nat de la Stat et des Etudes Economiques
FIN03	Finland	Health 2000	2000	KTL Kansanterveyslaitos
I01	Italy	Health Conditions and the Use of Health Services	1999	Istituto Nazionale de Statistica / ISTAT
N01	Norway	Survey on Living Conditions	1998	Statistik Sentralbyra
NL01	Netherlands	Continuous Quality of Life Survey	1998	Centraal Bureau voor de Statistiek / CBS
P01	Portugal	National Health Survey	1995	Instituto Nacional de Saude Dr Ricardo Jorge
POLS	Netherlands	POLS Health and Labour	1998	Centraal Bureau voor de Statistiek / CBS
UK03	UK	Disability Survey	1997	Office for National Statistics / ONS
UKL1	UK	Continuing Care Project	1997	University of Leeds
UK04	UK	General Household Survey - Elderly Follow Up	1994	Office for National Statistics / ONS

Table 5.1 Potential data sources for European data on personal care.

An e-mail was sent to each contact address to try and obtain the relevant data. If no response had been received within 6 weeks of the original e-mail, then another attempt was made by sending another reminder e-mail. If, still, nothing had been heard after a reasonable amount of time, then an attempt was made to contact someone else within the same institution.

Some institutes sent the relevant information, while others responded with information on how to gain access to the relevant data via databanks. The final list of the microdata sources that were available for analysis can be found in Table 5.2.

Code	Country	Year	Ages	Size	Items
A02	Austria	1995	all	60000	1
B01	Belgium	1997	all	10221	4
D05	Germany	1998	18-79	7124	1
DK01	Denmark	1994	16+	4668	1
I01	Italy	1999	all	180000	3
P01	Portugal	1995	all	49718	3
POLS	Netherlands	1998	all	9921	3
UK04	UK	1994	>65	1426	5

Table 5.2 European data sources for which we obtained data on personal care.

For various reasons (institution's lack of response to data requests, no microdata available for certain studies, time constraints to receive datasets, etc.), the original data matrix had to be amended to allow for a lack of available data, removal of unlinked items and for the inclusion of previously unidentified items in the studies. The final list of the items that were included from the information that was received can be found in Table 5.3.

5.2.2 Equivalence Assumptions

Items were grouped in sets that were seen as equivalent. The last column of in Table 5.3 shows the item groups where we assumed equivalence. Figure 5.1 gives a pictorial representation of the linkage matrix. Table 5.4 contains the exact recodes applied to obtain the values for the 13 items to be analysed.

Question Code		UK	P01	I01	POLS	A02	D05	DK01	B01	Item code
bath_it	Italy	I		Y						bath1
bath_uk	UK	Y		I						
dresdif5	NL				Y				I	Dress1
dresdif7	Belgium				I				Y	
dreshelp10	Portugal	I	Y							Dress2
dresseas	UK	Y	I							
dreshelp6	Italy			Y						Dress3
eatdif4	NL				Y	I				Eat1
eatdif5	Austria				I	Y				
eateas	UK	Y								Eat2
eathelp7	Belgium								Y	Eat3
eathelp9	Portugal		Y							Eat4
gen3	Belgium						I	I	Y	Gen4
gen4	Germany						Y	I	I	
gen5	Denmark						I	Y	I	
toidif2	Belgium								Y	toi2
toieas	UK	Y								toi3
waseas	UK	Y								Wash2
washf1	Italy		I	Y	I					Wash4
washf3	Portugal		Y	I	I					
washf5	NL		I	I	Y					

Figure 5.1 Linkage matrix of personal care items. 'Y' indicates which items were observed in which studies. 'I' are considered equivalent.

Note that linking of the studies using equated items leaves the structure vulnerable. The equivalence assumptions used to equate the items are generally tested using differential item functioning (DIF) analysis. This implies that if a linkage item with DIF is found it may mean that the link between the studies might be broken.

Question code	Question	Categories	Country	Item
bath1	Can you bath/shower without help?	without difficulty, with a little difficulty, can only do it with the help of someone	Italy	bath1
bath2	ghs Bath easy	easy, difficult	UK	
Dresdif5	Dressing and undressing	without difficulty, with some difficulty, with great difficulty, only with the help of others, refuses/doesn't know	NL	dress1
Dresdif7	Can you dress and undress yourself on you own?	yes without difficulty, yes with some difficulty, only with someone to help me	Belgium	
Dresseas	ghs Dress easy	easy, difficult	UK	dress2
dreshelp10	Are you able to get dressed and undressed?	alone without difficulty, alone but with some difficulty, only with help	Portugal	
dreshelp6	Can you get dressed without help?	without difficulty, with a little difficulty, can only do it with the help of someone	Italy	dress3
eatdif4	Eating and drinking	without difficulty, with some difficulty, with great difficulty, only with the help of others, refuses/doesn't know	NL	eat1
eatdif5	eating, drinking	possible to do it without the help of others, possible only with the help of others, not possible at all	Austria	
Eateas	ghs Feed easy	easy, difficult	UK	eat2
Eathelp7	Can you, without the help of someone else, feed yourself and cut up food by yourself?	yes without difficulty, yes with some difficulty, only with someone to help me	Belgium	eat3
Eathelp9	Are you able to eat (cut your food and lift and bring drinks to your mouth)?	alone without difficulty, alone but with some difficulty, only with help	Portugal	eat4
gen3	Bathing, showering or dressing yourself	yes limited a lot, yes limited a little, no not limited at all	Belgium	gen4
gen4	Bathing or dressing yourself?	yes limited a lot, yes limited a little, no not limited at all	Germany	
gen5	Bathing or dressing yourself?	yes very much restricted, yes a little restricted, no not restricted at all	Denmark	
toidif2	Can you get to and use the toilet on your own?	yes without difficulty, yes with some difficulty, I can only get to and use the toilet with someone to help me	Belgium	toi2
Toieas	ghs Toilet easy	easy, difficult	UK	toi3
waseas	ghs Wash easy	easy, difficult	UK	wash2
washf1	Can you wash hands and face without help	without difficulty, with a little difficulty, can only do it with the help of someone	Italy	wash4
washf3	Can you wash your hands and face on your own?	Yes without help difficulty, yes without help but with difficulty, only with help, don't know	Portugal	
washf5	Washing one's face and hands	without difficulty, with some difficulty, with great difficulty, only with the help of others, refuses/doesn't know	NL	

Table 5.3 European survey items on personal care.

Item	UK	Portugal	Italy	Netherlands	Austria	Germany	Denmark	Belgium	category
bath1	Easy		without difficulty						1
			with a little difficulty						2
	Difficult		only with help						
dress1				without difficulty				without difficulty	1
				with some difficulty				with some difficulty	2
				with great difficulty					
				only with help				only with help	3
dress2	Easy	alone w/out difficulty							1
		alone, with difficulty							2
	Difficult	Only with help							
dress3			without difficulty						1
			with a little difficulty						2
			only with help						3
eat1				without difficulty	done without help				1
				with some difficulty					2
				with great difficulty					3
				only with help	only with help				4
					not possible				5
eat2	Easy								1
	Difficult								2
eat3								without difficulty	1
								with some difficulty	2
								only with help	3
eat4		alone w/out difficulty							1
		alone, with difficulty							2
		Only with help							3
gen4						no, not limited at all	no, not restrict at all	no, not limited at all	1
						yes, limited a little	yes, little restricted	yes, limited a little	2
						yes, limited a lot	yes, very much restr	yes, limited a lot	3
toi2								without difficulty	1
								with some difficulty	2

				only with help	3
toi3	Easy				1
	Difficult				2
wash2	Easy				1
	Difficult				2
wash4		alone w/out difficulty	without difficulty	without difficulty	1
		alone, with difficulty	with a little difficulty	with some difficulty	2
				with great difficulty	
		Only with help	only with help	only with help	3

Table 5.4 Recoding table for 13 personal care item from different member states.

5.3 Data Analysis

For the purposes of the analysis, all response categories were consistently recoded into the same direction using consecutive integers, with 0 indicating the category with least impairment and a higher value indicating a category with more impairment. Once the recoding of items was completed, the data was read into the RUMM2020 computer programme to assess the estimate of the conversion key and to assess the fit of the Rasch model.

Due to the large sample size, the fit to the Rasch model as measured purely by statistical significance will be poor in general. For this reason, the main fit measure used here is the RUMM item fit residual statistic. The statistic is normally distributed if the model fits. The sign of the statistic matters. A negative residual fit is generally considered as less problematic than a positive fit statistic, and is often taken to indicate redundancy in the scale.

The initial fit of the items was fairly good, although one of the items (eat1) displayed a disordered threshold. This item was rescored from five to three categories, with the three middle categories being collapsed together. In the analysis of the resulting data set, all thresholds were ordered. Following rescoring, the fit of the Items to the scale remained fairly good. Item bath1 was split by country. As can be seen in Table 5.5, all of the residuals statistics are either negative, or under +3, which is seen as indicative of a good item fit to the model where equivalence is the main objective. We had data on a total of 165,717 persons. Of these, 156650 persons were removed as extremes, and thus could not contribute to the overall conversion key. This left 9067 persons in the final analysis. No DIF could be identified in any of the linking items, but we preferred to split the bath item because other UK item wash2 and and toi3 were found to be located much to the left of similar items. Therefore, it was decided to split the bath item across the two countries.

Item	Location	SE	Fit Residual
bath_it	-2.961	0.035	-2.77
bath_uk	-4.919	0.191	1.25
dress1	1.636	0.063	0.56
dress2	-3.519	0.053	1.05
dress3	-0.541	0.036	-0.06
eat1	6.098	0.118	-1.72
eat2	-0.434	0.286	-0.56
eat3	2.989	0.094	0.20
eat4	1.056	0.036	-13.53
gen4	-0.067	0.042	-1.79
toi2	3.207	0.096	-0.35
toi3	-3.073	0.177	2.05
wash2	-1.244	0.21	-1.30
wash4	1.773	0.03	-6.99

Table 5.5 Location estimates and residuals fit statistics of the final solution for the Personal Care key (n=9067).

Item	Thresholds		Recode values		
	0-1	1-2	0	1	2
bath_it	-5.454	-1.468	-3.221	-2.555	-0.903
bath_uk	-4.919		-3.066	-1.850	
dress1	-0.057	3.328	-2.380	-0.484	2.419
dress2	-3.519		-2.932	-1.633	
dress3	-2,378	1.296	-2.786	-1.649	0.861
eat1	2.988	9.209	-2.061	1.480	3.899
eat2	-0.434		-2.428	-0.441	
eat3	2.431	3.547	-2.112	0.768	3.229
eat4	0.009	2.103	-2.386	-0.649	1.903
gen4	-0.951	0.817	-2.565	-1.269	0.972
toi2	2.173	4.241	-2.125	0.772	3.343
toi3	-3.073		-2.870	-1.522	
wash2	-1.244		-2.563	-0.840	
wash4	0.272	3.274	-2.334	-0.336	2.489

Table 5.6 Thresholds estimates and conversion key recode values for personal care (lognormal prior).

Table 5.6 shows the results of the analysis. Recode values were derived under the lognormal prior. Though the number of items is quite small, some ordering in topics appears. Table 5.6 suggests that personal care topics are approximately ordered as follows: bathing, dressing, toileting, washing and eating. Thus bathing becomes already difficult at relatively low levels of disability in personal care, while eating difficulties occur only at higher levels. There are however some inconsistencies with respect to this general pattern. Both toileting items are located at quite different positions on the common scale. One explanation for this is that the question formulations of both items are different. Item toi3 is located much to the left of toi2. The description for toi3 is ‘Toilet easy’ with response categories ‘easy’ and ‘difficult’. Similar instances of this phenomenon occur in item eat2 ‘Feed easy’ and item wash2. Both items are considerably more to left than the other eating or washing items. This suggests that it is actually easier to answer ‘difficult’ for items toi3 and eat2 than the category ‘some difficulty’ on the other eating, washing and toileting items.

5.4 Country comparison

Figure 5.2 is a graphic representation of the differences between countries. As before, the country differences were estimated from the data used to construct the key. The age trend is quite convincing relative to the country differences. Thus, despite the scant linkage structure in the personal care data, the converted data turn out to be reasonably well behaved. Nevertheless, some anomalies appear. First, the (blue) curve for The Netherlands is higher than all the rest until the age of 50, indicating more personal care disabilities for the Dutch. This could be a real phenomenon, but it might also be an artefact. Perhaps, the Dutch items are 'too difficult' across the board. U.K. data were only collected for people over 65 years of age. Note that the U.K. data hardly show any age trend. All in all, it appears that the data on the common scale for Portugal, Germany, Denmark and Belgium are reasonably comparable, while The Netherlands,

Italy and the U.K. exhibit some peculiarities that need further investigation. The analysis suggests that floor and ceiling effects could play a role in the conversion process, especially if the number of items is small.

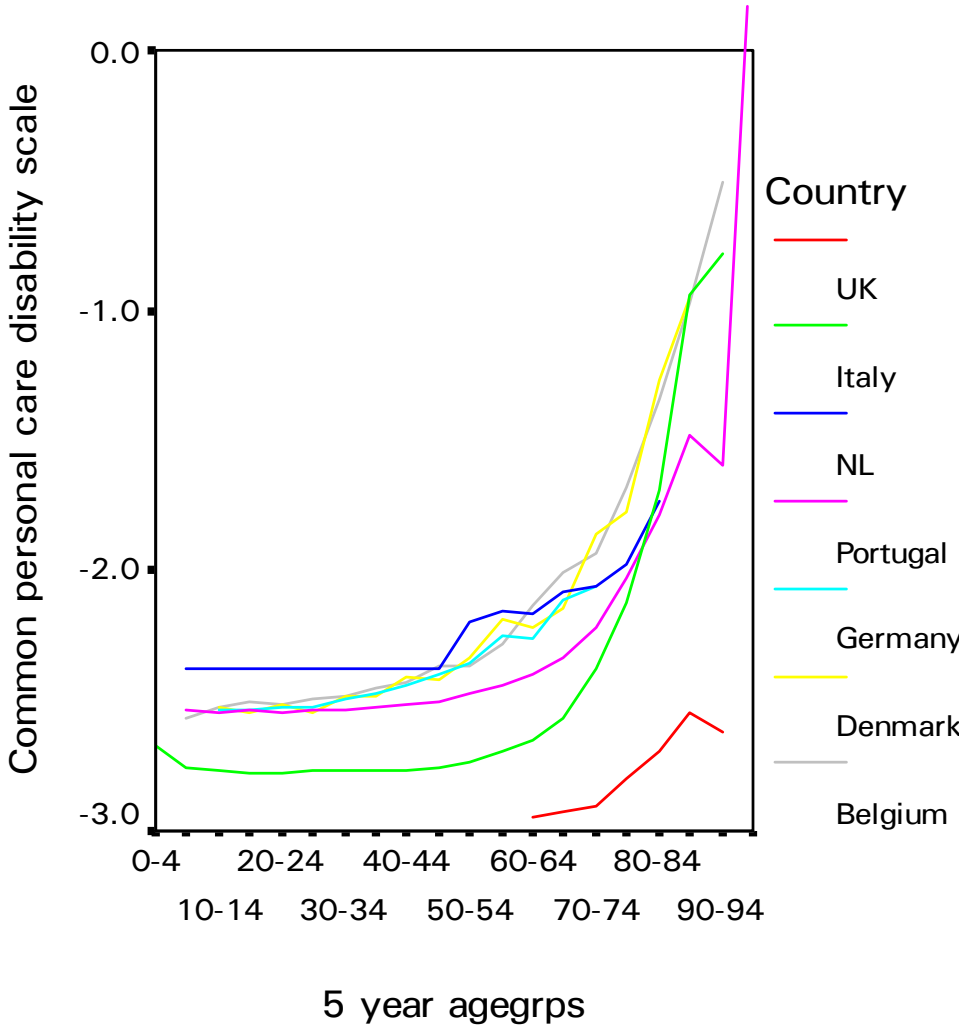

Figure 5.2 Personal care, common disability scale, split by age group and country

6 Sensory function and communication scale

Mike Horton, Alan Tennant

6.1 Trait to be measured

The proposed scale is a scale that which, under the ICF classification, combines the Impairment of Body Function parameters of Seeing, Hearing and Speech functions with the Activity Limitations and Participation Restriction parameters of Communication. This Communication domain is primarily concerned with communicating with (receiving) spoken messages and non-verbal messages, as well as with speaking and producing non-verbal messages. Thus the domain is clearly linked with an individual's ability to see, hear and speak, as well as communicating in general. Thus the proposed scale is a sensory function and communication scale, closely related to functional activity in everyday life. In the ECHI-list, the topic is classified as a health status measurement under code 2.4.3.

6.2 Method

6.2.1 Data sources

A number of sources were used to identify relevant items from questionnaires across Europe (e.g. Robine *et al*, 2000). Other information was obtained from the European Health Interview & Health Examination Surveys HIS-HES Database at https://www.iph.fgov.be/hishes/. A generic internet search was also used to identify independent data collection agencies and databanks that could be utilised. However, this mostly yielded similar information to that identified by searching the HIS-HES database for relevant European questionnaire items. This database was searched for all relevant items pertaining to sensory impairment and communication. More specifically, this included all items that we could find concerning sight/visual impairment, hearing/audio impairment, speaking/aural impairment and also any other general communication impairment items that could be found.

All of the original questions that were found on the database were categorised into 4 separate sections; Sight Items, Hearing Items, Speaking Items and Miscellaneous Communication Items. Following this stage, the initial data matrix was formulated by combining all of the questions in each of the categories. At this initial formulation stage, all items that were seen as equivalent across the studies were grouped together.

Once the initial data matrix, containing all of the potentially relevant items, had been formulated, all of the corresponding institutions were contacted to obtain the relevant microdata. Contact addresses were all either found on the HIS-HES database or on the websites of the relevant institutions. Table 6.1 contains all of the potential data sources that were contacted.

Code	Country	Name of Study	Year	Organisation
AT02	Austria	Disabled persons	1995	Statistik Austria
BE01	Belgium	Health Interview Survey	1997	IPH - Scientific Institute of Public Health
BE02	Belgium	Health Interview Survey	2001	IPH - Scientific Institute of Public Health
CH01	Switzerland	Swiss Health Survey	1997	Bundesamt fur Statistik
CH02	Switzerland	Swiss Health Survey	2002	Bundesamt fur Statistik
DE02	Germany	Survey on Living Conditions, Health and Environment	1998	Bundesinstitut fur Bevolkerungsforschung
DK01	Denmark	Danish Health and Morbidity Survey	1994	National Institute of Public Health
DK02	Denmark	Health and Morbidity in Denmark	2000	National Institute of Public Health
ES01	Spain	National Health Survey	1995	Ministerio de Sanidad y Consumo
ES02	Spain	National Health Survey	2001	Ministerio de Sanidad y Consumo
ES04	Spain	Impairments, Disabilities and Health Status Survey	1999	Instituto Nacional de Estadistica
FR01	France	Health and Care Interview Survey	1996	INSEE - Inst Nat de la Stat et des Etudes
FR02	France	Handicaps, Disabilities and Dependency Survey	1999	INSEE - Inst Nat de la Stat et des Etudes
FR05	France	Handicaps, Disabilities and Dependency Survey	2001	INSEE - Inst Nat de la Stat et des Etudes
FIN03	Finland	Health 2000	2000	KTL
FIN04	Finland	Living Conditions Survey	1986	Tilastokeskus
FIN05	Finland	Finnish Health Care Survey	1996	STAKES
IT01	Italy	Health Conditions and the Use of Health Services	1999	ISTAT - Instituto Nazionale de Statistica
IE01	Ireland	Survey of Lifestyle, Attitudes and Nutrition (SLÁN)	1998	National University of Ireland, Galway
IE03	Ireland	Survey of Lifestyle, Attitudes and Nutrition (SLÁN)	2002	National University of Ireland, Galway
NL01	Netherlands	Continuous Quality of Life Survey	1998	Centraal Bureau voor de Statistiek / CBS
NO01	Norway	Survey on Living Conditions	1998	Statistik Sentralbyra
NO02	Norway	Survey on Living Conditions & Health, Care and Social Relations	2002	Statistik Sentralbyra
PT01	Portugal	National Health Survey	1995	Instituto Nacional de Saude Dr Ricardo Jorge
PT03	Portugal	National Health Interview Survey	1999	Instituto Nacional de Saude Dr Ricardo Jorge
PT02	Portugal	National survey on disabilities, impairments and handicaps	1994	Instituto National de Estatistica
SE01	Sweden	Living Conditions Survey	1999	SCB - Statistika Centralbyran

SE02	Sweden	Living Conditions Survey	2001	SCB - Statistika Centralbyran
UK04	UK	Disability Survey	1997	ONS - Office for National Statistics
UK12	UK	The Health Survey for England	2000	National Centre for Social Research
UK06	UK	The Health Survey for England	2001	National Centre for Social Research

Table 6.1 Potential data sources for European data on communication disability and sensory function.

An e-mail was sent to each contact address to try and obtain the relevant data. If no response had been received within 6 weeks of the original e-mail, then another attempt was made by sending another reminder e-mail and a full translation (if possible) into the native language of the recipient. The final list of the 10 microdata sources that were available for analysis can be found in Table 6.2.

Code	Year(s)	Country	Ages	Sample size	Items
BE01	1997				
BE02	2001	Belgium	all	23556	4
FR02	1999	France	all	16945	5
FIN04	1986	Finland	15+	12057	2
IT01	1999	Italy	all	140011	5
NL01	01/02	Netherlands	all	19421	4
NO01	1998				
NO02	2002	Norway	16+	13952	4
PT03	1999	Portugal	all	48606	3
UK06	2001	UK	all	19640	10

Table 6.2 European data sources for which we obtained data on communication disability and sensory function.

For various reasons (Institution's lack of response to data requests, no microdata available for certain studies, time constraints to receive datasets, etc.), the planned data matrix was amended to allow for a lack of available data, removal of unlinked items and for the inclusion of previously unidentified items in the studies. The final list of the 37 items that were included from the information that was received can be found in Table 6.3.

Survey variable	Country	Question	Categories	Item block
il 24_1	Belgium	Can you see well enough to recognise a friend at a distance of four metres (across a road)?	-6 to -1 = invalid response, 1 = yes, 2 = no	sight1
disaba4	UK	Cannot see well enough to recognise a friend across a road (four yards away)	-9 to -1 = invalid response, 0 = no, 1 = yes	sight1
see1	Italy	Does he/she see enough to recognise a friend 4 metres away (on the other side of the street), using eye-glasses or contact lenses if necessary ?	1 = yes, 2 = no	sight1
oecd5	NL	Can you recognise a face at a distance of 4 meters? (if necessary, wearing glasses or lenses)	1 = Yes, no difficulty /2 = Yes, some difficulty / 3 = Yes, much difficulty /4 = No, I can't	sight1
seeing	Portugal	Can you see well enough to recognise a friend (with or without spectacles or contact lenses) at a distance of 4 metres (across a road)?	1 = yes, 2 or 3 = no	sight1
bsen2	France	Can you (he/she) recognise the face of someone 4 meters away? (with your glasses or lenses on if you wear any)	1 = Yes, without any difficulty/2 = Yes, but with some difficulty / 3 =Yes, but with much difficulty/4 = No/9 = Does not know	sight1
il 25_1	Belgium	Can you see well enough to recognise a friend at a distance of one metre (at arms length)?	-6 to -1 = invalid response, 1 = yes, 2 = no	sight2
see2	Italy	Does he/she see enough to recognise a friend 1 metre away (an arm's length away) ?	8 = yes, 9 = no	sight2
no armsee	UK	Can you see well enough to recognise a friend one yard away (at arm's length)?	-9 to -1 = invalid response, 1 = yes, 2 = no	sight2
bsen1	France	Can you (he/she) see well close to? (to read a paper, a book, draw, do crosswords, with your glasses or lenses on, if you wear any)	1 = Yes, without any difficulty/2 = Yes, but with some difficulty / 3 =Yes, but with much difficulty/4 = No/9 = Does not know	sight3
H20	Norway	Can you without difficulty, or wearing glasses if need be, see ordinary sized newsprint?	1 = yes, 2 = no	sight3
reading	Finland	Can you read ordinary text in a newspaper without difficulty (with or without glasses)?	1 = yes, 2 = no, 9 = no data	sight3
oecd4	NL	Is your sight good enough to read ordinary newspaper print? (if necessary, wearing glasses or lenses)	1 = Yes, no difficulty /2 = Yes, some difficulty /3 = Yes, much difficulty /4 = No, I can't	sight3
comvis	UK	Are your communication problems to do with your vision?	-9 to -1 = invalid response, 1 = yes, 2 = no	sight4
il22 _1	Belgium	Is your hearing good enough to follow a TV programme at a volume others find acceptable ?	-6 to -1 = invalid response, 1 = yes, 2 = no	hearing1

Survey variable	Country	Question	Categories	Item block
hear1	Italy	Does he/she hear enough to be able to watch a television program at a volume that does not disturb other people, using a hearing aid if necessary	1 = yes, 2 = no	hearing1
H21b	Norway	Is your hearing good enough, wearing a hearing aid if need be, to follow along with a TV programme with the sound at a level others find acceptable	1 = yes, 2 = no	hearing1
disaba 3	UK	Cannot follow a TV programme at a volume others find acceptable	-9 to -1 = invalid response, 0 = no, 1 = yes	hearing1
listenin	Portugal	Are you able to hear a tv or radio programme (with or without hearing aid) at a volume which does not disturb others?	1 = yes, 2 or 3 = no	hearing1
il23_1	Belgium	Can you follow a TV programme with the volume turned up ?	-6 to -1 = invalid response, 1 = yes, 2 = no	hearing2
hear2	Italy	If NO. Does he/she manage to hear a television program by raising the volume ?	8 = yes, 9 = no	hearing2
novol	UK	Can you follow a TV programme with the volume turned up?	-9 to -1 = invalid response, 1 = yes, 2 = no	hearing2
bsen3	France	Can you (he/she) hear what is being said in a conversation (if necessary with the assistance of your hearing aid)?	1 = Yes, with no difficulty, even if there are several people around /2 = Yes, if there is only one person speaking, even normally /3 = Yes, if there is only one person speaking aloud /4 = No/8 = Will not answer/ 9 = Does not know	hearing3
hearing	Finland	Can you hear without difficulty what is said in a conversation between several persons (with a hearing aid, if you use one)?	1 = yes, 2 = no, 9 = no data	hearing3
H21a	Norway	Can you without difficulty, or wearing a hearing aid if need be, hear what is said during the course of a normal conversation with at least two other people	1 = yes, 2 = no	hearing3
oecd1	NL	Are you able to hear what is said in a normal conversation between 3 persons or more? (if necessary, wearing a hearing aid)	1 = Yes, no difficulty /2 = Yes, some difficulty /3 = Yes, much difficulty /4 = No, I can't	hearing3
oecd2	NL	Are you able to have a conversation with one person? (if necessary, wearing a hearing aid)	1 = Yes, no difficulty /2 = Yes, some difficulty /3 = Yes, much difficulty /4 = No, I can't	hearing4
comhear	UK	Are your communication problems to do with your hearing?	-9 to -1 = invalid response, 1 = yes, 2 = no	hearing5
speak	Italy	Can he/she speak without difficulty ?	1 =YES, without difficulty / 2 = YES, with a little difficulty / 3 = YES, with great difficulty / 4 = NO, he/she is not able to	speak1
disaba5	UK	Cannot speak without difficulty	-9 to -1 = invalid response, 0 = no, 1 = yes	speak1
speaking	Portugal	Do you have difficulty in speaking?	1 = yes, 2 = no	speak1
bsen4	France	Do you have trouble speaking? (including stuttering)	0 = irrelevant ; dumb, 1 = not at all, 2 = yes, except with people who know me well, 3 = yes, much difficulty, 4 = does not speak ; autistic, 7 =	speak1

Survey variable	Country	Question	Categories	Item block
			irrelevant ; too young, 8 = will not answer, 9 = does not know.	
comspch	UK	Are your communication problems to do with your speech?	-9 to -1 = invalid response, 1 = yes, 2 = no	speak2
bcoh1	France	Notwithstanding problems linked to deafness, can you (he/she) communicate with relatives without any assistance?	0 = Irrelevant: does not communicate with people (autistic) / 1 = Yes, I communicate without any assistance and without any difficulty / 2 =Yes, I communicate without any assistance, but with some difficulty / 3 =Yes, I communicate without any assistance, but with ,much difficulty / 4 = No, I need some assistance / 7 = Irrelevant: too young / 8 = Will not answer / 9 = Does not know	commun1
comfam	UK	Do you have any problems communicating with close members of your family, that is, problems with understanding members of your close family or making them understand you?	-9 to -1 = invalid response, 1 = yes, 2 = no	commun1
disabb08	UK	Have problem communicating with other people - that is, have problem understanding them or being understood by them	-9 to -1 = invalid response, 0 = no, 1 = yes	commun2
H29d	Norway	Owing to permanent health problems or disabilities, have you: .. Had trouble establishing contact with or talking to other people?	1 = not possible, 2 = extremely difficult, 3 = somewhat difficult, 4 = not difficult	commun2

Table 6.3 European survey items on communication disability and sensory function.

Survey variable	Country	Belgium	France	Finland	Italy	Norway	NL	Portugal	UK	Item Block
il 24_1	Belgium	Y	I		I		I	I	I	sight1
bsen2	France	I	Y		I		I	I	I	
see1	Italy	I	I		Y		I	I	I	
oecd5	NL	I	I		I		Y	I	I	
Seeing	Portugal	I	I		I		I	Y	I	
disaba4	UK	I	I		I		I	I	Y	
il 25_1	Belgium	Y			I				I	sight2
See2	Italy	I			Y				I	
no armsee	UK	I			I				Y	
bsen1	France		Y	I		I	I			sight3
Reading	Finland		I	Y		I	I			
H20	Norway		I	I		Y	I			
oecd4	NL		I	I		I	Y			
Comvis	UK								Y	sight4
il22 _1	Belgium	Y			I	I		I	I	hearing1
hear1	Italy	I			Y	I		I	I	
H21b	Norway	I			I	Y		I	I	
Listenin	Portugal	I			I	I		Y	I	
disaba 3	UK	I			I	I		I	Y	
il23_1	Belgium	Y			I				I	hearing2
hear2	Italy	I			Y				I	
Novol	UK	I			I				Y	
bsen3	France		Y	I		I	I			hearing3
Hearing	Finland		I	Y		I	I			
H21a	Norway		I	I		Y	I			
oecd1	NL		I	I		I	Y			
oecd2	NL						Y			hearing4
Comhear	UK								Y	hearing5
bsen4	France		Y		I		I		I	speak1
Speak	Italy		I		Y		I		I	
Speaking	Portugal		I		I		Y		I	
disaba5	UK		I		I		I		Y	
Comspch	UK								Y	speak2
bcoh1	France		Y						I	commun1
Comfam	UK		I						Y	
H29d	Norway					Y			I	commun2
disabb08	UK					I			Y	

Figure 6.1 Linkage matrix of 37 items for measuring communication disability and sensory functioning. 'Y' indicates which items were observed in which studies, 'I' indicates items that were assumed to be equivalent.

The linkage matrix based on the available data can be seen in Figure 6.1.

6.2.2 *Preliminary data transformations*

Due to the nature of the items in the scale, the wording of many items was generic across studies, and these items were coded as equivalent. See Table 6.4.

Block	Countries	Equivalent Items
sight1	Be, Fr, It, NL, P, UK	il 24_1 = bsen2 = see1 = oecd5 = seeing = disaba4
sight2	Be, It, UK	il 25_1 = see2 = no armsee
sight3	Fr, Fin, No, NL	Bsen1 = reading = H20 = oecd4
sight4	UK	n/a
hearing1	Be, It, No, P, UK	il 22_1 = hear1 = H21b = listenin = disaba 3
hearing2	Be, It, UK	il 23_1 = hear2 = novol
hearing3	Fr, Fin, No, NL	Bsen3 = hearing = H21a = oecd1
hearing4	NL	n/a
hearing5	UK	n/a
speak1	Fr, It, P, UK	Bsen4 = speak = speaking = disaba5
speak2	UK	n/a
commun1	Fr, UK	Bcoh1 = comfam
commun2	No, UK	H29d = disabb08

Table 6.4 Preliminary equivalence assumptions of items for measuring communication disability and sensory functioning.

There were, however, some differences in the direction of the questioning across studies. For example, Item il 24_1 from Belgium asks, “Can you see well enough to recognise a friend at a distance of four metres (across a road)?”, in which case an answer of “Yes” would indicate no impairment. In contrast, Item disaba4 from the UK states “Cannot see well enough to recognise a friend across a road (four yards away)”, in which case a response of “Yes” would indicate that the respondent does have some impairment. For the purposes of the analysis, all response categories were consistently recoded into the same direction, with zero indicating the category with least impairment and a higher value indicating a category with more impairment.

A complication occurred in conditional responses. Two items of the UK dataset were originally taken up within the sight1 block as they asked the same question. However, one of the items was asked conditionally on the respondent NOT wearing spectacles or contact lenses. This item was, therefore, not utilised and only the unconditional version of the question was used in the analysis. Similarly, there were two equivalent items within the hearing1 block, but one was conditional on the respondent NOT wearing a hearing aid. Again, only the unconditional version of the question was used.

The construction of the questionnaires from which the items had been taken has to be taken into account. In some questionnaires the level of impairment is assessed by a number of items, where an ‘easier’ item precedes the more 'difficult' item, and where the respondent need not answer the difficult item depending on the answer on the easy item. This results in the ‘harder’ questions, which measure a higher level of impairment, being coded as ‘Not Applicable’, as the level of impairment of the respondent had already been equated by the ‘easier’ item. In these instances, if the respondent had scored 0 on the ‘easier’ item (thus indicating no impairment) and had then been coded as ‘Not Applicable’ on the ‘harder’ item, then these respondents were recoded to score 0 (no impairment) on both the ‘easier’ item and the ‘harder’ item.

Another issue concerned the scoring categories of some data. The manner in which the items in certain datasets were asked meant that the item responses encompassed more than one of the original item categories. These responses therefore were recoded to equate to the other items. Two of the items from the Portuguese dataset underwent this transformation and were recoded as follows:

The item; “Are you able to hear a TV or radio programme (with or without a hearing aid)?” with the following response categories;

1: Yes, at a volume which does not disturb others
2: Yes, but only at high volume
3: No, not even at high volume
N/A: Don't know

was recoded as

(listenin): “Are you able to hear a TV or radio programme (with or without hearing aid) at a volume which does not disturb others?”

1: Yes
2 or 3: No

The item “Can you see well enough to recognise a friend (with or without spectacles or contact lenses)?” with the following response categories;

1: Yes, at a distance of 4 metres (e.g. across the street)
2: Yes, at a distance of one metre
3: No, not even at one metre
N/A: Don't know

was recoded as

(seeing): “Can you see well enough to recognise a friend (with or without spectacles or contact lenses) at a distance of 4 metres (across a road)?”

1: Yes
2 or 3: No

Also, an item from the French dataset was recoded. The item; “Can you (he/she) hear what is being said in a conversation (if necessary with the assistance of your hearing aid)?” with the following response categories;

1: Yes, with no difficulty, even if there are several people around
2: Yes, if there is only one person speaking, even normally
3: Yes, if there is only one person speaking aloud
4: No

was recoded as

(bsen3): “Can you hear without difficulty what is said in a conversation between several persons (with a hearing aid, if you use one)?”

1: Yes
2, 3 & 4: No

Other items were also recoded depending on their answer categories. Table 6.5 contains the full list of recodes applied before data analysis. As a result of these transformations, the structure of the resulting data matrix simplifies. Figure 6.2 is a condensed version of the earlier linkage matrix.

Country	sight1	sight2	sight3	sight4	hearing1	hearing2	hearing3	hearing4	hearing5	speak1	speak2	commun1	commun2
Belgium	■	■			■	■							
Finland			■				■						
France	■		■				■			■		■	
Italy	■	■			■	■				■			
Netherlands	■		■				■	■					
Norway			■		■		■						■
Portugal	■				■					■			
UK	■	■		■	■	■			■	■	■	■	■

Figure 6.2 Condensed linkage matrix as a result of preliminary data transformations (note rows and columns are interchanged compared to Figure 6.1). The green colour indicates where data are available.

Item block	Belgium	France	Finland	Italy	Norway	NL	Portugal	UK	polytomous recode of response categories
sight1	1, yes	1, yes, no difficulty		1, yes		1, yes, no difficulty	1, yes	0, no	0
		2, yes, some difficulty				2, yes, some difficulty			1
		3, yes, much difficulty				3, yes, much difficulty			2
	2, no	4, no		2, no		4, no	2 & 3, no	1, yes	3
sight2	1, yes			8, yes				1, yes	0
	2, no			9, no				2, no	1
sight3		1, yes, no difficulty	1, yes		1, yes	1, yes, no difficulty			0
		2, yes, some difficulty				2, yes, some difficulty			1
		3, yes, much difficulty				3, yes, much difficulty			2
		4, no	2, no		2, no	4, no			3
sight4								2, no	0
								1, yes	1
hearing1	1, yes			1, yes	1, yes		1, yes	0, no	0
	2, no			2, no	2, no		2 & 3, no	1, yes	1
hearing2	1, yes			8, yes				1, yes	0
	2, no			9, no				2, no	1
hearing3		1, yes, no difficulty, even with several people	1, yes		1, yes	1, yes, no difficulty			0
						2, yes, some difficulty			1
						3, yes, much difficulty			2
		2, yes, one person speaking normally							
		3, yes, one person speaking aloud							
		4, no	2, no		2, no	4, no			3
hearing4						1, yes, no difficulty			0
						2, yes, some difficulty			1
						3, yes, much dif			2

						4, no			3
hearing5								2, no	0
								1, yes	1
speak1		1, no		1, yes, without difficulty			2, no	0, no	0
		2, yes, except with people I know well		2, yes, with a little difficulty					1
		3, yes, much difficulty		3, yes, great difficulty			1, yes	1, yes	2
		4, does not speak; autistic		4, no					3
speak2								2, no	0
								1, yes	1
commun1		1, yes, no assistance and no difficulty						2, no	0
		2, yes, no assistance but some difficulty							1
		3, yes, no assistance but much difficulty							2
		4, no, need assistance						1, yes	3
commun2					4, not difficult			0, no	0
					3, somewhat difficult				1
					2, extremely difficult			1, yes	2
					1, not possible				3

0 -> 3 = less to more impaired (ie. No limitation -> More limitation)

Table 6.5 Recoding table for communication and sensory functioning items.

6.3 Data Analysis

The data were fitted to the Partial Credit Model using the RUMM2020 software. At this stage, we decided to remove item blocks sight4, hearing5 and speak2 from the analysis as these items did not follow the same line of questioning as the other impairment items. Rather these items focussed on the main factor of a pre-determined impairment that the respondent already possessed. The main fit statistic used is the item fit residual. Preferably, it should be below a value of +3.0 for all items. Another fit diagnostic indicating fitting problem is the occurrence of reversed thresholds.

The initial fit of the items was poor, with some of the items displaying disordered thresholds. The items were rescored such that all thresholds appeared ordered in the analysis. The rescored categories can be found in Table 6.6.

Item block	0	1	2	3
sight1	0	0	0	1
sight2	0	1		
sight3	0	0	0	1
hearing1	0	1		
hearing2	0	1		
hearing3	0	1	1	1
hearing4	0	0	0	1
speak1	0	1	1	2
commun1	0	0	0	1
commun2	0	0	1	2

Table 6.6 Scheme to collapse categories to increase model fit.

Following rescoring, there were still high positive residuals on a number of items. Misfitting items tested for differential item functioning (DIF) by country, and split if DIF turned out to be problem. This resulted in 3 item blocks being split for DIF by country: sight1, hearing1 and hearing3. Following this item splitting, the items that were still misfitting were systematically removed one by one until a good fit to the model had been established. Of the original 294188 persons entered into the analysis, 277465 persons were removed as extreme, which left 16723 persons in the final analysis. The linkage matrix of the remaining items can be found in Figure 6.3.

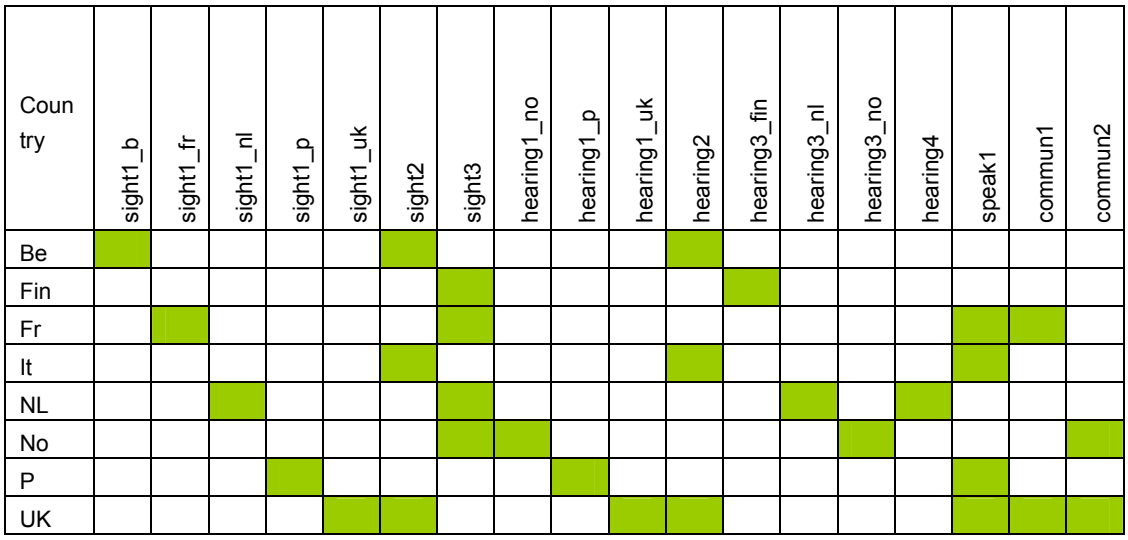

Figure 6.3 Linkage matrix corresponding to the final analysis.

The threshold estimates and the conversion key are given in Table 6.7.

Item	Thresholds		Recode values		
	0-1	1-2	0	1	2
sight1_b	-2.715		-2.815	-1.416	
sight1_fr	-0.076		-2.373	-0.252	
sight1_nl	-0.498		-2.438	-0.474	
sight1_p	0.298		-2.320	-0.046	
sight1_uk	2.188		-2.116	1.042	
sight2	0.277		-2.322	-0.058	
sight3	-0.817		-2.490	-0.636	
hearing1_no	2.194		-2.115	1.045	
hearing1_p	-0.404		-2.423	0.426	
hearing1_uk	1.390		-2.188	0.580	
hearing2	-0.195		-2.391	-0.315	
hearing3_fin	-2.027		-2.699	-1.173	
hearing3_nl	-3.961		-2.985	-1.722	
hearing3_no	0.764		-2.259	0.217	
hearing4	0.855		-2.248	0.270	
speak1	-0.585	-0.319	-2.566	-1.399	0.531
commun1	1.440		-2.183	0.609	
commun2	1.379	2.096	-2.223	-0.058	2.356

Table 6.7 Result: Threshold estimates and conversion key, derives using a lognormal prior.

6.4 Country comparison

The available data were recoded according to Table 6.7, and summary statistics by country and age group were calculated. Some of the data sources provided survey weights that correct for unit nonresponse according to known population totals, whereas

others did not. Different methods were used to derive these survey weights. In order to eliminate any differences due to different methods for estimation weights, we decided not to apply any weighting at all in the summary statistics.

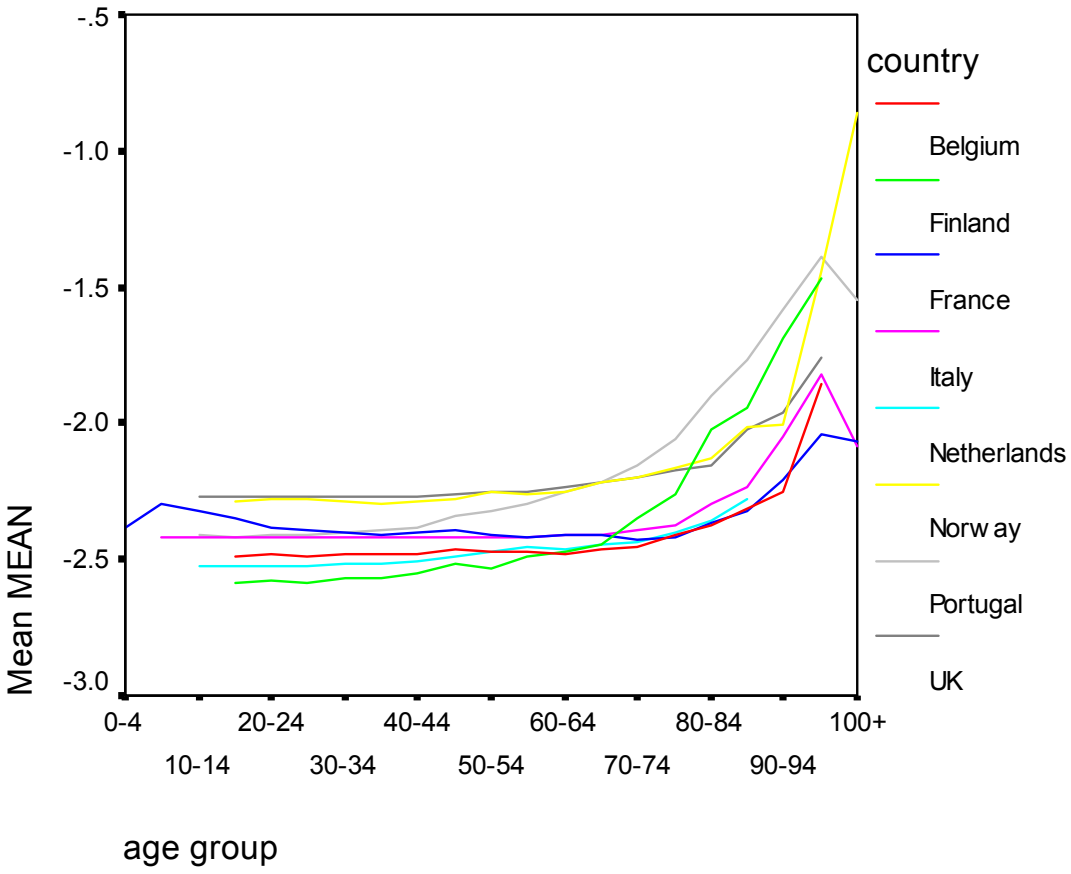

Figure 6.4 Mean communication disability and sensory function as a function of age for different European countries

Figure 6.4 is a representation of the mean communication disability and sensory functioning for five-year age groups by different European countries. The mean was calculated over all available items per country. So, for example, the position of Belgium is defined by the average of its three items. Higher values mean more disability. Note that, as expected, disability prevalence increases with age in all countries. The pattern for France (the blue line) is peculiar at the youngest age groups. Closer examination of the raw data reveals that in France the incidence of speaking problems (item speak1) is much higher in young children than in adults.

Substantial differences occur between the levels between age 20-60, and between the slope of the lines. The U.K. and Norway exhibit relatively high levels, whereas Finland and The Netherlands are low. It is instructive to examine the raw data on item sight3, which is shared by Finland, Netherland, Norway and France. It appears that, in the age group 20-24 y, only 1 out of 709 persons (0.15%) in the Dutch sample reports a problem, whereas 19 out of 1015 persons (1.87%) of the Norwegian sample report problems. So already in the raw data, the Norwegian incidence is about 12 times as large as the Dutch incidence. As the sight3 item is one of the linkage items, it is hardly

surprising to find that the Norwegian curve is located about the Dutch curve. One could conclude that, in hindsight, the equivalence assumption on sight3 should perhaps not have been made. On the other hand, assumptions of this type are often made implicitly, and hardly ever critically evaluated.

6.5 Conclusion

Substantial effort went into data collection, but despite these efforts, the linkages between countries are relatively thin. Various recode steps had to be applied. Application of the conversion keys identified some peculiarities in the raw data. For example, we found a Norwegian incidence that was 12 times as high as the Dutch incidence. Such peculiarities show up thanks to the systematic application of Response Conversion.

7 Quality of life: Physical well-being

Gert Jacobusse, Stef van Buuren, Jeanet Bruil, Ulrike Ravens-Sieberer for the KIDSCREEN group

7.1 Trait to be measured

Health-related quality of life (HRQoL) is an often-used measure for determining the effect of health interventions. “General quality of life” is listed under code 2.3.7 in the ECHI-2 indicator list, but has not yet been appropriately operationalised. Health-related quality of life is a multidimensional concept. Depending on the definition, it covers topic to physical, social, cognitive, emotional well-being. A common denominator in quality of life measurement is that it takes the viewpoint of the patient explicitly into account.

This chapter focuses on the physical domain of HRQoL. Physical well-being is an important aspect of quality of life, and involves issues like general health, physical activity, energy and fitness.

7.2 Method

7.2.1 Data

The conversion key will be based on data from the European KIDSCREEN project (Ravens-Sieberer *et al*., 2001; Rajmil *et al*, 2004). KIDSCREEN is a carefully designed study using state-of-the-art harmonisation methodology. The aim of the KIDSCREEN project is to develop a screening instrument for quality of life in children between 8-18 years. To achieve this aim, items were sampled across different countries using focus group interviews. The instrument was tested in eleven European countries: Austria, Germany, France, Netherlands, Spain, Switzerland, Czech Republic, Poland, Greece, Ireland and the United Kingdom. See www.kidscreen.de for more details.

Ten quality of life scales were constructed within the KIDSCREEN project: physical well-being, psychological well-being, moods and emotions, self perception, autonomy, parents relations and home life, peers and social support, school environment, bullying, financial resources. This chapter only concerns physical well being, a scale consisting of five items. In addition to these items, we studied 20 extra items, related to physical well being, and also collected in KIDSCREEN. Some items were developed by the KIDSCREEN team (pwb_1 to pwb_5, itm_1 to itm5), others were taken from existing instruments in the participating countries. Table 7.1 contains the description of the items analysed in this chapter.

itmcode	Parent/ Child	Question	Response Categories
pwb_1	C	In general, how would you say your health is?	Excellent, very good, good, fair, poor
pwb_2	C	Have you felt fit and well?	Excellent, very good, good, fair, poor
pwb_3	C	Have you been physically active (e g running, climbing, biking)?	Excellent, very good, good, fair, poor
pwb_4	C	Have you been able to run well?	Excellent, very good, good, fair, poor
pwb_5	C	Have you felt full of energy?	Excellent, very good, good, fair, poor
itm1	C	I am full of energy	Completely agree, mostly agree, agree a little, do not agree
itm2	C	I resist illness very well	Completely agree, mostly agree, agree a little, do not agree
itm3	C	When I get sick, I usually recover quickly	Completely agree, mostly agree, agree a little, do not agree
itm4	C	I am very physically fit	Completely agree, mostly agree, agree a little, do not agree
itm5	C	Felt ill	Never, seldom, sometimes, often, all the time
itm6	C	Felt strong and full of energy	All the time, often, sometimes, seldom, never
itm7	C	Have you had little energy?	Never, seldom, sometimes, often, all the time
itm8	C	Have you been in good physical shape?	Always, often, sometimes, rarely, never
itm9	C	I have low energy	Never, almost never, sometimes, often, almost always
itm10	C	I get a lot of headaches, stomach-aches or sickness.	Not true, somewhat true, certainly true
itm11	P	In general, how would you say your child rates his/her health?	Excellent, very good, good, fair, poor
itm12	P	Has your child felt fit and well?	Extremely, very, moderately, slightly, not at all
itm13	P	Has your child been physically active (e g running, climbing, biking)?	Extremely, very, moderately, slightly, not at all
itm14	P	Has your child been able to run well?	Extremely, very, moderately, slightly, not at all
itm15	P	Has your child felt full of energy?	Always, very often, quite often, seldom, never
itm16	P	In general, how would you say your childs health is?	Excellent, very good, good, fair, poor
itm17	P	Limited due to health problems: doing things that take a lot of energy	No, yes a little, yes some, yes a lot
itm18	P	Limited due to health problems: doing thins that take some energy	No, yes a little, yes some, yes a lot
itm19	P	My child seems to be less healthy than other children I know	definitely false, mostly false, don't know, mostly true, definitely true
itm20	P	My child has never been seriously ill	definitely false, mostly false, don't know, mostly true, definitely true

Table 7.1 Selected items from the KIDSCREEN study for measuring physical well being in children.

	D	ES	NL	A	UK	FR	CH	GR	CZ	IRL	PL
pwb_1	880	494	943	757	456	502	857	596	801	185	845
pwb_2	882	493	951	758	456	503	862	595	799	186	850
pwb_3	878	493	948	760	456	500	861	595	798	186	848
pwb_4	872	492	948	749	450	485	856	595	795	185	844
pwb_5	883	478	953	757	456	504	866	595	799	186	848
itm1	562	302	646	472	320	177	582	592	525		554
itm2	559	301	646	468	319	176	578	595	525		553
itm3	562	302	645	473	320	176	581	594	525		553
itm4	561	301	645	472	320	176	579	595	525		554
itm5	879	490		759			281	593			
itm6	879	490		759			278	593			
itm7		303				393					
itm8		302				391					
itm9					318					9	
itm10	561	301	645	475	78	176	275	580	524		550
itm11	874	381	930	732	449	467	859	525	798	128	829
itm12	876	381	931	730	449	469	861	524	798	156	826
itm13	877	380	932	730	449	471	863	524	797	156	825
itm14	856	377	922	731	446	449	849	525	789	156	820
itm15	873	378	932	732	448	470	861	523	799	155	829
itm16	870	385	932	739	449	471	865	522	802		830
itm17	871	384	928	733	447	469	862	512	800		823
itm18	868	380	926	732	445	468	857	509	793		816
itm19	866	386	934	731	445	471	859	505	795		806
itm20	861	388	932	731	448	470	860	510	792		797

Table 7.2 Realised sample size per item per country.

A total of 7381 children and parents completed the KIDSCREEN questionnaire. Some of these items were only administered in a subset of countries. Table 7.2 provides the sample per item and per country. The items are generally very well linked to each other. There are many common items across country, and each could potentially act as a bridge item.

7.2.2 Data analysis

All items were recoded so that the lowest value indicates the highest physical well-being. Threshold estimates were estimated under the partial credit model, using RUMM2020 (Rumm Laboratories, 2003).

We found that the standard fit residual statistics of RUMM depend on sample size. As large samples lead to more stringent statistical testing, large samples would lead to exclusion, or split up, of many items. Therefore we also computed ‘outfit mean square’ statistics (Wright and Masters, 1982) as a part of our model fitting strategy, which are almost invariant under sample size. Values between 0.5 and 1.5 are labelled ‘productive

for measurement' (Linacre, 2002), while Schulman, Trujillo and Karney (2001) report a much stricter "ideal range" of 0.9-1.2 for the outfit statistic.

The analysis of DIF is also influence by sample size. At these sizes, conventional statistical tests will indicate significant DIF, even though the absolute sizes of DIF could be small. In order to circumvent these problems, two additional methods for establishing DIF were applied. The first method assesses the ICC graph and infers whether all country specific trace lines occur within a certain band. The width of the band can be expressed in the unit of logits. Rules of the thumb criteria include 0.5 logit (for a relatively stringent test) and 1.0 logit (for a relatively loose assessment). Another method, practiced by the KIDSCREEN group, is to quantify the difference between two polytomous logistic regression models. The first is the response probability as a function of the total score (or a substitute for the total score). The second model adds a country effect and an interaction of country effect and total score. The difference between the models can be expressed as the amount of explained variance by the country and interaction effect. Jodoin and Gierl (2001) suggest that an increase beyond 3.5% in the percentage of explained variance indicates DIF.

7.3 Results

7.3.1 Model fitting

Based on an analysis of all items, we excluded items itm5, itm19 and itm20 from further analysis. Both model fitting statistics were in close agreement about the most deviant items, with Rumm residuals over 8 and an outfit mean square over 1.3.

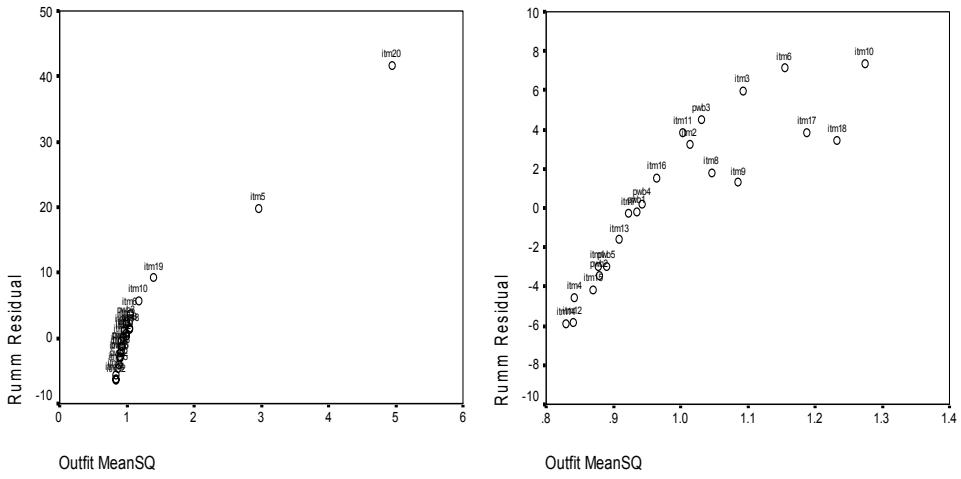

Figure 7.1 Model fit statistics: RUMM residuals plotted versus Outfit Mean Square before (left) and after (right) elimination of items itm5, itm19 and itm20.

7.3.2 Differential item functioning

As expected, nearly all items show statistically significant DIF between member states Exceptions were itm7 and itm9. Figure 7.2 plots the mean score of item pwb1 per country as a function of the total score based on all items. In general, all countries have

the same pattern. The band around the mean curve is about 1.0 logit, which is substantial and indicates DIF. It appears that two countries (Germany and The Netherlands) have relatively high scores, while the mean of the two other countries (France and Greece) is very low. The remaining countries are compact and have a band of well below 0.5 logits. The percentage of variance explained by DIF is 4.1%, thus over 3.5%. The figure and the analysis suggest a split the item into three groups.

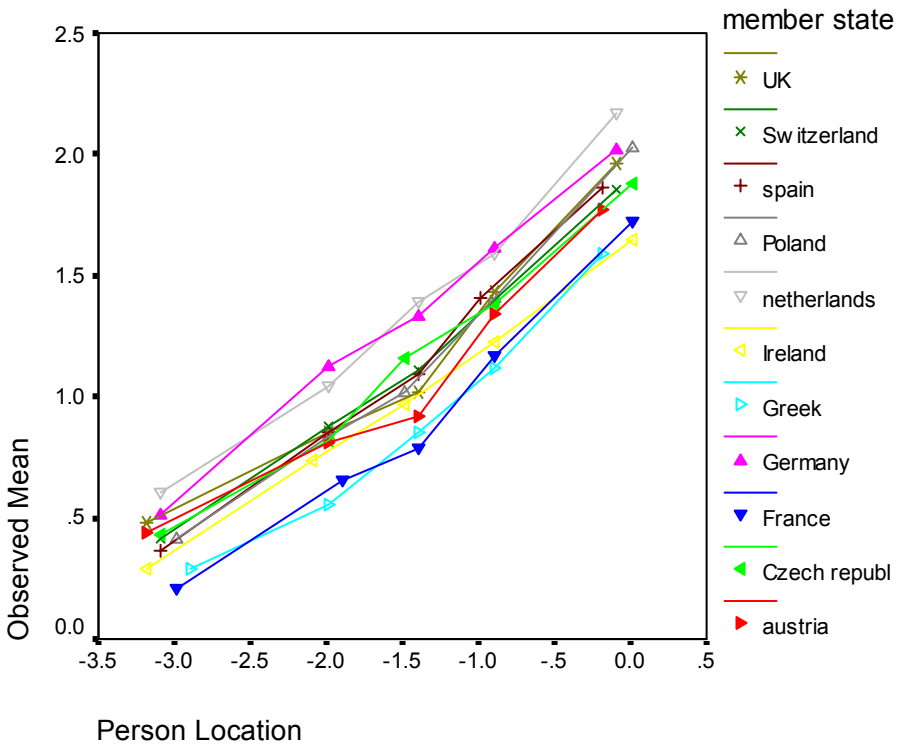

Figure 7.2 Differential Item Functioning of item pwb1 by country. The horizontal axis is the common scale based on all items, the vertical axis is de average observed score of item pwb1.

Figure 7.3 provides another view at the DIF in item pwb_1. The observed probabilities for each of the five response categories are given for France, Greece, Germany and Netherlands. The figure illustrates that that the French and Greek have a higher probabilities to respond 0 (excellent) or 1 (very good), while the German and the Dutch have higher probabilities to respond 2 (good), for all given person locations. Categories 3 (fair) and 4 (poor) are hardly ever chosen at all, though category 3 seems to be slightly more often chosen in The Netherlands and in Germany. These findings could provide input to re-evaluate the translations of the question, or to look for cross cultural differences that could explain the DIF.

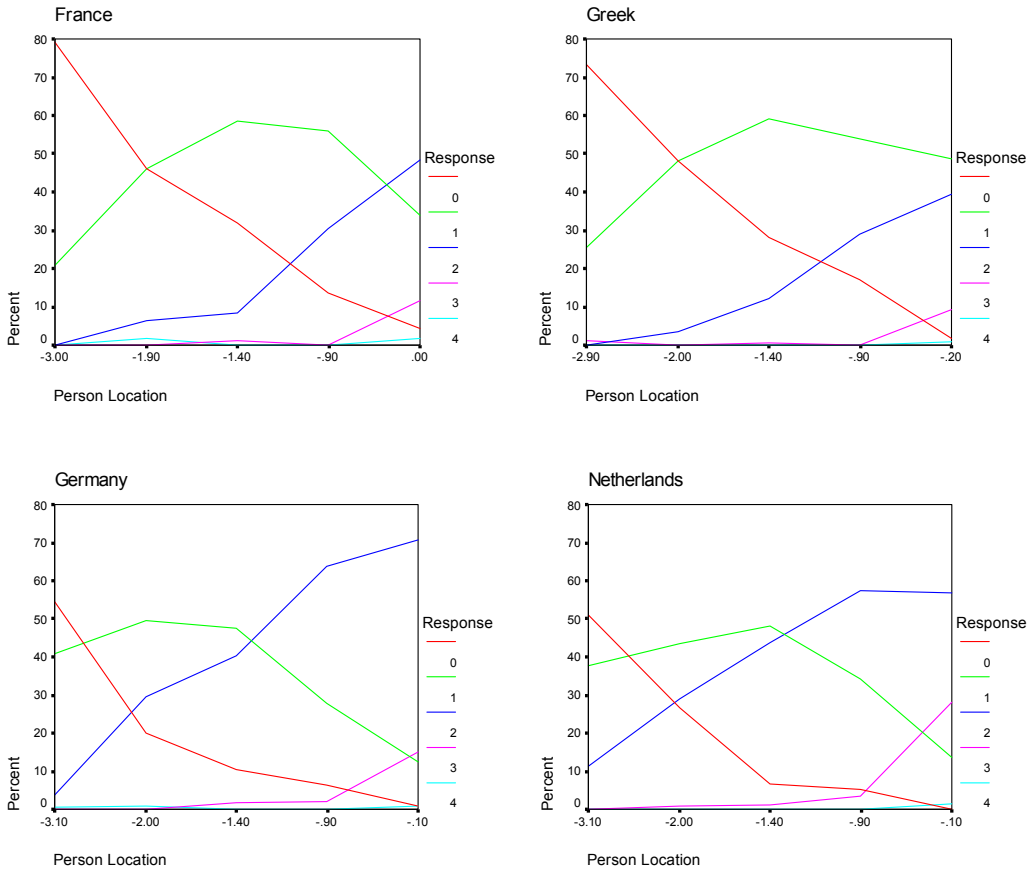

Figure 7.3 Percentage responses per category of item pwb_1 as a function of the common scale (person locations) for the most diverging countries. Note for example difference in the location where the green line (category 1: very good) intersects the red line (category 0: excellent) and the blue line (category 2: good).

Similar DIF analyses were done for all other items. In this way, we identified six items with country DIF (c.f. Table 7.3). These items were split across country groups, thus resulting in a total of 31 items to be used for estimating the model parameters.

Item	group1	group2	group3
pwb_1	D, NL	FR, GR	Rest
itm1	FR	Rest	
Itm4	FR	Rest	
itm11	D, NL, ES	GR, UK	Rest
itm14	FR	Rest	
itm16	D, NL, ES	FR, GR, UK	Rest

Table 7.3 Scheme for splitting 6 items for measuring physical well being.

7.3.3 Conversion key

The conversion key, based on the solution without items itm5, itm19 and itm20, and split according to Table 7.3 is given in Table 7.4. Though the most severe sources of DIF have been removed in the modelling process, one should be aware that DIF could still play a role in the comparisons between member states, especially when the comparison is based on a single item.

Item	Countries	0	1	2	3
pwb1a	D,NL	-3.025	-1.690	1.266	2.505
pwb1b	FR,GR	-2.117	-0.187	1.469	2.200
pwb1c	Rest	-2.826	-0.793	1.442	2.281
pwb2	All	-2.237	-0.588	0.648	1.007
pwb3	All	-1.813	-0.970	-0.276	0.290
pwb4	All	-1.721	-0.605	0.412	0.366
pwb5	All	-2.879	-0.796	0.570	2.086
itm1a	FR	-1.023	0.190	1.458	
itm1b	Rest	-2.143	0.203	1.713	
itm2	All	-1.725	-0.130	0.921	
itm3	All	-1.443	-0.111	0.806	
itm4a	FR	-1.021	0.037	2.680	
itm4b	Rest	-1.738	-0.223	1.154	
itm6	All	-2.732	-0.726	-0.368	0.512
itm7	All	-1.477	-0.291	1.080	1.657
itm8	All	-2.123	-0.107	0.418	0.929
itm9	All	-0.922	-0.236	1.040	2.654
itm10	All	-0.261	0.784		
i11a	D,NL,ES	-3.110	-1.542	1.589	1.601
i11b	GR,UK	-1.433	0.407	1.718	5.216
i11c	Rest	-3.155	-0.930	1.519	2.834
itm12	All	-2.511	-0.248	1.499	1.241
itm13	All	-2.293	-1.115	0.040	0.209
i14a	FR	-3.103	-0.742	-0.094	-0.526
i14b	Rest	-2.032	-0.500	0.876	0.087
itm15	All	-2.742	-0.554	0.993	2.591
i16a	D,NL,ES	-2.495	-0.900	1.427	3.052
i16b	FR,GR,UK	-1.347	0.364	1.635	3.691
i16c	Rest	-2.809	-0.417	1.586	3.089
itm17	All	0.608	0.135	0.089	
itm18	All	1.130	0.391	0.262	

Table 7.4 Conversion key for Physical Well Being items for Quality of Life for children. Based on KIDSCREEN data using a N(0,1) prior.

7.3.4 Country comparison

Figure 7.4 provides a box plot on the common scale for each country.

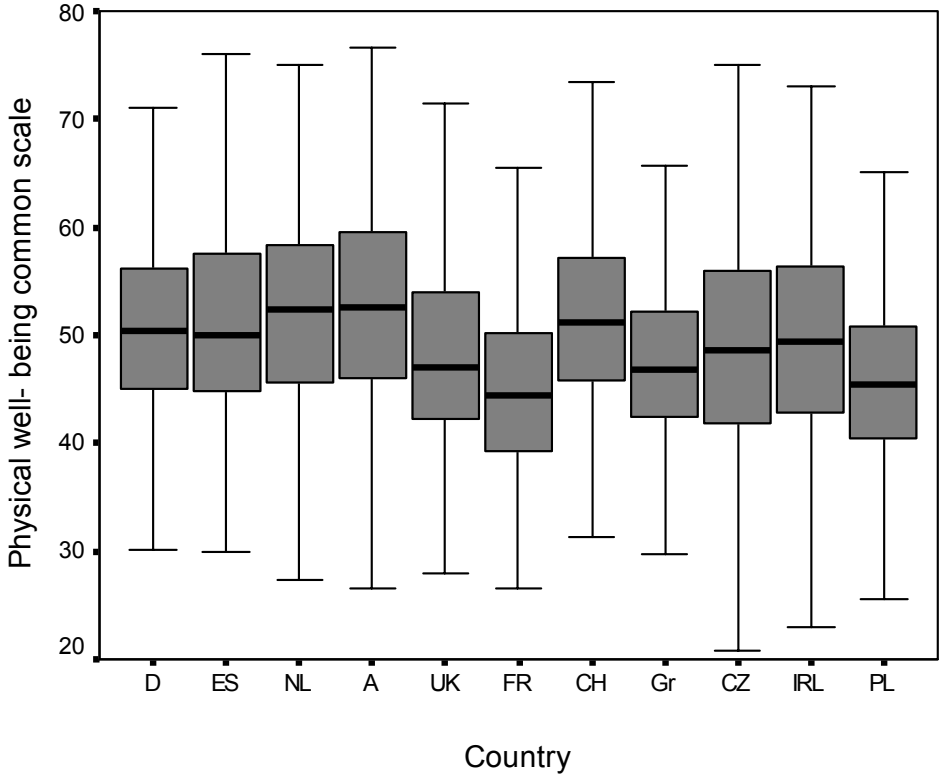

Figure 7.4 Box plots of the distribution of Physical well being in 11 European countries based on the KIDSCREEN data set. The common scale was standardised with overall mean 50 and s.d. 10 by the linear function -8.347*θ +35.90, where θ is the RUMM ability estimate.

Data are scaled such that a high score means a high level of physical well being. The differences in levels between countries are fairly large, with low average values for France and Poland. Note that the 75 percentile of the French distribution, i.e., the top line of the gray box, is even lower than the Swiss mean. Also, differences in spread occur. For example, the spread is low in Greece and high in Czech Republic. Such differences are of interest when a proportion below a certain value, e.g. 30, is needed.

7.4 Conclusion

The KIDSCREEN project is a carefully planned study using state-of-the-art pe-harmonisation methodology for improving comparability. The linkage structure between items for the personal well being scale is very strong. Yet, even after correcting for DIF, we observe substantial differences between countries in personal well-being, both in level and in spread. The differences might reflect true population differences, but it would be comforting if other sources would confirm these findings. More detailed analysis could perhaps reveal systematic sources of variation that could account for part of the country effect. All in all, Response Conversion analysis brings out new insights that would have been difficult to obtain otherwise.

8 Conclusion

Stef van Buuren

Response Conversion is a way to improve comparability in existing data. All assumptions in RC are explicit. The conversion process takes small steps, is fully repeatable, and leads to verifiable quantitative results. Application of the method helps to evade some common pitfalls when dealing with cross-cultural comparability.

8.1 Main results

The project produced the following results:

- A detailed description of the statistical methodology and the accompanying modelling issues (Ch. 2);
- A web site site http://www.tno.nl/responseconversion with some on line tools to support the RC methods (Ch. 2);
- A list of 26 ECHI-indicators for which RC could be potentially be useful (Ch. 3);
- New conversion keys for
    - Physical Activity (Ch. 4)
    - Personal Care Disability (Ch. 5)
    - Communication Disabilities and Sensory Functioning (Ch. 6)
    - Quality of life: Physical well-being for children (Ch. 7).

In addition to these results, numerous new technical insights were obtained throughout the analyses in Chapters 4-7. More in particular, new lessons include:

- floor and ceiling effects may affect comparability (Ch. 6);
- obtaining data from the statistical offices in the different MS is a tedious and long-winded process (Ch. 5 and 6);
- it is preferable to work with data from planned studies that include a cross-cultural dimension. The stronger linkage structure of such studies allows better statistical analyses (Ch. 4 and 7);
- different diagnostic item fit statistics may lead to different conclusions (Ch. 7);
- recoding variables into a small number of categories turned out to be superior to modelling the large number of categories (Ch. 4);
- DIF occurred in items thought to be of as cross-culturally invariant (Ch. 4 and 7);
- appropriate recoding may be used to deal with conditional questions (Ch. 5);
- recoding tables can be used to systematically portray all decision taken by the analyst (Ch. 5 and 6).

8.2 Suggestions for further work

8.2.1 Application issues

The soft point of the RC methodology as currently practiced is that there is no clear-cut and explicit procedure for helping the uninitiated decide when a specific item can act as

a bridge item. Objective procedures need to be developed, such as the absence of differential item functioning. In principle, each bridge item should adhere to minimal standards for comparability. Such standards need to be developed. In particular, it would be useful to have independent quantitative measures of comparability that can be used as a yardstick for the success of a particular application.

It would be useful to compare the estimates on the common scale with raw country differences, and with to data obtained from other sources. For example, for Physical Activity, we found that none of the 10 common items passed through all tests. Four items were deleted because they did not fit the model, the remaining six were splits to account for country DIF. The analysis on the common scale represents a technical improvement over the raw analysis. It would be interesting to investigate whether the new analysis of country differences would lead to different insights.

It is certainly not always clear why certain items display DIF. It would be very helpful if our analyses were complemented by more qualitative approaches looking into translation and interpretation issues, performing in-depth systematic analysis of the item content, and so on. Such qualitative work would especially be needed to complement our evaluation of DIF in bridge items.

The EU enlargement process brings with it new member states with their own different statistical systems, reporting traditions, sample coverage, languages, and question formulations. The currently developed conversion keys need updating to include the new Member States.

Depending on the priorities of the EC, the development of new keys would be useful. The analysis in Chapter 3 identified 26 potential topics where Response Conversion could improve comparability.

8.2.2 Implementation issues

It would be useful if the conversion key values could be coupled to the existing on-line HIS-HES and ECHI-databases. For example, pointers to conversion keys could be added to these meta-databases.

In order to stimulate wider acceptance, some independent body or organisation should evaluate and endorse the developed keys. Preferably, there should be a formal scheme of approval, as well as a kind of mechanism for updating, numbering and extending conversion keys. In addition, it would be useful if appropriate links to the relevant EUROSTAT task forces and structures could be formed.

8.2.3 Technical issues

Our analyses made clear that different diagnostic measures may result in different decisions. More insight is needed into the dependencies among different measures.

In Chapter 2, we proposed that the choice of the prior distribution be guided by the population distribution on the common scale. Some further work is needed to verify whether this choice is optimal, and how alternatives choices affect the final result.

Response Conversion may result in a loss of information as a consequence of the conversion process. In order the account for this, the standard errors of parameters in the common scale estimates may need to be adjusted. There is good Bayesian theory for deriving such estimates, but this theory needs to be adapted to the current context.

In those cases where multiple survey items are used to position a person on the common scale, we either used the RUMM estimates (in Chapters 4 and 7) or the average over recoded values of the individual items (in Chapters 5 and 6). The RUMM estimation method is scientifically appropriate, but the use of the software is not a very logical choice for users that do not construct conversion keys themselves. The mean calculation method can be done by anyone by standard software, but a disadvantage is the method is more sensitive to the choice of the prior than needed. It would therefore be useful to have software that can calculate common scale estimates from multiple items.

8.3 Final comment

Response Conversion is an evolving technique. We submitted a new application for funding as a response to the Call for Proposal 2004 of the EC Public Health programme 2003-2008. This project is titled "Comparability Methods for Community Health Indicators" (COMET), and extends the current project team with a larger and more diverse group of experts on comparability methodology. The prospective project draws together expertise on translation issues, statistics, cross-cultural psychology and public health. The following main outputs are envisaged:

- a report on the evaluation of current practice in dealing with comparability the PH programme;
- improvements and extensions of Response Conversion;
- a set of standards for comparability.

We believe that further developments along these lines will strengthen the information system that is needed to advance European health policy.

A References

[1] ANDRICH D. Rasch models for measurement. Newbury Park: Sage, 1988.

[2] ANDRICH D, LUO G. Conditional Pairwise Estimation in the Rasch model for Ordered Response Categories using Principal Components. Journal of Applied Measurement 2003, 4, 205-221.

[3] ANDRICH D, DE JONG JHAL, SHERIDAN BE. Diagnostic opportunities with the Rasch model for ordered response categories. In J Rost, R Langeheine, Eds, Applications of Latent Trait and Latent Class Models in the Social Sciences (pp. 59-70). New York: Plenum Press, 1997.

[4] BOCK RD, MISLEVY RJ. Adaptive EAP Estimation of Ability in a Microcomputer Environment. Applied Psychological Measurement 1982, 6, 431-444.

[5] CURTIS DD. Misfits: People and their problems. What might it all mean? International Education Journal 2001, 2 (4), 91-100.

[6] ECHI WORKING GROUP. Design for a set of European Community Health Indicators, 2001. pdf

[7] ECHI WORKING GROUP. Proposed draft list of EC Health Indicators (long list), 2004a. pdf.

[8] ECHI WORKING GROUP. Draft list of recommended 'First Phase Core Indicators' (short list), 2004b. pdf

[9] EMBRETSON SE, REISE SP. Item Response Theory for Psychologists. London: Lawrence Erlbaum, 2000.

[10] EUPASS. Final report to the European commission. Agreement reference number VS/1999/5133 (99CVF3-502), 2001. pdf.

[11] FRANCHET Y. Memorandum from Mr Yves Franchet, Director General, Eurostat. 1998. html.

[12] GIFI A. Nonlinear multivariate analysis. New York, Wiley, 1990.

[13] GRAIS B. Statistical Harmonisation and Quality: The case of Social Statistics. Paper presented to the fourth Mondorf Seminar, Eurostat, 26/27 March 1998.

[14] GÜNTHER R. Report on compiled information. CHINTEX – The Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel. Working paper #19, 2003. pdf.

[15] HATTIE J. Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement 1985, 9, 139-164.

[16] HARKNESS JA. Questionnaire translation. In JA Harkness, FJR van de Vijver, PPh Mohler, eds., Cross-Cultural Survey Methods (pp. 35-56). New York, Wiley, 2003.

[17] HOPMAN-ROCK M, ODDING E, HOFMAN A, KRAAIMAAT FW, BIJLSMA JWJ. Physical and Psychodisability in Elderly Subjects in Relation to Pain in the Hip or Knee. Journal of Rheumatology 1996, 23, 1037-1044.

[18] HOLLAND PW, WAINER H (Eds.). Differential Item Functioning. New York: Lawrence Erlbaum, 1993.

[19] JODOIN MG, GIERL MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education 2001, 14(4), 329-349.

[20] KOLEN MJ, BRENNAN RL. Test equating: Methods and practices. New York: Springer, 1995.

[21] KRAMERS PGN. The ECHI project. European Journal of Public Health 2003, 13, Supplement 1, 101-106.

[22] LINACRE JM. What do Infit and Outfit, Mean-square and Standardized mean? Linacre JM. Rasch Measurement Transactions, 2002, 16:2 p.878. html.

[23] MASTERS GN. A Rasch Model for Partial Credit Scoring. Psychometrika 1982, 47, 149-174.

[24] MONTSERRAT A, SICARD F, Building a European Health Survey System: Improving information on self-perceived morbidity and chronic conditions. Paper presented at the Working Party Morbidity and Mortality, Luxembourg 20 January 2004. pdf.

[25] ODDING E, VALKENBURG HA, ALGRA D, VANDENOUWELAND FA, GROBBEE DE, HOFMAN A. Association of Locomotor Complaints and Disability in the Rotterdam Study. Annals of Rheumatic Diseases 1995, 54, 721-725.

[26] RASCH G. On Specific Objectivity: An attempt at Formalizing the Request for Generality and Validity of Scientific Statements. In M. Glegvad (eds.), The Danish Yearbook of Philosophy 1977. Copenhagen: Munksgaard, 58-94.

[27] RAJMIL L, HERDMAN M, FERNÁNDEZ DE SANMAMED MJ, DETMAR S, BRUIL J, RAVENS-SIEBERER U, BULLINGER M, SIMEONI M-C, AUQUIER P, AND THE EUROPEAN KIDSCREEN GROUP. Generic Health-related Quality of Life Instruments in Children and Adolescents: A Qualitative Analysis of Content. Journal of Adolescent Health 2004, 34, 37-45.

[28] RAVENS-SIEBERER U, GOSCH A, ABEL T, AUQUIER P, BELLACH B-M, DÜR W, RAJMIL L AND THE EUROPEAN KIDSCREEN GROUP. Quality of life in children and adolescents: a European public health perspective. Social and Preventive Medicine 2001, 46, 297-302.

[29] ROBINE J-M, JAGGER C, EGIDI V. Selection of a Coherent Set of Health Indicators. Final draft. A First Step Towards A User's Guide to Health Expectancies for the European Union. Montpellier (France), Euro-REVES, 2000. pdf.

[30] RUMM LABORATORIES. Rumm 2010. Rasch Unidimensional Measurement Models, 2001. http.

[31] RUMM LABORATORIES. Rumm 2020. Rasch Unidimensional Measurement Models, 2003. http.

[32] RÜTTEN A, ZIEMAINZ H, SCHENA F, STAHL T, STIGGELBOUT M, AUWEELE YV, VUILLEMIN A, WELSHMAN J. Using Different Physical Activity Measurements in Eight European Countries: Results of the European Physical Activity Surveillance System (EUPASS) Time Series Survey. Public Health Nutrition 2003, 6, 371-376.

[33] RÜTTEN A, VUILLEMIN A, OOIJENDIJK WT, SCHENA F, SJOSTROM M, STAHL T, VANDEN AUWEELE Y, WELSHMAN J, ZIEMAINZ H. Physical activity monitoring in Europe. The European Physical Activity Surveillance System (EUPASS) approach and indicator testing. Public Health Nutrition 2003, 6, 377-84.

[34] SCHULMAN JA, TRUJILLO MJ, KARNEY BR. Facets: Computer software for evaluating assessment tools. American Journal of Health Behavior 2001, 25(1):75-77.

[35] SMITH TW. Developing comparable questions in cross-national surveys. In JA Harkness, FJR. van de Vijver, PPh Mohler, eds., Cross-Cultural Survey Methods (pp. 69-91). New York, Wiley, 2003.

[36] UNECE. ENECE/WHO/EUROSTAT Meeting on Health Statistics. Geneva, 24-26 May, 2004. http://www.unece.org/stats/documents/2004.05.health.htm.

[37] VAN BUUREN S, HOPMAN-ROCK, M Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory. Statistics in Medicine 2001, 15;20(7):1061-76.

[38] VAN BUUREN S, EYRES S, TENNANT A, HOPMAN-ROCK M. Response conversion: A new technology for comparing existing health information. TNO report 2001.097. Leiden: TNO Prevention and Health, 2001. pdf.

[39] VAN BUUREN S, EYRES S, TENNANT A, HOPMAN-ROCK, M. Assessing compara-bility of dressing disability in different countries by Response Conversion. European Journal of Public Health 2003, 13, Supplement 1, 15-19.

[40] VAN BUUREN S, EYRES S, TENNANT A, HOPMAN-ROCK M. Improving comparability of existing data by Response Conversion. Journal of Official Statistics, under review, 2004.

[41] VAN DETH JW (Ed.). Comparative politics. The problem of equivalence. London: Routledge, 1998.

[42] VAN DE VIJVER FJR, LEUNG K. Methods and data analysis for cross-cultural research, Thousand Oaks: Sage, 1997.

[43] WAINER H, THISSEN D. Estimating Ability with the Wrong Model. Journal of Educational Statistics 1987, 12, 339-368.

[44] WARM TA. Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika 1989, 54, 427-450.

[45] WRIGHT BD, MASTERS GN. Rating scale analysis: Rasch measurement. Chicago, MESA Press, 1982.