United Nations Statistics Division
United Nations Children's Fund
Statistical Office of the European Communities
Centres for Disease Control and Prevention
of the United States of America

ESA/STAT/AC.81/3-2
24 May 2001

International Seminar on the Measurement of Disability

New York
4-6 June 2001

J.L.A. van Rijckevorsel:
Cross population comparison of surveys: A review of new technologies.

TNO Prevention and Health

Division Public Health
Wassenaarseweg 56
P.O. Box 2215
2301 CE Leiden
The Netherlands

www.tno.nl

T +31 71 518 1752
F +31 71 518 19 21
JLA.vanRijckevorsel@pg.tno.nl

TNO report

2001.108

Cross population comparison of surveys: A review of new technologies

Date June 2001

Authors J.L.A. van Rijckevorsel
S van Buuren
M.W. de Kleijn – de Vrankrijker

Authors: J.L.A. van Rijckevorsel
S van Buuren
M.W. de Kleijn – de Vrankrijker

This report can be ordered from TNO-PG by transferring ƒ 110.00 (excl. VAT) to account number 99.889 of TNO-PG Leiden. Please state TNO publication number PG/VGZ/2001.108.

Summary

This report is a non-technical review of new statistical techniques that can be used for cross population comparison of health surveys. It is based on the work of van Buuren et al. (2001) and of Murray et al. (2000). A distinction is made between pre- and post harmonization of surveys and how additional exogenous information is used for the comparison between nations, member states or socio-economic groups. Ten different methods are discussed, which fall apart into three groups: comparable scale construction, fixed ability comparisons and response conversion. Most methodology leans strongly on psychometric theory of latent traits. This kind of technology needs intensive counselling by statistical experts in order to be able to be used in a realistic setting.

Contents

1 Introduction

1.1 The purpose of this report

This report is an abstract for a representation on the international seminar of the measurement of disability on 4-6 June 2001, New York, organized by the UNDP. It gives a short overview of new statistical initiatives for cross-population comparison of health surveys for the layman. Cross-population means across place: nation to nation, different socio-economic groups within nations and across time: nation in year 1 vs nation in year 2. For technical discussion and treatment see the list of references. The main technical references are the following: van Buuren et al. (2001) discuss statistical comparison for general health status measurement by surveys in the context of the health monitoring program (HMP) of the European commission (EC), Murray et al. (2000) and Hopman et al. (2000) restrict the comparison to field to WHO disability surveys and Kolen et al. (1995) discuss statistical comparison techniques from the psychometric point of view.

1.2 What is the problem of cross-population comparability?

In the EU the objective of the Health Monitoring Program is to set up a system in which the health of different Member States in the European Union can be compared. This system will have to be based on existing population surveys. This requirement introduces new issues regarding the comparability of information across Member States. The present section outlines some complexities of the comparability problem.

Suppose that we are interested in comparing two populations, and that we have access to one survey for each population. Each survey provides information on a sample of respondents. Survey instruments typically consist of a standardised set of *questionnaire items*, like the SF-36 or the OECD disability indicator. For a given field of health, we may be able to identify specific instruments or items that measure the aspect of health.
If both studies use equivalent instruments/items, there would be (in principle at least) no problems regarding the comparability of content. In practice however all studies so far cannot compare such studies satisfactorily. Murray (2000) and Sadana (2000) raised questions about the comparability. The studies could still differ in their sampling methods, in their ways for collecting data (e.g. interview, self-report), or in other ways. Those differences have to be accounted in any valid comparison.
This report also concerns the problem that target studies may contain measurements of the same thing, but using different instruments or items. Let A and B denote two target items that measure the same characteristic. In general, responses on A and B can only be meaningfully compared if the scales on which they are measured have the same origin and the same unit. If A and B are different, it is not informative to directly compare their responses since differences in the response distribution of A and B may be due to:
1 real differences between populations;
2 systematic differences between the items;
3 a combination of both.
In practice, interest focuses on comparing (sub)populations, which presupposes that possibility 1 is true. Without any additional information or assumptions, it is however impossible to distinguish between the three possibilities. Thus, we generally do not

know whether differences between the responses on A and B reflect real population differences.

1.2.1 Some types of comparability problems

In general nations contribute to the development of the measurement of health status based on their own specific policies, but these initiatives have not always been co-ordinated in any major way. This has resulted in consequences that data and information are often of limited comparability between countries and sometimes of medium or poor quality.
A cross-population health monitoring system will bring together data collected in different nations. It will be clear that any differences in data collection methodologies should be accounted for before these data can be used to provide comparative information across nations. Incomparability may occur at different levels:

- Appropriate data may not be collected at all in some nations;
- Some nations collect appropriate data for specific sub samples, or with special designs;
- The definition of diseases or disabilities may differ between nations, e.g. by using different classifications; and/or levels. e.g. impairment vs. activity.
- The (meaning of the) wording of the question or the formulation of the response categories can differ.

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

For example, for 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 200yds" and "200yds or more". By contrast, 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 of the respondent, but it is far from clear how the answer on the U.K.-item can be compared with those on the Dutch item.

1.2.2 Pre-harmonisation and post-harmonisation

There are two broad strategies to deal with incomparability: pre-harmonisation and post-harmonisation according to van Buuren et al. (2001).
Pre-harmonisation is the royal road to solve comparability problems. The idea is that, once and for all, all nations will start collecting comparable data. The major advantage is that comparability is guaranteed since every office works in the same way using the same instrument. As easy as this may sound however, it is not trivial to actually achieve this in practice. The national data collecting agencies of the individual nations will generally be very reluctant to change their sampling methods and instruments. Their major argument is that a change of the current practice will break the comparability to historic data. In that case, pre-harmonisation does not solve the problem, but puts it on a different level, that is, at the level of the national offices of the nations.

Methods for cross-cultural comparison fall apart in to two groups: (1) methods that require pre-harmonization e.g. that different surveys have the same items and (2) post harmonization methods that permit different wording and number of categories per item for each population. The first group of methods consists of comparable scale construction and fixed ability comparisons and is discussed in section 3.1 and 3.2. The second group of methods is called response conversion and is discussed in section 3.3.

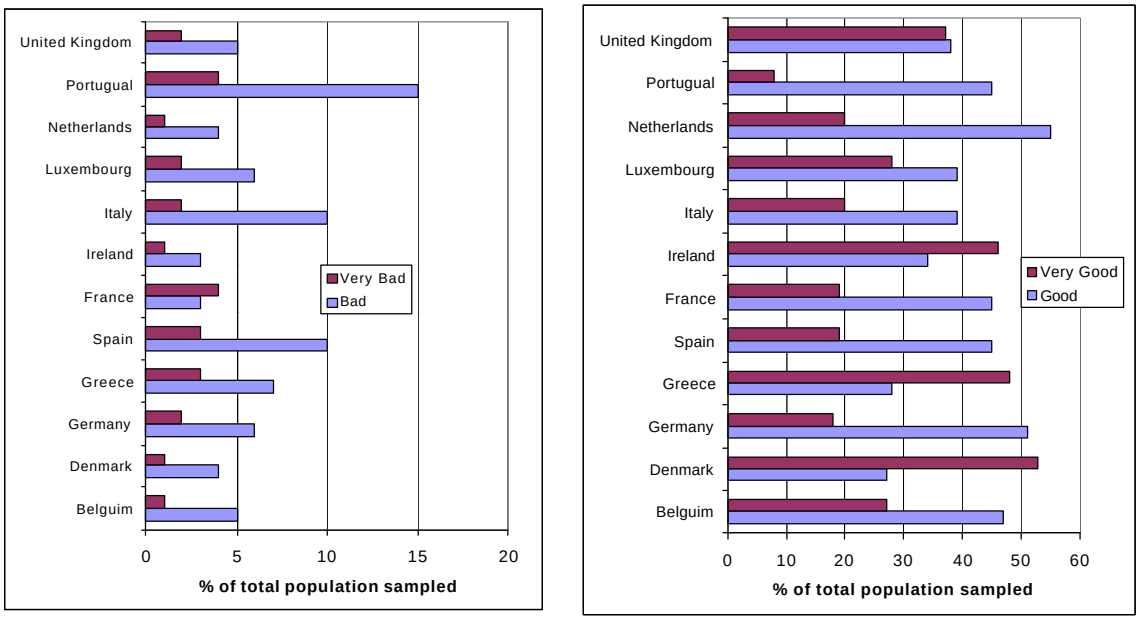

Figure 1 Response on the question "How is your health in general?" in 12 European countries (Source: Sadana et al. 2000).

By its nature pre-harmonisation will only work for new, and not for existing data. In addition, even if done well, pre-harmonisation could still yield implausible results that will raise comparability issues. Consider the single question “How is your health in general?” on a five point Likert response scale “very good, good, fair, poor, very poor”. This question was posed (after translation) in 12 countries of the European Union, using on the same survey and methods within the context of the 1994 European Community Household Survey (Eurostat, 1997). Figure 1 is taken from Sadana *et al*. (2000) and contains the age-sex-standardized proportions of the responses per country. Note that the category ‘very good’ health is reported by as much as 53% of the Danish and as little as 8% of the Portuguese population. Also, nontrivial differences occur for the bad and very bad categories. These differences can be context or horizon related and methodologically they suggest that pre-harmonisation may not be enough to solve all comparability problems.

Post-harmonisation is the murky way to solve comparability problems. The idea is that we can somehow transform incomparable data into a comparable version, and use the latter in our analyses. The big advantage is that we can use existing data. The disadvantage is that we often do not know what the transformation should be, and whether applying it will affect the results. In addition, it is sometimes simply impossible to transform the data into a comparable form without making strong, untestable assumptions. On the other hand, post-harmonisation is often the only option if we are to make any progress. Given that situation, we should try to use the best avail-able scientific technology to make post-harmonisation work. This implies that we should be explicit about the concepts, assumptions and limitations of the method.

There exist combinations of less formal methods that are used in the revision of the ICIDH. Trotter et al. (2000) call this the cross-cultural applicability research (CAR) methods.. Such methods are a mix of ethnographic, rapid assessment and statistical methods to identify most stable items across cultures and to identify those items that are problematic. This information is hence used in the revision. CAR is a post harmonization technique that is not discussed in this paper because of its substantial non-statistical content.

1.3 Contents

Chapter I introduces the problem of cross-cultural comparability. It describes the essential concepts and main assumptions of the method. Chapter 2 discusses the general principles behind these survey comparability methods. Chapter 3 addresses the particular methods. Chapter 4 concludes this reports, and discusses the usefulness of the methodology.

2 Key words of survey comparison methods

2.1 The common scale

Health is a multidimensional concept and each dimension like vision, hearing etc. is considered to be a scale on which the true attainment of a person is measured. Statistical techniques obtain formal, mathematical representations of such scales. The best-known representations of this kind are the principal components and latent traits. Comparison methods assume the existence of a continuous *latent trait* θ that underlies all items. The latent trait θ can be interpreted as walking disability. A latent trait is a theoretical construct with some of the following properties. A latent trait varies continuously and can take on all values. The ability level of each person in the sample can be characterised by a position θ_i on the trait. The trait is latent, which means that it cannot be observed directly. So the "true value" of θ_i for person *i* is not known, and can only be observed through the manifest item responses.
The main idea is that the value of the latent trait governs the probability of responding in a specific response category. For low θ_i (e.g. no disability), the probability of answering in the most severe disability response categories is low.

The characteristics of such a scale or trait are always that you cannot observe the scale directly, that each person has a position on the scale and that the scale can take al values. All techniques discussed in this paper use the concept of a common scale.
Latent traits or common scales can be obtained by a multitude of models. From a practical point of view, the actual differences (when fitted to data) are usually not that large. All models do more or less the same, but the results have different theoretical properties.
Such models will be not discussed in this report. Appropriate references are given where needed.

2.2 Response category cut-points

The best example to illustrate what a response category cut-point is measuring age in different countries. If one asks the age of a person in country A and the answer is 20 and the same question is asked in country B and the answer is also 20, then the assumption of cross-cultural consistency of response category cut-points is that the answers are comparable because, in spite of cultural differences, both answers coincide with the response category cut-point of 20 years. The actual ages can be different like 20 and two months in country A and 20 years and 9 months in country B. So there is one category cut-point for the age of twenty for both countries.
The idea of response category cut points applies directly to the common scale. Because if in a health status survey a person has a certain position on a common scale that reflects his/her ability to perform a certain task, then you can define some cut points in that scale that separate persons with a certain degree of ability from other persons with another degree of that ability. Such points are called response category cut points and they are representative for a population
A response category refers to the group of persons between two successive cut points on the common scale. The response category labels the individuals into one group. Response categories are the keystones of comparison techniques. *The general assumption is that different populations have the same response category cut-points*.

Empirical evidence that confirms the assumption is only available if one is prepared to make additional assumptions on the comparability of exogenous measurements.

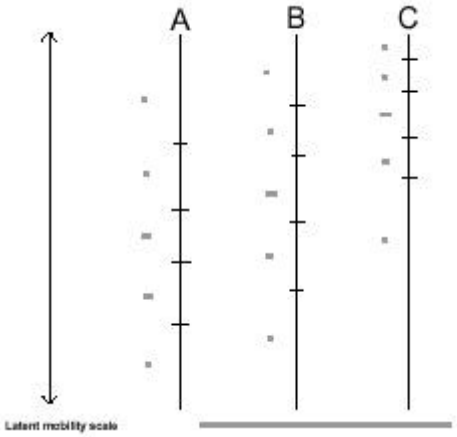

Figure 2 Mapping from latent mobility variable to categories
(Source: Murray et al. 2000)

2.3 Calibration by exogenous information

The role of information exogenous to the survey plays a vital role in the comparison of studies because one is always looking for an element that two surveys have in common. That is the natural starting point for comparison. Such information has an anchoring or calibrating role. The surveys to be compared have always some kind of reference to this anchor.

In general an anchor can be one common item and the same cut-points; the most extreme case of this kind of anchoring is when two surveys have identical sets of items. More strictly one could say that an item is common if one is willing to assume that it has the same cut-points in different populations. If one is not willing to make this assumption beforehand, for instance because empirically the assumption cannot be confirmed, than one should conclude that the items are different and that there is no comparison possible. In psychometric test research such items are removed from the set as differential items. In psychological tests this is never a problem because there exits an abundant number of items. In health surveys it is a problem because there are few candidates that could serve as “common” items.

Other cases of anchoring are when there exists an extraneous measurement of the very phenomenon that the surveys are supposed to measure like a physical measurement. Murray calls this measured tests and he implicitly assumes that different populations have the same cut-points. Or that two surveys refer to a homogeneous group albeit that one survey is sampled in nation A and the other survey in nation B. When exogenous information is used the outcome of the comparison sometimes can be evaluated by inspecting the plot of the estimate of the domain scale based on the exogenous

information against the estimate of the domain scale obtained by self-reports for each population or group to be compared. All these cases and more will be discussed in Chapter 3.

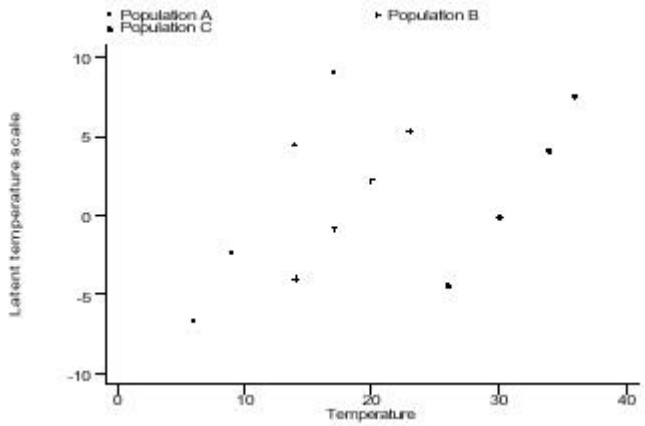

Figure 3 True cut-points versus estimated using Rasch model (PCM) for Q2 in all 3 populations (Source: Murray et al. 2000).

3 Methods for comparison

Methods for cross-cultural comparison fall apart in to two groups: methods that require that different surveys have the same items and methods that permit different wording and number of categories for each population. The first group of methods consists of comparable scale construction and fixed ability comparisons and is discussed in sections 3.1 and 3.2. The second group of methods is called response conversion and is discussed in section 3.3.

3.1 Comparable scale construction

.

3.1.1 Item response theory (IRT)

IRT is a psychometric technique that establishes response category cut-points, usually called thresholds, on a common latent scale. The attractivity is that the cut-points and the individual response have positions on that scale and for each item there is a probability function that shows the probability of an individual's response as a function of the response category cut-point on the scale. Each item has its own curve on the scale.

The concept for comparison is that surveys in different populations have the same common scale and have the same cut-points. According to Murray et al. (2000) the latter need not hold in practice. Since cross-cultural consistency is at the heart of IRT, Murray raises doubts about the usefulness of IRT for solving comparability problems.

Characteristics:

- ✓ Pre-harmonized data
- ✓ Same items in different populations
- ✓ No exogenous information

3.1.2 Measured tests and the HOPIT model

The combination of exogenous information obtained by other tests, other than self reported items, like the Snellen's chart for visual acuity, with a statistical representation of the response category cut-points is way to calibrate the common scale. Murray (2000) gives several examples where the hierarchical ordered probit model (HOPIT) is used to estimate the common scale with Snellen's E chart. In an artificial temperature sample he also shows that the common scale calibrated by real measurements of temperature is superior to uncalibrated common scale. In a case study on vision from WHO pilot surveys however the cut points have different positions per population on the common scale.

Characteristics:

- ✓ Pre-harmonized data
- ✓ Same items in different populations
- ✓ Raw Exogenous information

3.2 Fixed ability comparisons

3.2.1 Principal components analysis

PCA or factor analysis is a technique that looks for the common denominator or scale among a set of items per country based on linear combinations of the items. Items and individuals that scored on the items have positions on the common scale. Sadana et al.

(2000) use PCA for cross-cultural comparison and anchor the scales per country in such a way similar to health utility indexing, that the most healthy persons across-cultures are placed in the same position on the extreme of the scale. An identical procedure is followed for the most unhealthy persons. This creates one scale on which the health status of different cultures is represented

The technique assumes that the most extreme status of health is equal across cultures and that this status occurs in the study sample. Sadana et al. conclude that using this method it is difficult to create a common comparative scale, although the scales per country are reliable

Characteristics:

- ✓ Pre-harmonized data
- ✓ Same items in different populations
- ✓ Scaled anchor information

3.2.2 Using vignettes

If measured tests do not exist Murray proposes to fix the level of the health status on a domain and assess the variability of the responses on such an item by different populations or groups that are to be compared by the HOPIT model. Such a fixed status serves as an anchor and is called a vignette. In health utility assessment a similar strategy is used. A vignette is a standard case with a fixed level of ability on a given domain to be evaluated by the respondent. Murray report a good level of agreement

Characteristics:

- ✓ Pre-harmonized data
- ✓ Same items in different populations
- ✓ Scaled anchor information

3.2.3 Using comparable homogeneous groups

If the same items are used in two populations to be compared one can draw homogenous samples from both populations that are comparable on all other relevant aspects. This theoretically excludes the effect of covariates, provided the criteria for homogeneity are appropriate

Characteristics:

- ✓ Pre-harmonized data
- ✓ Same items in different populations
- ✓ Exogenous information by post harmonizing grouping of respondents

3.3 Response conversion

The present section is primarily concerned with ways to cope with differences in wording and categories and depends on the assumption of cross-cultural consistency of cut-points. Van Buuren et al. (2001) distinguishes three major strategies to address to comparability issues if items are not the same across populations: by fiat, by linking item and linking study. In practice, one typically makes a combination of these options. Response conversion is a method that in the second and third strategy by systematically exploits any information overlap between different studies. Overlap can occur in items, in samples, or in both, leading to different linked data matrices.

The major tasks in the practical application of response conversion consist of

1. Identification and construction of the linked data matrix;
2. Construction of a conversion key that place different items on a common scale;
3. Application of conversion key to estimate disability on a common scale.

Steps 1 and 2 need to be done only once, where step 2 results in a conversion key. A separate conversion key is needed for each topic. Once a conversion key is

available, applying it to new data is cheap and easy, and can be done on a routine basis.

3.3.1 By fiat

Assume a common score system, recode the responses using by panels of experts into a common system, and compare;

Characteristics:

- ✓ Post-harmonization technique
- ✓ Different items in different populations
- ✓ No exogenous information

At first sight, the first strategy (*by fiat*) may seem appealing. If we would have a way to recode the responses on both items into a comparable system, then we can simply use the recoded data to gain insight into differences in walking disability between both samples. We have solved the comparability problem by "assuming away" any systematic differences that might exist.

Some comments are in order on this strategy. First, it is only possible to move into the direction of the item with the lowest number of response categories. This will inevitably lead to a loss of information for items that have more refined response systems. In principle, one could try to solve this problem by splitting crude category into refined sub-categories. It is however difficult to see how such splitting proportions should be chosen. The whole procedure relies on arbitrary and untested criteria, and could therefore generate considerable debate. There is no way of knowing whether the chosen cut-point are actually correct. The *by fiat* strategy should therefore only be chosen in cases where 1) the possibility of dispute is relatively small, 2) the response categories are finely grained, and 3) a clear authority can endorse the system.

3.3.2 Link by item

Identify additional items on walking disability (within both studies) that are common to both studies, and exploit the overlap to compare studies;

Characteristics:

- ✓ Post-harmonization technique
- ✓ Different items in different populations
- ✓ Anchoring information

If two studies contain additional items on walking disability that are common to both studies II and I, then this information provides a link between both studies. An item that connects two studies is called a *bridge item.*

The item C provides a means to compare both studies. Simple visual inspection of the category frequencies for both studies tells us that, say, the population of study I is more disabled than in study II. We have simply replaced an incomparable set of items (A and B) by a comparable item (C) that happened also to be administered in both studies. Of course, we could have started with the C right away, and not be concerned with either A or B at all. In this sense, A and B are redundant.

Now imagine that we have two *new* studies, where the first contains only A (but not C) and the second contains only B (but not C). The interesting question then is: It is possible to use the information contained in the category frequencies of the first two studies in such a way that we can validly compare the two new studies, even in the absence of bridge items? The answer is yes, given that both of the following assumptions hold:

- the bridge item measures the same characteristic as the target items;

- the bridge item is equivalent in both studies, implying cross-cultural consistency of cut-points;

If true, it is possible to define a statistical model for converting observed scores into a comparable form. In later applications, this model can be used to convert information without the need of any bridge items.

3.3.3 Link by study

Look for other (third) studies that contain both items, and use the relation between both items in comparing both original studies.

Characteristics:

- ✓ Post-harmonization technique
- ✓ Different items in different populations
- ✓ Exogenous information

The third strategy (*link by study*) is the logical complement of the second. Suppose a third study is available, that administers both target items to a third population. Such a study is called a *bridge study.*

Item	Description	Response categories	Study		
			ERGOPLUS	BRIDGE	EURIDISS
			n=306	*n=300*	*n=292*
SI01	I walk shorter distances or often stop for a rest.	0 = No	276	215	
		1 = Yes	28	85	
GAR9	Can you, fully independently, walk outdoors (if necessary, with a cane)?	0 = no difficulty		150	145
		1 = some difficulty		105	110
		2 = much difficulty		34	29
		3 = only with help		11	8

Table 2 contains an example of observations from a hypothetical bridge study. The sample size (n=300) of the bridge study is chosen to be similar to the target studies for ease of comparison. Equality of sample sizes is not a requirement in actual application. The bridge study administers both SI01 and GAR9 to a third population. The comparison of the score distributions on GAR9 suggests that the disability of the bridge population is almost equal to that in de EURIDISS study. In contrast, the difference on SI01 with the ERGOPLUS study is substantial. Combining these two findings suggests that, like before, the ability level in ERGOPLUS is higher than in EURIDISS.

The validity of the link-by-study strategy depends on the following assumptions:

- the items in the bridge study are equivalent to those in the target studies;
- the relation between both items does not depend on the ability level of the sample.

It is important to observe that it is not required that the ability level of the bridge study is comparable to one of the target studies. The second assumption implies instead that the relation between the items is assumed to be of the same in all studies. This condition is much weaker.

4 Discussion

Several conclusions can be drawn given the present state of the art:

1. The comparison problem not only pertains to the comparison between populations but also to groups differing in time, place or socio-economic status within one population.
2. The need for a statistical representation of an ability on a dimension as a latent trait or common scale is generally recognized.
3. The key to measurement, and thus comparability is cross-cultural constancy of response category cut-points.
4. There exist various different appropriate models like IRT and HOPIT that in practice produce response category cut-points.
5. For cost-effective and political reasons there exist the need to compare surveys where not all items are identical or have the same wording between populations.
6. The optimal cross-cultural comparison does not necessarily mean all nations share exactly the same surveys.New statistical comparison techniques allow for different items and wording provided some items are identical.
7. From an IRT point of view the only important requirement for cross cultural comparison is the constancy of response category cut-points. Wording and response categories may differ.
8. If the response category cut-points are not constant there exists a validity problem: Do we measure what we want to measure ?
9. To solve the validity problem one makes extra assumptions on exogenous information to constant in different populations.
10. In the (re-) design of surveys the possibility of adjustment, linkage and use of exogenous information should be build in as a default.
11. The cross-cultural comparison requires generous access to micro-data and correctly and fully documented data.
12. Cross-cultural comparison requires considerate statistical modelling expertise that differ from the statistical level of survey design.

5 References

1. EUROPEAN COMMISSION, *Programme of community action on health monitoring. Work programme 1998-1999.* Art. 5.2.b of Decision 1400/97/EC. Luxembourg: European Commission, 1998.
2. EUROSTAT, Self reported health in the European Community. *Statistics in Focus, Populationand social conditions.* ISSN 1024-4352.
3. HAMBLETON RK, SWAMINATHAN H, ROGERS JH. *Fundamentals of item response theory*. Newburg Park: Sage, 1991.
4. KOLEN MJ, BRENNAN RL. *Test equating: Methods and practices*. New York: Springer, 1995.
5. MURRAY CJL, TANDON A, SALOMON J, MATHERS CD. *Enhancing Cross-Population Comparability of Survey Results.* GPE. Discussion Paper series: No. 35. Geneva: World Health Organization, 2000.
6. SADANA R, MATHERS CD, LOPEZ AD, MURRAY CJL, IBURG KM. *Comparative analyses of more than 50 household surveys on health status.* GPE Discussion Paper Series: No.15., Geneva: World Health Organization, 2000.
7. TROTTER RT, REHM J, CHATTERJI S, ROOM R, ÛSTUN TB. Cross-cultural Applicability Research in: ÛSTUN TB et al. (eds): *Disability and Culture: Universalism and Diversity.* ICIDH-2 Series. Seattle: Hogrefe & Huber, 2001
8. VAN BUUREN S, EYRES, TENNANT A, HOPMAN-ROCK M. Response conversion: a new technology for comparing existing health information. Leiden: TNO Report 2001.097. To appear.
9. VAN BUUREN S, HOPMAN-ROCK M. Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory. *Stat Med 2001, 20,* 1061-76.