

Background Paper for the Technical Consultation on Responsiveness Concepts and Measurement

**13–14 September 2001
Geneva, Switzerland**



**World Health Organization
Geneva**

© World Health Organization 2001

This document is not issued to the general public, and all rights are reserved by the World Health Organization (WHO). The document may not be reviewed, abstracted, quoted, reproduced or translated, in part or in whole, without the prior written permission of WHO. No part of this document may be stored in a retrieval system or transmitted in any form or by any means—electronic, mechanical or other—without the prior written permission of WHO.

The views expressed in documents by named authors are solely the responsibility of those authors.

Department of Health Financing and Stewardship
World Health Organization
1211 Geneva 27
Switzerland

Contents

TABLE OF FIGURES	V
INTRODUCTION AND RATIONALE FOR WHO'S WORK ON RESPONSIVENESS	1
1. UNDERSTANDING AND MEASURING RESPONSIVENESS	2
1.1. A conceptual framework for understanding responsiveness	2
1.1.1. <i>The domains</i>	2
1.1.2. <i>The population</i>	3
1.1.3. <i>Scope of the health system and legitimate expectations</i>	3
1.1.4. <i>Preferences</i>	3
1.2. Principles of measuring responsiveness	3
1.2.1. <i>An underlying latent variable and response category cut-point shift</i>	4
1.3. Overview of method of calculating the level of responsiveness	6
1.3.1. <i>The raw data - self reports</i>	6
1.3.2. <i>The raw data – the set of vignettes</i>	7
1.3.3. <i>Hopiting the raw data</i>	8
1.3.4. <i>Raw versus adjusted scores</i>	10
1.3.5. <i>Other methodological steps involved in the calculation of the measure for a country</i>	11
2. OVERVIEW OF THE SURVEY PROGRAM AND FINDINGS	12
2.1. Description of the survey program	12
2.2. The responsiveness questionnaire	13
2.2.1. <i>The instruments in different household modes</i>	13
2.2.2. <i>Questionnaire structure</i>	13
2.2.3. <i>Response scales</i>	14
3. FINDINGS FROM ANALYSES OF 50 SURVEYS	15
3.1. Overall scores	15
3.2. Domains showing the highest responsiveness, on average	18
3.3. Responsiveness and health system expenditure	18
3.4. Importance of various domains	19
3.5. Respondents reporting personal health care encounters in the last 12 months (averaged over 50 surveys)	19
3.6. Results of adjustment on rankings	21
3.7. Scores by age groups	22
4. REVISITING ISSUES ON WHO'S RESPONSIVENESS RESEARCH AGENDA	23

4.1.	WHO research strategy aims	23
4.2.	Qualitative research	23
4.3.	Specific research questions and answers	24
4.3.1	<i>Domains and terminology</i>	24
4.3.2	<i>Health system activity</i>	25
4.3.3	<i>Operationalizing domains</i>	26
4.3.4	<i>Response categories</i>	28
4.3.5	<i>Users and survey methods</i>	29
4.3.6	<i>Experience reference period</i>	30
4.3.7	<i>Vignettes and the scale</i>	31
4.3.8	<i>Biases in different modes</i>	33
4.3.9	<i>External calibrators and validity checks</i>	33
4.3.10	<i>Linkages to functions (or determinants of better responsiveness performance)</i>	34
	BIBLIOGRAPHY	35
	Appendix 1: Identifying the Components of Individuals' Satisfaction: Review of the Literature	35
	Appendix 2: Elaboration on the Understanding of the Responsiveness Domains	36

Table of Figures

FIGURE 1: MAPPING FROM THE LATENT VARIABLE TO CATEGORIES	5
FIGURE 2: ESTIMATED CUT-POINT SHIFT FROM NHANES DATA	6
FIGURE 3: COMPARING FREQUENCIES IN THE RAW DATA FOR TWO COUNTRIES	7
FIGURE 4: COMPARISONS OF VIGNETTE RESPONSES IN THE RAW DATA	8
FIGURE 5: A COMPARISON OF THE CUT-POINTS FOR EGYPT AND THE UK	9
FIGURE 6: OUTPUT FROM HOPIT MODEL REGRESSION	10
FIGURE 7: RAW VERSUS ADJUSTED SCORES FOR OUTPATIENT DIGNITY	10
FIGURE 8: OVERLAPPING SURVEY DESIGN	12
FIGURE 9: OVERVIEW OF THE QUESTIONNAIRE	13
FIGURE 10: INPATIENT AND OUTPATIENT RESPONSIVENESS SCORES	15
FIGURE 11: OVERALL RESPONSIVENESS SCORES (N=50)	16
FIGURE 12: AVERAGE DOMAIN SCORES, ACROSS COUNTRIES (N=50)	18
FIGURE 13: THE IMPORTANCE OF THE DOMAINS (AVERAGE RESULTS FROM THE SURVEYS)	19
FIGURE 14: PERCENTAGE OF RESPONDENTS ON AVERAGE	20
FIGURE 15: COMPARISON OF UTILIZATION RATES BY SEX	20
FIGURE 16: OUTPATIENT DOMAINS: RAW VERSUS ADJUSTED EFFECT ON RANKS	21
FIGURE 17: RESPONSIVENESS SCORES FOR OUTPATIENT DIGNITY BY AGE GROUP	22
FIGURE 18: CORRELATION OF THE COUNTRY LEVEL DOMAIN SCORES	25
FIGURE 19: COGNITIVE INTERVIEW TESTING IN 2000	27
FIGURE 20: RAW VERSUS ADJUSTED SCORES FOR 6 DOMAINS (N=50)	28
FIGURE 21: RANK CORRELATION COEFFICIENTS ACROSS COUNTRIES FOR THE SET OF DIGNITY VIGNETTES	32
FIGURE 22: MEAN CORRELATION SCORES FOR DIFFERENT SETS OF VIGNETTES (N=65)	32

Introduction and rationale for WHO's work on responsiveness

The *World health report 2000* introduced a new framework for the assessment of health system performance and initiated a process of more rigorous scientific work and analysis on this subject. Within this framework, the goal of the responsiveness of health systems to the legitimate expectations of the population was identified as a key goal, along with achieving better health (average level and a more equal distribution) and greater fairness in the way health systems are financed.

The emphasis on the goal of having responsive health systems is in line with WHO's own *Strategy on Health and Human Rights (Draft, September 2000)* and other more general UN covenants such as the *International Covenant on Economic, Social and Cultural Rights*, to which many of WHO members are signatories. While there is both WHO and UN-wide organizational precedence for reviewing and monitoring the process of care, there is little guidance and precedence for how this should be formalised, both conceptually and methodologically.

Accordingly, consultation work leading up to and since the launch of the *World health report 2000*, has focused on the development and improvement of the responsiveness concept and its measurement. Several steps were taken towards ensuring that the research on responsiveness was of the highest quality. The main ones involving outside experts were a responsiveness meeting of experts from 35 different countries and the secondment of a quality of care expert from the U.S. *Agency for Health Research and Quality* to help develop the concepts and methods.

The research completed by WHO has been documented in several Global Programme for Evidence Department's *Discussion Papers* and work is underway to document the findings of the most recent set of household and key informant surveys. These papers are available on the WHO website (<http://www.who.int/evidence>) and have been widely read and discussed by critiques of the WHO Responsiveness Indices and performance measurement approach. Two of the more relevant papers have been distributed to delegates of this technical consultation: the conceptual paper, documenting the research behind the development of the concept; and the strategy paper documenting the research strategy for improving responsiveness concepts and methods.

This technical consultation represents a continuation of WHO's effort to consult experts in this field on how to improve the concept and measurement. The main aim of the work to-date on responsiveness has been to develop a standardized approach to measuring responsiveness that will be useful to countries, both to their populations and to policy makers. The main objectives of the consultation are:

- to discuss the conceptual and methodological challenges to responsiveness research;
- to prioritise the conceptual and methodological challenges for responsiveness research;
- to outline potential policy applications (e.g. sub-national analysis); and
- to outline a future research plan.

1. Understanding and measuring responsiveness

1.1. A conceptual framework for understanding responsiveness

Responsiveness has been defined as a measure of how well the health system responds to the population's legitimate expectations of non-health aspects of health services. Research on the concept of responsiveness began in 1999, and development of the concepts drew on a broad literature review of materials in the patient satisfaction and quality of care literature. The studies reviewed relevant to the responsiveness domains are referenced in Appendix 1.

The boundaries of the health system, defined by all actions whose primary intent are to improve health in the health system performance framework, applies in the context of responsiveness. Therefore we are concerned about the responsiveness of all actors in the health system who interact with the population. This includes personal and non-personal health services, as well as the population's interaction with insurers and other administrative arms of the health system.

The focus on non-health aspects of the health system does not imply there is no causation between better responsiveness experiences and better health outcomes. The literature has shown an association between better health outcomes and better communication, for example (1, 2, 3, 4, and 5). At a conceptual level, the aim of having a separate responsiveness goal is to measure the increase in welfare derived from having health system processes, which are responsive, as the measurement of health outcomes will capture the improvement in health resulting from more responsiveness treatment or health system processes.

1.1.1. The domains

Through a literature review and the 1999 meeting of experts on patient satisfaction and quality of care, the following 8 domains were identified: dignity, autonomy, confidentiality, communication, prompt attention, access to social support networks during care, quality of basic amenities and choice. Over the past year and a half, the definitions have been subject to scrutiny by ethicists, professionals with training in international human rights and health professionals with various other types of formal training. The conclusions of all of this work are not yet finalized, but some of their findings will be presented in the current technical consultation. Consequently, the results presented in this paper are still in the process of development. This technical consultation will most likely cause them to change further.

The definitions of the domains are elaborated upon in Appendix 2, but they can be summarized as follows:

Dignity: The state of being worthy or esteemed. The aspects of dignity described in Appendix 2 have been chosen so as to preserve not only the esteem of others but self respect as well.

Autonomy: Having the right to be involved in making decisions regarding ones own health and treatment.

Confidentiality: Having total control over how your personal information is stored and to whom your personal information is communicated.

Communication: Having the right to have communication take place in a way that is clear and which respects the right of individuals to questions matters pertaining to their own health

Prompt Attention: Having your medical needs attended to within sufficiently short time intervals so as not to cause you undue discomfort or distress, in addition to the

detrimental affect of any delays in responses on the health problem for which you sought care originally.

Support for Social Integration or Social Consideration: Having health services give consideration to the impact of the health care process and the illness on the patients social network and access to social networks before, during and after care.

Quality of Basic Amenities: Having access to physical amenities which are of a generally acceptable standards, that is conducive to health (cleanliness) and reasonable levels of comfort.

Choice and Continuity: Being able to choose to continue care with a particular provider or institution or to change care from one provider or institution to another. The main criteria for identifying responsiveness domains are to ensure that they are parsimonious, yet comprehensive and cross-culturally appropriate. Techniques for aggregation of domain scores will need to be discussed under measurement issues.

1.1.2. The population

At a conceptual level, the population to which health systems are responding includes the *de facto* population of a country. As is to be expected, it has proved difficult to apply this gold standard of measurement in practice, and part of the purpose of assembling this group of technical experts is to address some of these challenges, to be discussed in more detail below.

1.1.3. Scope of the health system and legitimate expectations

The health system is defined as all acts whose primary intention is to improve health. In practice, the application of that definition becomes tricky when applied to certain traditional medical practices and, or, alternative medical practices. In different countries, different practices may be considered part of the health system. The best definition of the scope of the health system is those services and interventions that are recognized by the government of a country or internationally as contributing to the improvement of health. The international scope of interventions defined as contributing to the improvement of health may be different than the national ones. These would override the national definitions only when there is clear evidence that these activities contribute to the improvement of health.

1.1.4. Preferences

Conceptually, it has been hypothesized that it may be possible to order the domains according to those that are more and less important in different societies. Some work from the surveys, to be discussed below, has been conducted in attempt to measure people's preferences.

1.2. Principles of measuring responsiveness

The main objective is to measure **what happens** when a health system and the persons it serves interact. This objective seeks to report the behaviour, event or action of the health system. That is, measure what happens. This measurement is done from the perspective of the person the system is designed to serve. This self-report measurement like all self-report measurement, is confounded by the expectations and perceptions of the person reporting on the event. As a result, a particular measurement challenge is to assess the role of the respondent's **expectations and perceptions of the event**. Respondents are asked to report on and evaluate the health system, but not asked to indicate whether they are satisfied with the responsiveness of

the system. The purpose is to try to reduce the role of expectations, which are systematically built into satisfaction questions.

Given the conceptual underpinnings of responsiveness, as being the process of care experienced by the population, the best way to measure it is to ask people about their experiences. As we are wanting to create estimates of a country's health system responsiveness, the people asked need to be representative of the population in that country.

However, as an international organization concerned with building up an evidence base useful for policy making in health sectors in different countries, it is necessary to develop a comparable measure of health system responsiveness. To understand the challenge of cross-population comparability, we need to review the basic paradigm for measurement of health status in multiple dimensions. The problem with the comparability of results is formulated in terms of response category cut-point shift, sometimes called differential item functioning in the psychometric literature. The main part of the paper presents a series of alternative methods to address the problem of response category cut-point shift in order to enhance cross-population comparability of results.

1.2.1. An underlying latent variable and response category cut-point shift

The approach to constructing a cross-population measurement of responsiveness mirrors the approach used in constructing cross-comparability measures from health state self reports. For this reason, some reference will be made to experiences with health state self reports.

Most descriptions of the an individual's experience of responsiveness are based on the concept that responsiveness is multidimensional. Practical implementations of this concept require a limited number of domains of responsiveness be identified and measured in a reliable, valid, and cross-population comparable fashion. While there is certainly scope for debate on the domains that are part of the conception of responsiveness, we believe that there are a core set of domains that feature in nearly everyone's understanding of responsiveness of health systems. Some of the contentions around missing domains have already been discussed in the previous section and there will be time to discuss these issues in more detail in the consultation. The responsiveness domains include dignity, autonomy, confidentiality, communication, prompt attention, access to social support, quality of basic amenities and choice.

For the purposes of comparable measurement of individual and population responsiveness, we need to find reliable, valid, cross-population comparable, and cost-effective methods to measure this level of attainment. For financial reasons, the mainstay of measuring individual responsiveness experiences will necessarily depend on using self-reported information on key domains of responsiveness. Other measurements such as measured tests may be used to enhance cross-population comparability in sub-samples but the goal must be to find methods to increase the comparability of self-reported results.

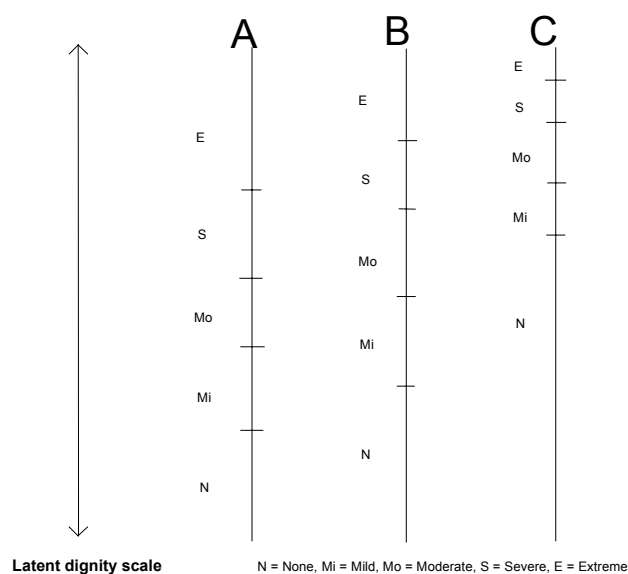
For self-reported questions to be used as an effective strategy to measure the level of a particular domain, the questions need to satisfy certain basic requirements.

Individuals need to interpret the meaning of questions in an identical fashion, retrieve all the relevant information with no memory loss, process the information to form

integrated judgements and convey the judgement in each survey identically. Cross-population comparability, however, requires even more.

Figure 1 illustrates the main challenge of using self-reported levels on a domain even when reliability and within population validity have been well established. For each domain, there is some true or latent scale for a domain. Imagine that *Figure 1* shows the latent scale for mobility. In a survey, consider a general question on dignity, such as “how was your experience of being treated with dignity at the health services in the last 12 months?” The response categories are “very good”, “good”, “moderate”, “bad”, and “very bad”.

Figure 1: Mapping from the latent variable to categories

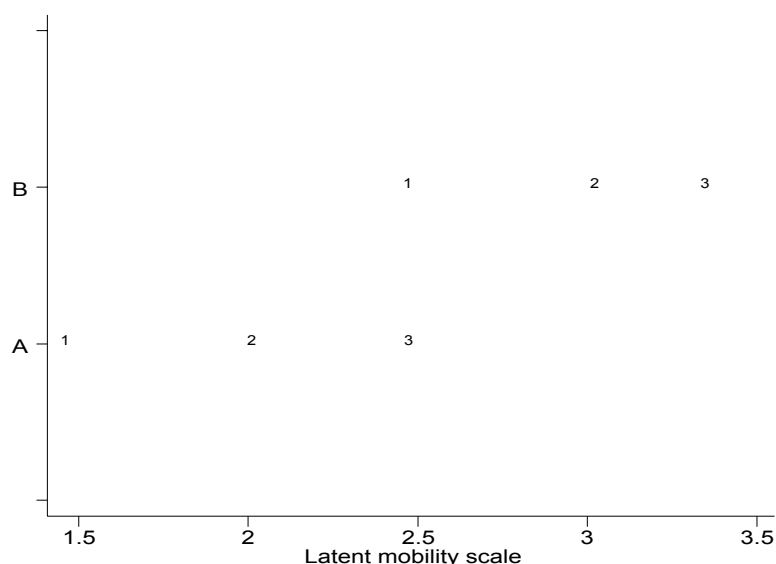


The second column in *Figure 1* shows for population A, the response category cut-points. These cut-points are the levels of dignity at which an individual will shift from one response category to another. The lowest cut-point on the figure shows the transition from answering "very good" to "good". In population B, the response category cut-points are shifted so that a higher level of dignity is associated with each of the response categories. A third example is also shown with even more shift in the cut-points. The implication is dramatic. A response of a good experience of being treated with dignity maps to a different level of responsiveness in population A as compared to B and C. In this example, survey results are reliable and valid within each population but the results cannot be compared across populations without adjustment.

We can hypothesize that cut-points may vary between populations for many reasons. These reasons might be due to different expectations held by different socio-economic groups. Another reason might be because of different cultural factors. Therefore cut-points is likely to vary within and across a cultural group. The cut-points for more educated individuals may shift as their expectations for a domain increases with increasing education. Recently, researchers working on improving comparability of patient satisfaction reports on health care plans in the US have proposed using “case-

mix” models adjusting for age, health status and education. Their proposal is to standardise satisfaction reports based on evidence of case-mix effects. Response category cut-point shift can make crude comparisons of results across populations nearly meaningless, even when exactly the same questions are used. How extensive a problem is it? Data from the US National Health and Nutrition Examination Survey III showed that even within one population there is substantial response category cut-point shift for different racial and socio-demographic groups. To turn to another health example, on the question of self-reported difficulty with regard to walking for a quarter of a mile (1=No difficulty, 2=Some difficulty, 3=Much difficulty, 4=Unable to do), the response of “much difficulty” in a black, poor, less educated, older female (individual A in *Figure 2*) was equivalent to the response of “no difficulty” for a white, rich, well-educated, younger male (individual B in *Figure 2*). In other terms, there was an estimated cut-point shift of almost two full categories.

Figure 2: Estimated cut-point shift from NHANES data



1.3. Overview of method of calculating the level of responsiveness

This section takes the survey results for a single domain and step by step walks through the process of calculating the domain score for a country. The domain we have chosen is outpatient dignity and the countries we chose to illustrate the approach are the Egypt and the UK.

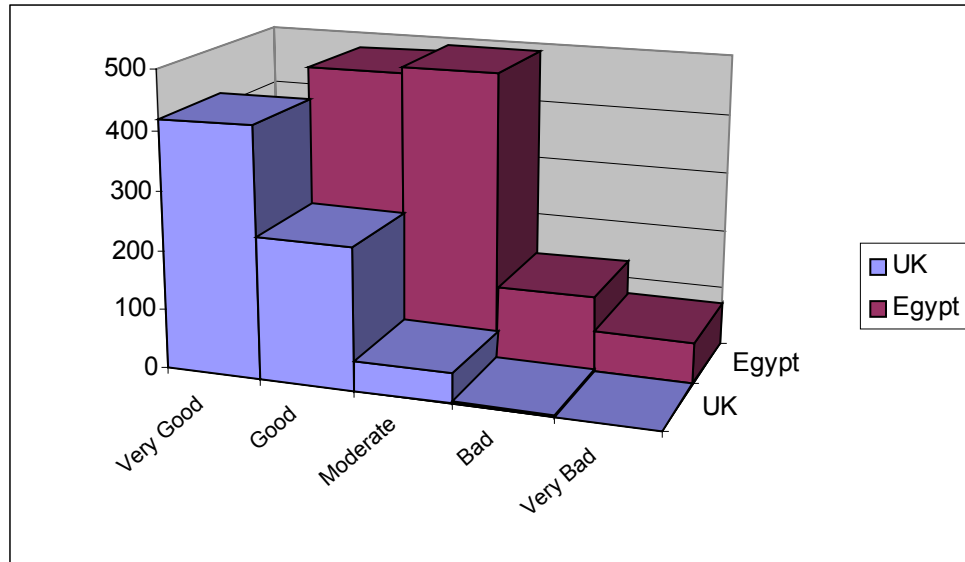
1.3.1. The raw data - self reports

We start off by looking at the raw data obtained from the survey. As the adjustment method uses the reports from vignettes to adjust the self-reports, we start off by reviewing the self-report and vignette report raw data.

The graph in Figure 3 shows the frequency distribution of answers to the question “How would you rate your experience of getting treated with dignity at the health

services in the last 12 months?” The response categories to the question were very good-good-moderate-bad-very bad.

Figure 3: Comparing frequencies in the raw data for Two countries



1.3.2. The raw data – the set of vignettes

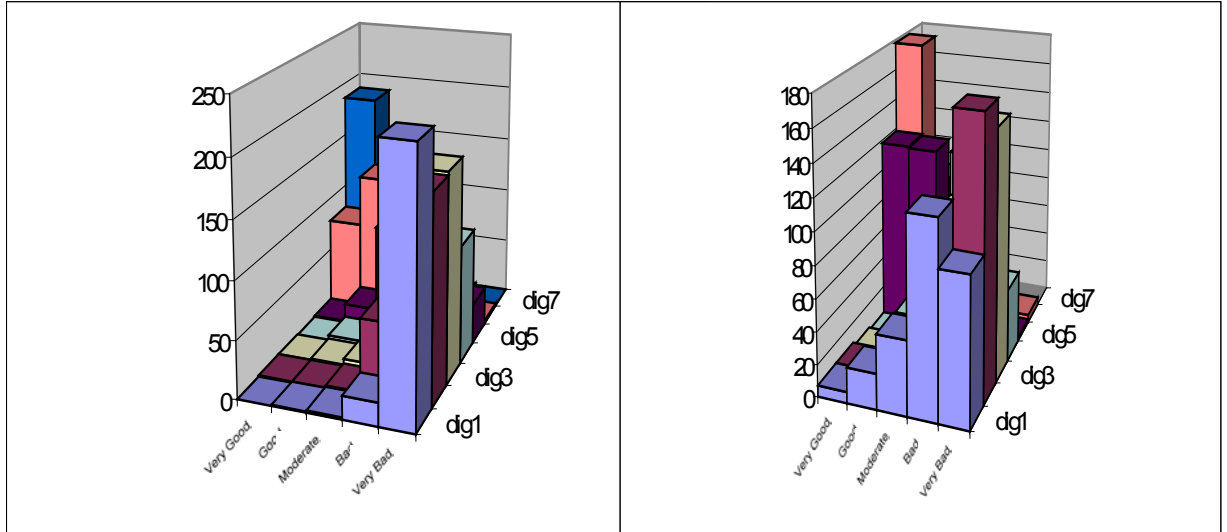
Figure 3 showed the frequency distribution of self-reports. However, while any two people from different backgrounds might have responded “good” to the same question, it is possible that they are not referring to the same level of the experience of the latent variable dignity. This is because it is possible that people use different categorical responses when referring to the same experience. Hopit adjusts for this shift in cut-points. Prior to running Hopit, we can review the raw data, namely, the categorical responses to the vignettes. Remember that for the vignette questions, respondents were asked to rate the experiences of the hypothetical person using the same phrasing and scale as that asked on the main domain self report question.

Figure 4 shows the frequency distribution of responses to the dignity set of vignettes for two countries. a group of countries. shows the same graph but this time for two different countries: China and India. These graphs show how the vignettes, while ranked in a similar fashion, are placed into different rating categories.

Figure 4: Comparisons of vignette responses in the raw data

(a) UK

(b) Egypt

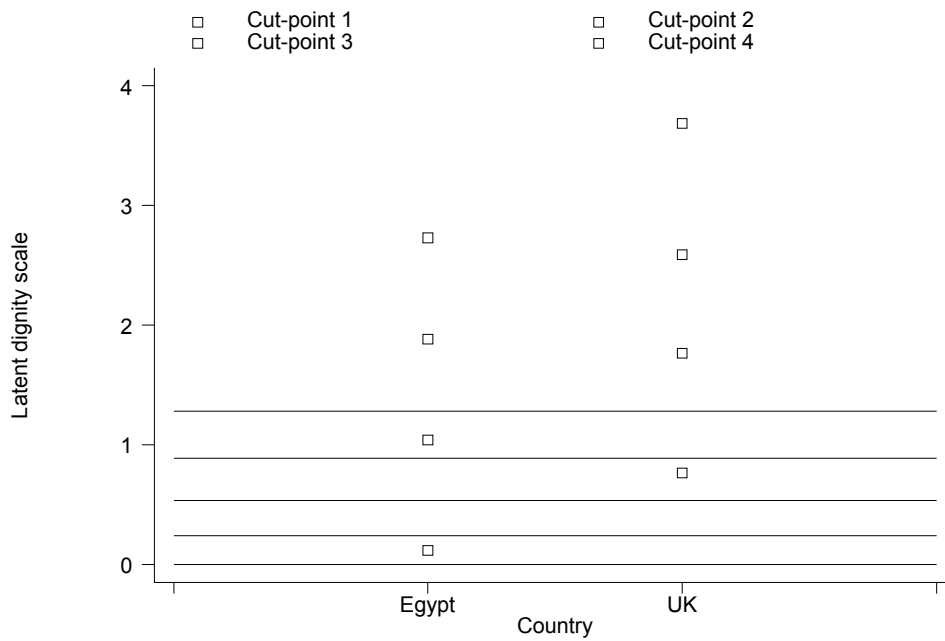


1.3.3. Hopiting the raw data

Once the hopit model is applied to the data, we have a set of results consisting of the cut-points on the latent variable for the different groups we have specified as having different cut-points. In this example, we have only specified the variable country as having an impact on the cut-points. In the full model we currently specify age, education and sex. Figure 5 shows a graph of the cut -points generated by Hopit for Egypt and the UK.

The cut-points between Egypt and the UK differ such that respondents in the UK tend to be harsher raters than respondents in Egypt. This is illustrated by the observation that every UK cut-point on the latent variable occurs higher up on the latent variable scale. The implication of a higher cut-point is that it would require a better experience of dignity (a higher level on the latent variable scale) to cause someone to shift their response to a more positive category, e.g., bad to moderate.

Figure 5: A comparison of the Cut-points for Egypt and the UK



The Hopit model also generate the output shown in Figure 6 . The first row of numbers shows the mean scores on the latent variable for each of the vignettes. This is followed by the mean of the self-report scores for the UK and Egypt. The rows headed “cut1” through to “cut4” refer to the values of the cut-points on the latent scale. In each case, as could be seen from Figure 6 the UK cut-points are higher than the Egyptian ones. To derive the cut-points for the UK, one needs to add the number to the base-case, Egyptian score. For cut-point 1, the UK cut-point value would be $0.12 + 0.65 = 0.77$.

Figure 6: Output from Hopit model regression

Number of obs	=	5407	Wald chi2(5)	=	1854.48
Log likelihood	=	-6067.6896	Prob > chi2	=	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
VIGNETTES					
_Ivignette_2	.239812	.0704762	3.40	0.001	.1016811 .3779428
_Ivignette_3	.5320174	.0692892	7.68	0.000	.3962131 .6678216
_Ivignette_4	.8872628	.0684542	12.96	0.000	.753095 1.021431
_Ivignette_5	1.278855	.0684994	18.67	0.000	1.144598 1.413111
_Ivignette_6	2.977407	.0777309	38.30	0.000	2.825057 3.129757
MEAN					
+UK	1.448953	.0830254	17.45	0.000	1.286226 1.611679
EGYPT	2.55911	.0751221	34.07	0.000	2.411874 2.706347
CUT1					
+UK	.6448705	.0485247	13.29	0.000	.5497638 .7399772
EGYPT	.1163824	.0551541	2.11	0.035	.0082823 .2244824
CUT2					
+UK	.7222357	.0518412	13.93	0.000	.6206288 .8238425
EGYPT	1.040691	.0566181	18.38	0.000	.9297213 1.15166
CUT3					
+UK	.7062425	.0607415	11.63	0.000	.5871913 .8252938
EGYPT	1.881387	.060802	30.94	0.000	1.762218 2.000557
CUT4					
+UK	.9540547	.0709587	13.45	0.000	.8149782 1.093131
EGYPT	2.727567	.0697237	39.12	0.000	2.590911 2.864223
s					
_cons	-.0297768	.0386339	-0.77	0.441	-.1054979 .0459443

1.3.4. Raw versus adjusted scores

Figure 7 shows the resultant score derived after making the cut-point adjustments. The latent variable scale was converted to a 0-100 scale by transforming all individual scores on the latent variable scale that were below the 2.5th percentile to 0 and all scores above the 2.5th percentile to 100. The distribution of self-report was also similarly rescaled.

Figure 7: Raw versus adjusted scores for outpatient dignity

Country	Score after adjustment	Raw score
Egypt	34.71	86.26
UK	74.32	88.34

1.3.5 Other methodological steps involved in the calculation of the measure for a country

Aggregation across types of services

The current approach to measurement identifies separate domain scores for two types of health system activities: outpatient and inpatient encounters. The current approach to the derivation of an overall domain score would be to average the two scores, without weighting them. The reasoning behind this approach is that to accord greater weight to either score would support the approach that one type of service is more desirable or valued than another and there is insufficient evidence to assess how populations value the different types of services.

Standardisation for cross-country comparisons

What implications does the utilization and availability of health services have for responsiveness conceptualisation and measurement? Would responsiveness be considered the same in two countries if those reporting on responsiveness gave the same reports but one country had a higher utilization rate than another, or one part of the population used the services more than another part. The proposal is that utilization rates do matter and that for comparability purposes, measures would need to be standardised for certain utilization rates of inpatient and outpatient health services across sex and age groups. Standardization with respect to non-personal services and other types of health system encounters would be more difficult as far less is known about the rate of utilization or exposure to these sorts of services. Other standardisation questions relate to the standardization of the responsiveness scores according to any other factors e.g., age-sex-education composition of the sample.

Overall responsiveness scores

An overall responsiveness score could be calculated by the summation of domain scores. Currently the thinking is to weight the domain scores according to mean preferences reported across survey countries. In this sense the weighting would be an empirical rather than a conceptual exercise. The only conceptual aspect of the weighting would be to hypothesize that some domains are considered more important than other domains and that this would be a generalisable result. Another alternative is to accord each domain an arbitrary weight from a distribution of weights.

Proposed inequality measures

Inequality measures can be classified into two main groups: those focused on inter-individual differences and those focused on individual-mean differences. The Gini coefficient is a measure of income inequality that falls into the inter-individual class of measures. The selection of an inequality measure for responsiveness has not yet been made but the general idea would be to assess the inequality of a global responsiveness score. Traditional inter-group comparisons for within country examination work, using the level of responsiveness pertaining to different groups in a society will be possible to the extent that the socio-demographic variables exist. Currently, the survey includes socio-demographic variables related to age, education, health insurance status, income and sex.

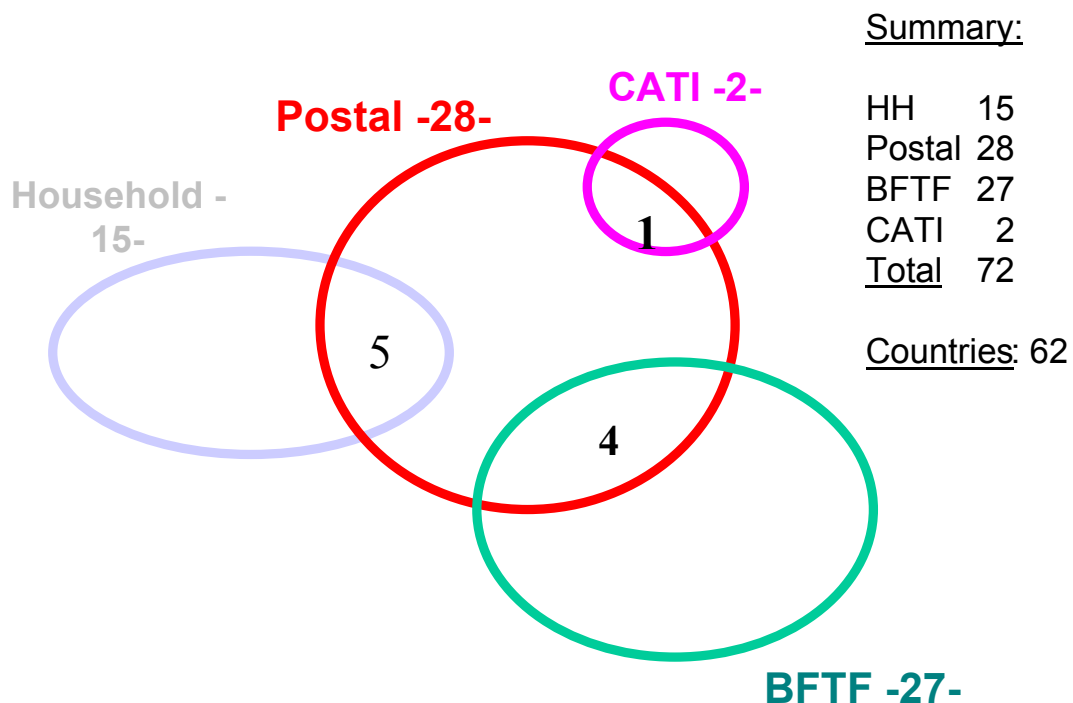
2. Overview of the Survey Program and Findings

2.1. Description of the survey program

The 2000 World Health Organization multi-country household survey program for responsiveness to-date has covered 117,960 respondents. Survey modes included face-to-face interviews (15 long and 27 brief), postal surveys (28) and telephone surveys. The survey program has been designed to test whether different survey modes yield comparable results.

The survey group is currently analysing the different cost implications of the different modes as well as the impact of different modes on sampling and representativity. Figure 8 shows the number of different survey modes, the total number of responsiveness surveys conducted in each region.

Figure 8: Overlapping survey design



2.2. The responsiveness questionnaire

2.2.1. The instruments in different household modes

The surveys all used a standardised questionnaire, with only minor differences between the long and brief household interviews. The inperson household and postal instruments generally consist of similar questions, there being a few additional questions in the long household questionnaire than are in the brief face-to-face or postal survey' (about 74 items, including general demographic questions, but excluding the household roster). The key informant survey consists of slightly more questions than those asked to the general population and the same items are occasionally worded differently in order to have the question make sense to "experts" rather than "patients" or "consumers".

2.2.2. Questionnaire structure

The first set of questions act as a filter to determine which respondents have had personal health care encounters in the previous 12 months. The subsequent self-report questions are grouped to correspond with different domains, first the outpatient section followed by the inpatient section.

Figure 9 sets out the main sections of the questionnaire, listing and comparing the number of items in each section in the household and postal questionnaires.

Figure 9: Overview of the questionnaire

Question Sections	Number of items	
	HOUSEHOLD	POSTAL
Population/Utilisation Filter	8	8
Outpatient Section	26	26
Inpatient Section	12	12
Questions - All Respondents	19	17
Other	5	2

The questions asked to all respondents include basic demographic questions, a question on which domains are most and least important, questions, a question on whether patients did not use health services for financial reasons and the vignettes.

The “vignettes” are set of hypothetical scenarios corresponding to different domains. There are 7 of these scenarios per domain. The main purpose of these scenarios is to elicit information about people’s cut-points. Using this information, it is possible to correct for differences in cut-point, which might have arisen for any number of reasons, including differing expectations and values. The resultant measure of responsiveness will thus be rendered comparable within and across countries.

In all surveys, vignettes for two sets of domains are rotated in order to keep the length of the questionnaire short. Therefore, although a respondent is asked for self-reports on all domains, they are only asked to evaluate the hypothetical scenarios for 2 randomly assigned domains. There are no fewer than 2 items per domain.

2.2.3. Response scales

All questionnaires shared common scales but scales varied between items within questionnaires and within domains. The two most common, categorical scales used were the never-sometimes-usually-always scale, and the very good-good-moderate-bad-very bad scales. The first type of scale is generally referred to in the patient satisfaction literature as a "report" of what happened, while the second very good-to-very bad 5-point scale is referred to as a "rating" scale.

3. Findings from Analyses of 50 Surveys

Results on the overall level of responsiveness have for 50 countries have been calculated using the approach described when walking through the Egypt and UK example. Domain scores were calculated for all the respondents to the survey, both for those having a visit and for those who did not have a visit. (Earlier in the section on measurement, we raised the point of alternatively calculating the index based only on the users of the services). Scores for the non-users were predicted according to a regression model based on age, education, sex, country and cut-point interactions. The domain scores were rescaled from 0-100 and standardised according to a standard age-sex population profile (50 % males 50 % females). These domain scores were then summed and weighted using the weights from last year, as an one approach to derive a final score for the level of responsiveness. As mentioned in the Section 1.3.5, equal weight was given to scores arising from the outpatient and inpatient health system encounters.

3.1. Overall scores

Overall scores for inpatient and outpatient are shown in Figure 10. The countries exhibit bunching on two sides of the inpatient scale. After examining the survey modes, survey companies and regions surveyed, there is no obvious explanation for the slight bunching tendency.

Figure 10: Inpatient and outpatient responsiveness scores

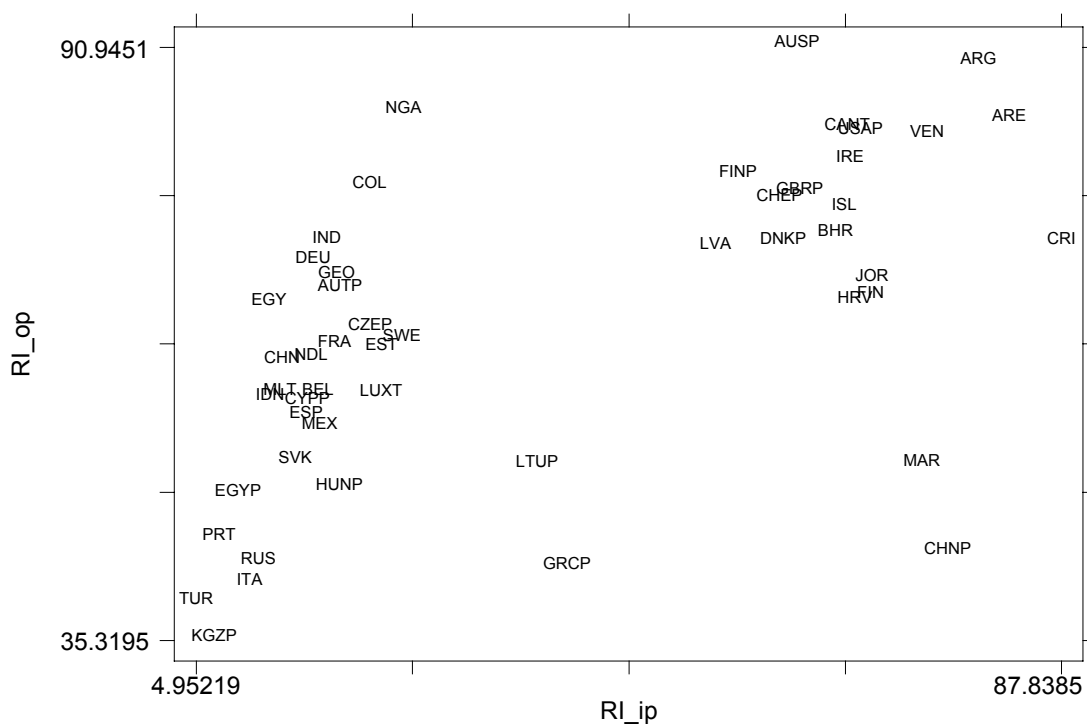


Figure 11: Overall responsiveness scores (n=50)

Country	IP-score	OP-score	Overall Score
AFRO			
Nigeria (h)	84.8	24.8	54.8
AMRO			
Argentina (b)	89.4	79.9	84.7
Canada (t)	83.1	67.3	75.2
Columbia (h)	77.8	21.6	49.7
Costa Rica (b)	72.5	87.8	80.2
Mexico (h)	55.2	16.7	36.0
USA (p)	82.8	68.6	75.7
Venezuela (b)	82.5	75.0	78.8
EMRO			
United Arab Emirates (b)	84.0	82.8	83.4
Cyprus (p)	57.5	15.7	36.6
Egypt (h)	66.8	11.9	39.4
Egypt(p)	48.9	9.0	28.9
Jordan (b)	69.0	69.7	69.3
Morocco (b)	51.6	74.5	63.0
EURO			
Austria (p)	68.1	18.8	43.5
Bahrain (b)	58.3	16.7	37.5
Belgium (b)	73.3	66.2	69.7
Switzerland (p)	76.5	60.8	68.7
Czech (p)	64.5	21.7	43.1
Germany (b)	70.7	16.2	43.4
Denmark (p)	72.5	61.2	66.8
Spain (b)	56.2	15.5	35.8
Estonia (b)	62.6	22.7	42.7
Finland (p)	67.4	69.6	68.5
Finland (b)	78.8	56.8	67.8
France (b)	62.9	18.3	40.6
Great Britain (p)	77.2	62.8	70.0
Georgia (h)	69.3	18.4	43.9
Greece (p)	42.0	40.5	41.3
Croatia (p)	67.0	68.1	67.6
Hungary (p)	49.4	18.7	34.1
Ireland (b)	80.2	67.6	73.9
Iceland (b)	75.6	67.1	71.4
Italy (b)	40.5	10.1	25.3
Kyrgyzstan (p)	35.3	6.7	21.0
Lithuania(p)	51.5	37.6	44.6
Luxembourg (t)	58.2	22.7	40.4
Latvia (p)	72.0	54.7	63.4
Malta (b)	58.4	13.0	35.7
Netherlands (b)	61.6	16.0	38.8
Portugal (b)	44.8	7.2	26.0
Russia (b)	42.5	10.9	26.7
Slovakia (p)	51.9	14.5	33.2
Sweden (b)	63.4	24.7	44.1
Turkey (h)	38.8	5.0	21.9
SEARO			
Indonesia (h)	57.9	12.0	35.0
India (h)	72.6	17.5	45.1
WPRO			
Australia (p)	90.9	62.5	76.7
China (h)	61.3	13.2	37.3
China (p)	43.4	77.0	60.2

Figure 11 shows the inpatient, outpatient and total scores arising from 50 surveys. Although, the group of surveys contained some overlapping modes, a comparison of modes has not yet been performed.

3.2. Domains showing the highest responsiveness, on average

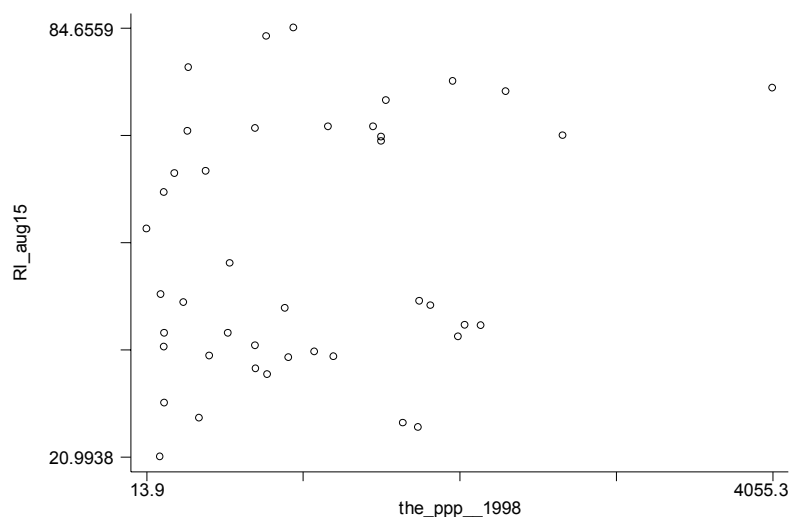
On average, health systems appear to exhibit the highest attainment in the responsiveness domain of outpatient choice of health care provider. Health systems on average perform less well in the domains of dignity and autonomy, and least well in prompt attention. With respect to inpatient domains, social support attains the highest score, on average, with prompt attention and confidentiality obtaining the lowest scores, across the populations, on average.

Figure 12: Average domain scores, across countries (n=50)

Domain	OP	Domain	IP
Choice	61.06	Social Support	44.05
Confidentiality	59.79	Dignity	42.99
Communication	58.08	QBA	41.50
Dignity	54.44	Choice	40.83
Autonomy	53.15	Communication	39.20
QBA	48.90	Autonomy	38.24
Prompt Attention	48.32	Confidentiality	37.25
Social Support		Prompt Attention	33.29

3.3 Responsiveness and health system expenditure

It is interesting to assess whether there appears to be any relationship between responsiveness scores and levels of health system expenditure. The scatterplot below does not seem to indicate that a very strong association exists, if at all. The correlation coefficient for a simple correlation between the responsiveness score and total health expenditure is about 0.4.

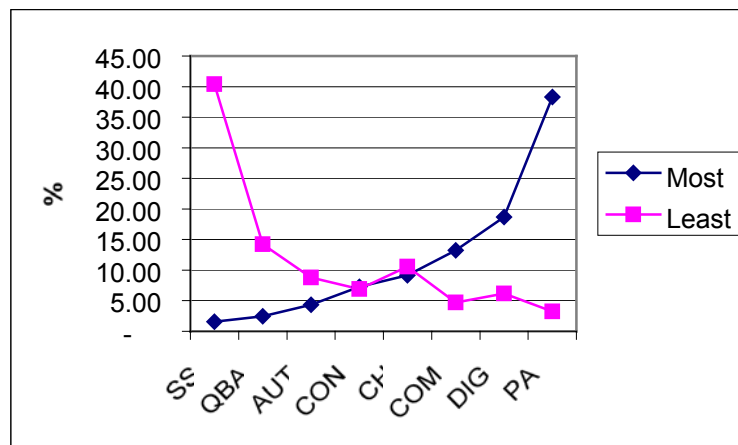


3.4 Importance of various domains

Respondents were asked to read descriptions of responsiveness domains provided to them and then to say which aspects of responsiveness they considered to be the most important.

Taking the answers they gave, and assigning a weight of 8 to domains ranked the most important and a weight of 1 to domains ranked the least important, it is possible to calculate a simple weight for each domain based on the number of times it was mentioned in first place, and according more points if it was. The resultant ranking seen in Figure 13 shows that of all the domains, prompt attention and dignity seem to be accorded the most importance, while social support and quality of basic amenities were accorded the lowest rankings, on average.

Figure 13: The importance of the domains (average results from the surveys)



3.5 Respondents reporting personal health care encounters in the last 12 months (averaged over 50 surveys)

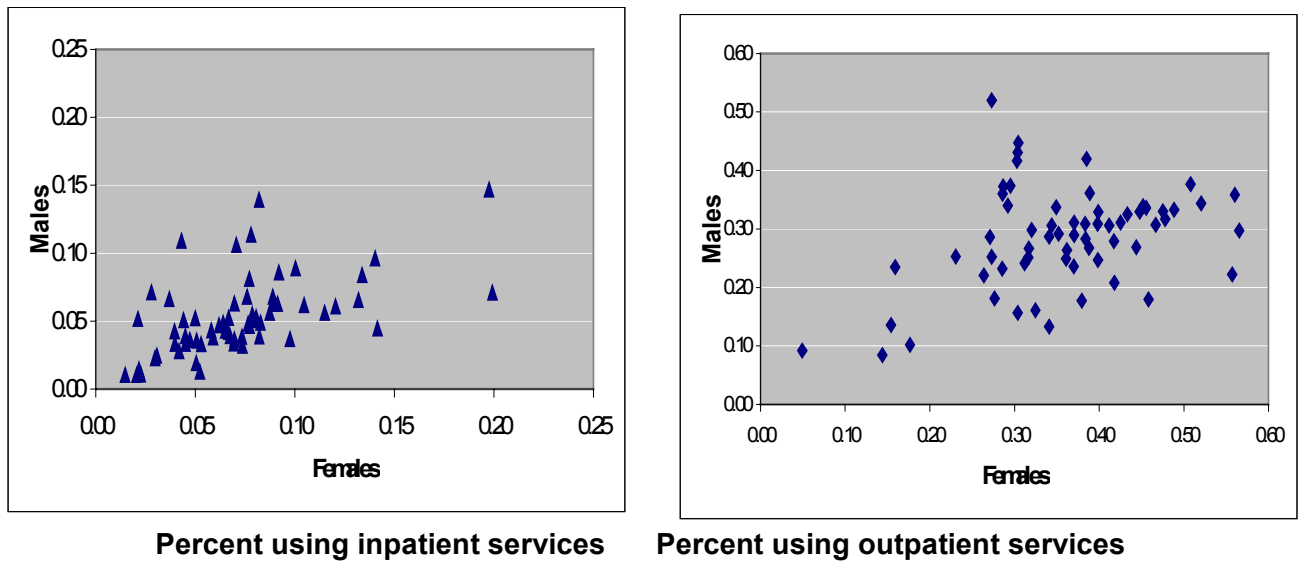
On average, as could be expected, outpatient encounters were much more common in the general populations surveyed (50.3 %) than inpatient encounters (11.7 %). Home care, only measured for a subset of countries running the long version of the household survey, was reported by a very small proportion of the populations (4 %).

Figure 14: Percentage of respondents on average

Population Filter and Vignette Results	
Percent Outpatient Visit	50.26
Percent Home Care Visit	3.97
Percent Inpatient Visit	11.70
Percent having test done	34.09
Percent of persons who received any health care	60.72

The above results apply to a subset of the 50 surveys initially reviewed. If the additional surveys, many of which are for European countries are added such that 65 surveys are included, the average outpatient encounter rate (the proportion reporting they have used an outpatient service at least once in the past 12 months) rises to 65 % and the inpatient rate to 13.6 %. On the whole, more respondents reporting having health system encounters were females than males, as can be seen in Figure 15.

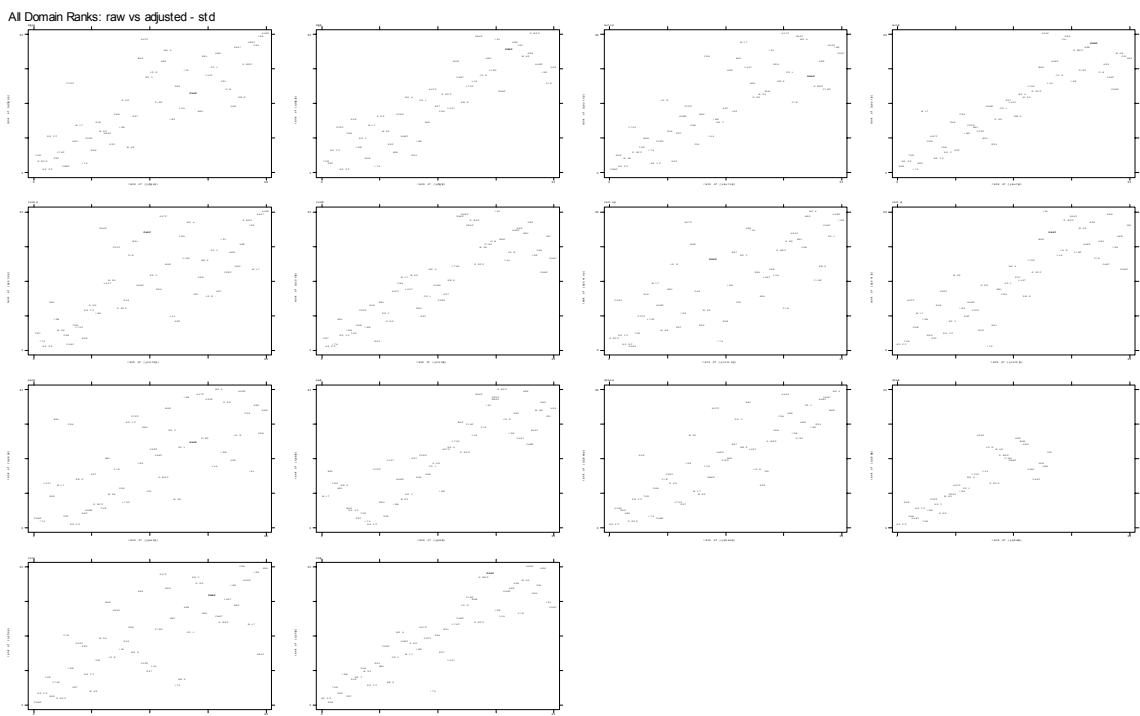
Figure 15: Comparison of utilization rates by sex



3.6 Results of adjustment on rankings

The Hopit adjustment would have a significant impact on the rankings of the countries, if the responsiveness scores calculated for the 50 countries had been used to rank. While the ranks of the adjusted scores are increasing functions of the raw scores, there is still variation in the ranks as a result of the cut-point adjustment. This is the type of result we would hope to see, otherwise the adjustment would not be large enough to make a difference.

Figure 16: Outpatient domains: raw versus adjusted effect on ranks

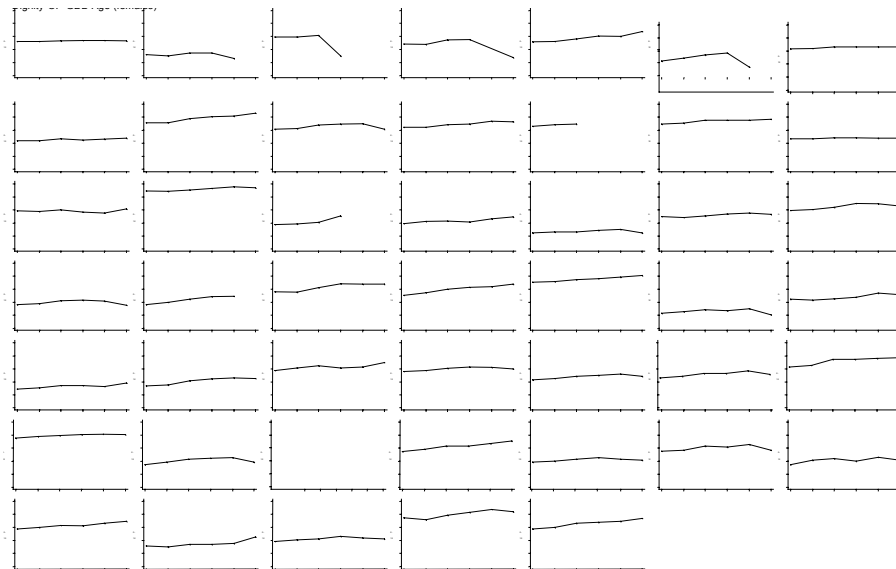


3.7 Scores by age groups

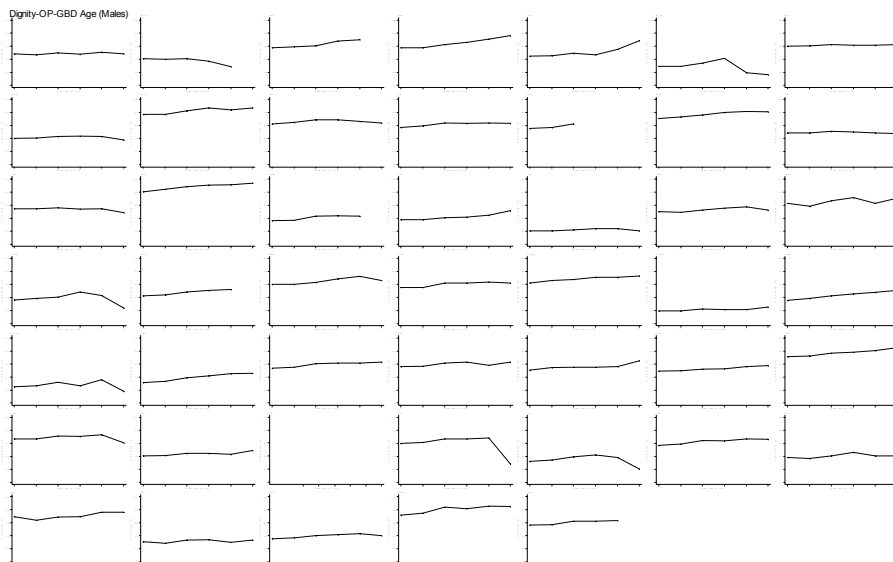
The pattern shown in Figure 17 for outpatient dignity scores across increasing age groups is typical of the pattern observed in the other domains. Overall, there seems to be a slight increase in responsiveness score. However, the variation around the mean at the higher age groups is larger, due to smaller sample sizes. More systematic work needs to be done to draw conclusions about which variables are associated with responsiveness at the individual level. The work presented here is merely descriptive and exploratory.

Figure 17: Responsiveness scores for outpatient dignity by age group

(a) Females



(b) Males



4. Revisiting Issues on WHO's Responsiveness Research Agenda

The research program consists of two main arms. There is the survey program, which began in 1999 and has culminated recently with the completion of 72 household surveys and more than 20 key informant surveys. (As discussed in the previous section, the purpose of running the key informant surveys is to assess whether their results are comparable in any systematic way with the household survey results.) The second arm of the research program is a program of qualitative work. Each of these are discussed in more detail as they relate to the different research agenda items below.

4.1. WHO research strategy aims

The responsiveness strategy document outlined a number of important goals and issues for improving the responsiveness conceptualization and measurement. The core were the following:

1. To provide a clear conceptual framework for responsiveness;
2. To measure responsiveness reliably and validly;
3. To keep the costs and burden of collecting data on responsiveness as low as possible; and
4. To link the measurement of responsiveness with its improvement.

The work outlined in this paper included a large survey program that included two main alternative survey modes, household surveys (conducted both inperson and by mail or telephone) and key informant surveys. Results from these surveys would then be compared, using pre-assigned overlap mode countries to assess the cheapest most reliable alternative for measuring responsiveness. The survey program that grew out of this research strategy extended to cover a number of short face-to-face interview as well as telephone interviews where these offered viable alternatives cost-wise. Subsequent work on the responsiveness research agenda led to the development of a series of research questions and priorities, which are listed below. These agenda items will form the basis for the final discussion around future research agenda priorities. Many of these items can be addressed using the data from the 2000-2001 round of surveys. Where data have been analysed to address these items, these will be discussed in the next section of the paper.

4.2. Qualitative research

Further to the surveys described above, qualitative work on the domains of responsiveness is underway. A focus group instrument has been developed and is in the process of being run in 3 countries. It is hoped that there will be funding to extend this work to a number of other countries. In the focus group, participants are asked to describe encounters with health care providers. The aim of this exercise is twofold:

(1) to assess the validity of the current list of domains, and
(2) to elicit descriptions of encounters using words and phrases provided by participants. These will be used to modify and build up question items.

Other work on responsiveness domains includes the commissioning of ethicists to write short commentaries on various domains, in particular with reference to cross-cultural applicability and understanding. A senior ethicist has been requested to synthesize the various papers into a longer paper and presentation, to be made at the technical consultation.

The facility surveys mentioned briefly above are going to be piloted in 3 countries: China, Egypt and Indonesia. The instrument is currently under review by a WHO health services working group. Prior to running the survey, it will be tested in a few sites in each country.

The testing of vignettes is an exercise that will be underway within the next few weeks. We will have some discussion regarding the proposed method for testing as part of the technical consultation's group sessions.

4.3. Specific research questions and answers

Over the course of the research on responsiveness, a number of issues have been raised, both within WHO and within the broader research and political community. Some of the main issues needing to be addressed are highlighted below. It will not be possible to address all of these at the Responsiveness Consultation due to time constraints. In addition, as the household survey is considered to be the gold standard by which responsiveness measurement should take place, you will see in the Consultation agenda that most of the sessions are geared towards improving this instrument. Nevertheless it is important to be aware of all the various research questions we are faced with in order to participate fully in the discussion in the final session on responsiveness research priorities. The paragraphs below elaborate on the different research agenda items and suggested approaches to addressing these.. In some cases analyses have already been carried out and results are reported. In other cases work is in progress or has not yet begun. The general approach has been to highlight those research items that seemed to be most urgent.. Delegates will be asked to help provide suggestions on how to approach the research items and to reassess research priorities as the last agenda item of the Responsiveness Technical Consultation.

4.3.1 Domains and terminology

The main questions under this heading centre around domain comprehensiveness, definitions of domains and general terminology surrounding our understanding of responsiveness (e.g., legitimate expectations, perceptions, welfare/well-being). A number of other domains have been proposed, the most prominent of which are domains for:

- caring;
- a domain encompassing the notion of “call to rescue”: people value knowing that health systems will do all they can for them or their loved ones in times of need; and

- a domain encompassing knowing that environmental procedures are in place to ensure that air quality is maintained, the integrity of food is maintained, there are safety procedures at work, at school at college and in the transport industries.
- With respect to the proposal to include a caring domain, it is possible that the current domains of dignity, communication autonomy, and confidentiality may address it. Caring has been equated in other studies with interpersonal aspects of care which include these four elements of responsiveness.
- The other proposed domains are more difficult to accommodate and might necessitate the revision or addition of domains.

Methods for addressing the research question on domains

Experts in the fields of ethics and human rights have been commissioned to review the comprehensiveness of the domains based on ethical theory, while bearing in mind a cross-cultural perspective. The results of this work will be presented at the Consultation. The domains are also being investigated through a series of focus groups aimed at getting participants description of the encounters they have had with health care providers. It is also hoped that a better understanding of some of the general terminology on responsiveness will emerge as a result of discussion at this Technical Consultation. Once detailed descriptions of the domains and mappings to human rights and ethics literature have been formulated, formal definitions will be developed.

One of the questions relating to overlapping domains can be partially answered by reviewing the correlation matrix below. The first point to note is that the outpatient and inpatient scores are not highly correlated. Within outpatient domains, there seems to be a fairly high correlation between dignity, autonomy, confidentiality and communication.

Figure 18: Correlation of the country level domain scores

	dig_y_op	dig_y_ip	aut_y_op	aut_y_ip	con_y_op	con_y_ip	com_y_op	com_y_ip	pa_y_op	pa_y_ip	qba_y_op	qba_y_ip	ss_y_ip	ch_y_op
dig_y_op	1.00													
dig_y_ip	0.53	1.00												
aut_y_op	0.81	0.20	1.00											
aut_y_ip	0.51	0.95	0.20	1.00										
con_y_op	0.71	0.49	0.63	0.50	1.00									
con_y_ip	0.49	0.92	0.13	0.95	0.57	1.00								
com_y_op	0.90	0.40	0.87	0.40	0.66	0.37	1.00							
com_y_ip	0.53	0.98	0.20	0.98	0.52	0.95	0.42	1.00						
pa_y_op	0.68	0.38	0.63	0.29	0.56	0.29	0.69	0.34	1.00					
pa_y_ip	0.49	0.96	0.16	0.92	0.51	0.93	0.37	0.96	0.42	1.00				
qba_y_op	0.58	0.49	0.41	0.44	0.39	0.45	0.60	0.50	0.39	0.46	1.00			
qba_y_ip	0.50	0.95	0.23	0.96	0.54	0.92	0.41	0.97	0.35	0.94	0.51	1.00		
ss_y_ip	0.40	0.94	0.09	0.91	0.45	0.87	0.28	0.93	0.39	0.93	0.41	0.90	1.00	
ch_y_op	0.78	0.26	0.77	0.22	0.65	0.18	0.74	0.23	0.78	0.24	0.21	0.22	0.23	1.00
ch_y_ip	0.56	0.95	0.24	0.94	0.58	0.92	0.42	0.94	0.43	0.93	0.37	0.93	0.91	0.39

4.3.2 Health system activity

The questionnaire focuses on personal health care services. The goal of the initiative is to include nonpersonal services as well. The questions in this area did not perform

well. Thus assistance is sought in the consultation for how to better report on these services.

1. In the household surveys, the responsiveness of outpatient and inpatient personal services are evaluated for each domain (home services are also included in the long version). Should other services be included (bearing in mind the objective of the measure of giving an overall score for the whole health system).
2. Should distinct measures be developed for government and non-government services – how do you do this? Should we leave this for a later stage? What are the challenges in doing this, for example are respondents able to distinguish between services received from government and nongovernment sources
3. Should we include an evaluation of other aspects of health services, e.g., non-personal services e.g. anti-smoking campaigns, administration services.

Method for improving the coverage of health system activities

It is hoped that the Technical Consultation will address the issues listed above, as well as the extent to which the evaluation of non-personal health services and other types of health system encounters can be made evaluated under existing domains or whether additional domains will need to be added.

4.3.3 Operationalizing domains

The main questions and issues around the domain item pools are the following:

1. Which questions are most valid across cultures?
2. Which questions are most reliable?
3. How might the number of items be reduced?
4. How might the equivalence and appropriateness of translations into different languages be improved.
5. Expand list of items, for example, to include reports on the experiences of the population with non-personal health services and with other health system actors.

Methods for improving the operationalization of domains

A series of cognitive tests were run on the item pool. Figure 19 shows the number of cognitive interviews and the countries in which these cognitive interviews were conducted. The cognitive interviews took the form of personal debriefings of the respondents, after they had completed the questionnaire.

Figure 19: Cognitive interview testing in 2000

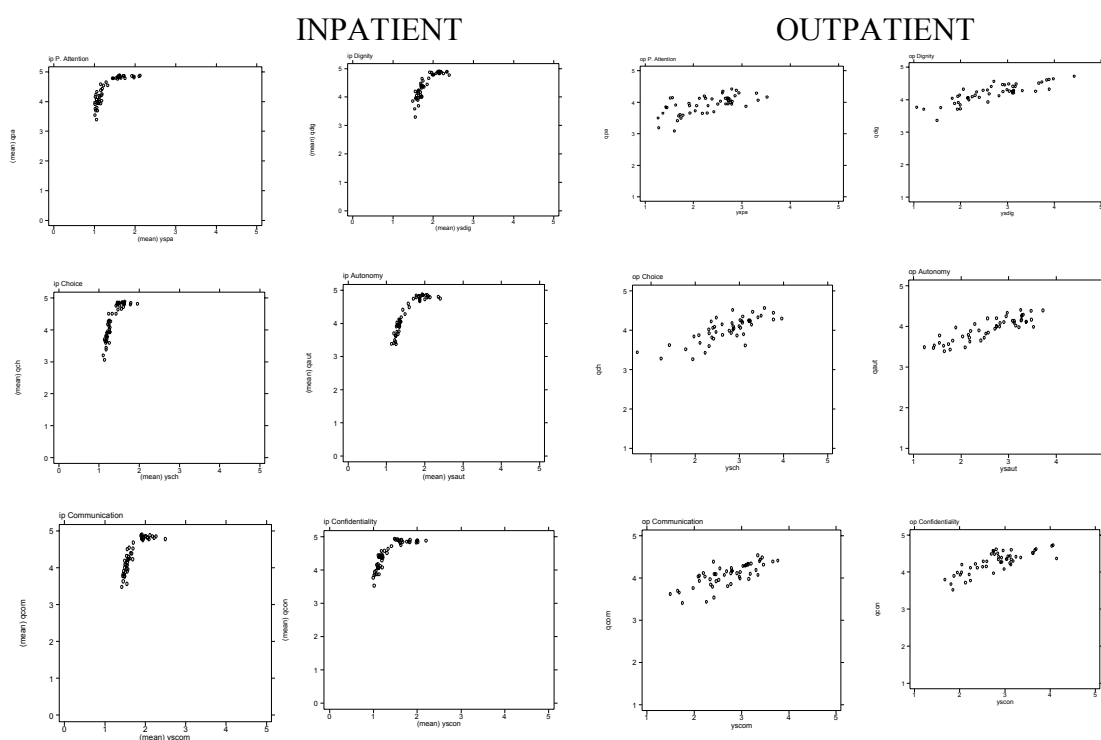
Country	Number of Interviews
China	38
Egypt	30
Georgia	30
India	23
Indonesia	5
Nigeria	20
Slovakia	33
Total	179

Over the next few months, cognitive testing will be substantially expanded. The aim is to conduct between 150 and 200 cognitive debriefing interviews in at least 8 different countries. In addition to this a few smaller research projects focused on particular domains are under discussion with institutions interested in responsiveness research. Three sets of focus groups are currently being run in 3 different countries. In addition to deriving information about the comprehensiveness of domains, the qualitative research aims at developing a pool of items built up from people's discussions of health service encounters. This work builds on similar studies in the field .

Finally, it is hoped that the results of the first set of group work in the Technical Consultation will lead towards a better idea of how to select the appropriate items. During the consultation, delegates will be handed information on Kappa scores and variables with high missing. There will be group work focusing on reviewing the validity and reliability of different items, and approaches for improving the quality of the survey instrument.

From the survey results, one observes that the inpatient items are behaving in such a way as to create a ceiling effect. This point is illustrated by the graphs below. These graphs show plots of adjusted versus raw scores for several countries' inpatient domain scores. The mean raw score is plotted on the y-axis and the mean latent score is plotted on the x-axis. From the shape of the scatter diagram, it is possible to see that the measurement using the raw score results in a bunching at the top end of the raw scale. The application of the hopit model to the raw data enables the one to discern between differing levels for the countries on the true latent scale.

Figure 20: Raw versus adjusted scores for 6 domains (n=50)



4.3.4 Response categories

1. Simplification of the response categories across different domains if possible.
2. Use of a never-to-always 4-point scale versus the very bad-to-very good 5-point scale. What are the advantages of one or the other? To answer this we need to discuss the number of response options on the scale, as well as the nature of the scales as report vs. rating items.

The current questionnaires use a variety of scales, described in the section on surveys. To improve the ease of completion of the questionnaire, a further attempt will be made to word the question items in such a way as to accord similar response scales to all question items. The literature on distinguished reports scales from rating scales. While report scales, such as the never-to-always scale are considered harder to translate, they are thought of as less affected by patient expectations, or values.

Methods for improving response category uniformity and appropriateness

In the drafting of the questionnaires fielded in the 2000-2001 surveys, much attention was paid to the standardization of the response scale but in the end there were nevertheless 3 different scales used for different domain items. The no problem-extreme problem scale, the never-to-always scale and the very bad-to-very good scale. With more tie to reflect on the questions, it is hoped that it will be possible to reduce

the variance in response scales and to select the most effective scale for providing the reports of experiences of respondents. A number of issues must be considered including the cross-cultural comparability, translatability, and performance of the response scales.

The inclusion of at least two different scales in the 2000-2001 survey questionnaire was intentional as there was much debate in the literature about the ability of never-to-always scales to be less expectations driven than rating scales, like a 0 to 10 rating scale or a very bad-to-very good rating scale.

In the responsiveness survey an additional attempt to capture expectations and differing values was made by using the hypothetical scenarios and the modelling technique contained in hopit. In this paradigm, the scale of the latent variable is fixed by the vignettes. The main issue becomes to resolve how the cut-points of different items and their response categories map to the latent scale and distribution of the latent variable. Some of the discussion in the Technical Consultation will focus on a comparison of the use of the different scales, but it is important to see the effect that scale differences might have in terms of cut-point shifts. Some discussion time at the Consultation will be allocated to the discussion of this topic.

4.3.5 Users and survey methods

1. Should the defined population be changed in any ways? (currently it is people who have used health services (any) in the last 12 months, who are over 18 years of age) For example, should we include proxy reporting by people who attended with relatives or friends.
2. What about institutionalised population (e.g. prisons, hospitals, old age homes)?
3. How do we include people who have not used health services in the last 12 months but who have had some other type of health system encounter e.g., environmental health?
4. How do we take into account people who wanted to use health care services but did not, for whatever reason?

Methods for improving the definition and sampling of the population

Once again, on this subject, we are hoping that the Technical Consultation will give us some direction as what users to include. Current household face-to-face survey techniques randomly identify respondents in the household to answer the questionnaire. What this means for responsiveness is that if the respondent is only asked about personal health care services s/he might have used in the past 12 months, the sample of people reporting on their experience of responsiveness in health services is much less than the original sample drawn.

Both from a pragmatic level and on a conceptual level, there has been much support for including reporting, on health services for children. Parents or other person caring for a child would provide their perspective on the health services for the child. This would not be a proxy response for the child, but rather the adults perspective on the health services for the child.

Other comments have focused on institutionalised populations and it is here that the experts to the Technical Consultation need to provide guidance. Implications of the operations of surveys when including institutionalised populations are significant. There are significant issues of cognitive ability of respondents in some institutions would be a cause of concern (e.g. homes for the mentally handicapped or facilities where many elderly persons reside who have cognitive impairments.) Proxy reporting might need to be used with families of these persons, but should probably be viewed as representing their views rather than as proxies.

The 2000-2001 surveys included a question on the non-use of services due to financial barriers. A number of people indicated that financial barriers had prevented them from using health services. In this case, how should the responsiveness indicator take into account non-use of health services, and are there some reasons for non-use that should result in lower responsiveness scores than others.

4.3.6 Experience reference period

1. Currently, a 12 months recall is used and respondents are asked about all services in the last 12 months. Is this recall period satisfactory? Is the frame of which services to include adequate?
2. Instead of being asked to think of all visits in the last year, respondents could be asked about the last visit or the usual place of care. Which would be the best reference?

The major criticism of the introductory questions has been that they might exclude some types of health treatment (e.g. visit by midwife) in favour of visits to a subset of recognized facilities. This problem can be redressed by having a more open set of introductory questions, which start at the point of asking about what happened the last time someone was ill. It might be possible to allow the person to plot a pathway towards talking about the set of experiences they wish to comment on, whether it was their last visit or their usual place of care. The time reference might also be a variable that could be recorded, but not used to exclude patient reports. The problem with extending the time period beyond 1 year means that time would be taken to collect the data for previous years but these data would need to be reported on separately for different years of responsiveness assessment. On the other hand, if surveys are conducted every few years, it might prove a valuable way of collecting time series data.

Experts in the consultation will be asked to consider an approach in which commonly recognized sets of health care services. These may vary by country. What are the implications of this for the survey. Also experts will be asked to consider the reference period. What implications would use of last visit have on the survey data. Can one get a representative sample of visit types this way. What is the impact if respondents are using very different time periods to include the last visit. If the last visit is more than say a year what is the impact on recall. In terms of health system accountability what are the implications of this approach.

Methods for improving the frame of reference

Currently, there are two sources to draw upon for improving the experience reference. First, the experience of the 2000-2001 round of surveys showed that the current arrangement of skips has worked, though not perfectly, as evidenced by the slight discrepancies between different questions in the pattern of skips. Second, in this Technical meeting we will be looking for guidance on developing different options as part of one of the group work sessions. Finally, any best alternative introductory questions will be tested in cognitive interviews.

4.3.7 Vignettes and the scale

1. Improve vignettes (are they triggering the correct responses from people? are people rating vignettes in the same way they are their own personal experiences? what tests can we use to ensure that cross-cultural adjustments are being made? – i.e. the vignettes are working (we currently have some suggestions but need more))
2. Which situations described in vignettes are most relevant across different countries and settings?
3. Describe what the latent scale means to policy makers by reference to the worst vignette in each domain.

Methods for improving the vignettes

Results from the 2000-2001 surveys have once again proved invaluable in the development of techniques for adjusting for expectations. These results will be discussed on the relevant survey section. Vignettes have been reviewed according to the extent to which people within the same population have similar understandings of the vignettes, as well as with respect to the extent to which vignettes capture the full range of experiences possible on the latent scale.

Figure 21 shows the average Spearman's rank correlation coefficient for the ranking of the vignettes for the set of dignity, across the surveys. The calculation is based on taking a correlation score for each individuals rankings of the vignettes compared with the mean ranking within a particular country. The closer the mean of this distribution of scores is to 1, the greater is the proportion of people in the population who agree on the ranking of the vignettes. This, indirectly, is a test of how the vignettes are people perceived by the populations across different countries. In Figure 21 the vignettes in Turkey, Hungary and Indonesia look particularly problematic. There might be problems with the translations of the vignettes or the relevance of the contexts described in the vignettes.

Figure 21: Rank correlation coefficients across countries for the set of dignity vignettes

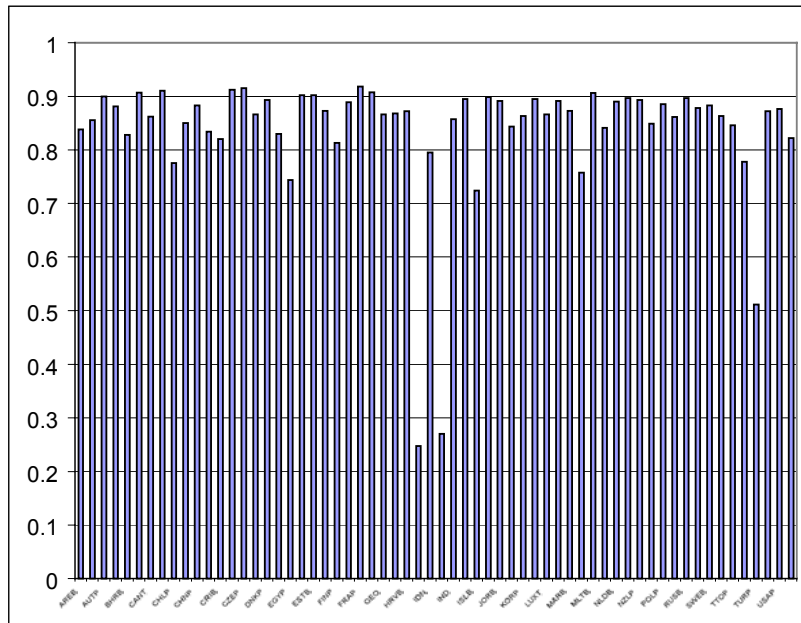
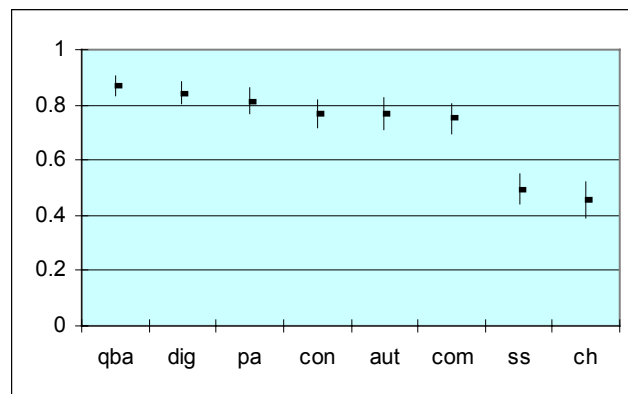


Figure 22 shows that the best functioning sets of vignettes are quality of basic amenities, dignity and prompt attention. The averages of the mean Spearman's correlation coefficients were taken for each domain, across countries. The domains where the vignettes are performing the worse are for choice and social support (or what we might potentially refer to as the concept of social integration).

Figure 22: Mean correlation scores for different sets of vignettes (n=65)



Vignettes will form a major part of the cognitive testing to take place over the next two months. How to approach this testing will be discussed in the Technical Consultation.

4.3.8 Biases in different modes

As one of the key objectives of the WHO research plan is to provide the least costly ways of collecting valid information on responsiveness, various survey modes with different costs are being assessed (The draft survey document reviews the various cost alternatives and some aspects of mode effect) . Accordingly some of the key questions to be answered are:

1. How biased are key informant surveys compared with household surveys?
2. How do biases in self-administered and face-to-face modes (at home and at facilities) compare?

Methods for reviewing mode bias

The literature shows that self-administered questionnaire on patient satisfaction tend to yield more negative ratings than face-to-face administered questionnaires (ref.). This hypothesis will be tested a review of the overlap countries forming part of the 2000-2001 survey program. In addition, a study comparing the key informant survey results with the household surveys will be informative in determining the extent to which key informant surveys act as least cost alternatives to household surveys.

4.3.9 External calibrators and validity checks

1. What techniques can be used to validate and calibrate the household self-reports?

While the vignettes offer some means of external calibration, another form of calibration could take the form of a facility survey, where information on responsiveness is gathered through observation.

Methods for determining external calibrators and validity

External validity checks will take the form of comparisons of certain of the data collected in the current round of surveys, e.g., utilization rates, with other known series covering the same variables.

External calibration and validity checks will also be developed through a survey program aimed at reviewing responsiveness conditions at facilities through observation and comparing this with patient reports on responsiveness at the same facilities.

A facility survey instrument is in the process of being developed. The survey instrument consists of four components: an interview with senior management, staff interviews, observation of the facility and certain process aspects (e.g. waiting time) and exit interviews with patients.

4.3.10 Linkages to functions (or determinants of better responsiveness performance)

How can responsiveness work be made more relevant to policy makers?

In order for responsiveness assessment to be relevant to policy-makers it is necessary to understand what factors lead to better responsiveness.

Methods for improving understanding of linkages between responsiveness and functions

The facility survey instrument mentioned above will be used to assess the association between different service characteristics such financing arrangements and staff working conditions and greater responsiveness.

5. Closing Note

A number of research issues have been highlighted in this paper. While the time afforded the consultation process will not permit all the issues touched on in this paper to be explored in depth, hopefully the broad outline of the research issues will stimulate fruitful discussion on many of the most serious conceptual and methodological questions.

The consultation closes with a session to discuss how to take the WHO responsiveness research agenda forward so this will provide a structure for future work and hopefully make up for some of the issues not covered during the 2 day consultation. At this session, delegates are invited to suggest research agenda items and the prioritization of different research questions, as well as to suggest methods for conducting the research.

Extensive notes from the Technical Consultation will be written up after the meeting and distributed to the Consultation delegates. Anyone wishing to contribute to ongoing debates and work on improving this field of study will be welcome to continue interaction on the subject.

Bibliography

- Aharony L, Strasser S.** Patient satisfaction: what we know about and we still need to explore. *Medical Care Review* 50: 49-79, 1993.
- Darby C, Valentine N, Murray C & de Silva A.** *WHO Strategy Paper*. Geneva, World Health Organization, 2000 (Global Programme on Evidence for Health Policy Discussion Paper No. 23)
- De Silva A, Murray CJL. 2000.** A Framework for Measuring Responsiveness. (Global Programme on Evidence for Health Policy Discussion Paper No. 32).
- Gilson L, Alilio M, Heggenhougen K.** Community satisfaction with primary health care services: an evaluation undertaken in the Morogoro region of Tanzania. *Social Science and Medicine* 39: 767-780, 1994.
- Iburg, K. M., Salomon, J., Murray, C. J. L., & Tandon, A.** 2000, "The measurement and interpretation of health in social surveys," in *Summary measures of population health*, C. J. L. Murray et al., eds..
- McPake B.** User charges for health services in developing countries: a review of the economic literature. *Social and Science and Medicine* 36: 1397-1405, 1993.
- Murray CJL & Frenk J.** A WHO Framework for Health System Performance Assessment. Geneva, World Health Organization, 2000 (Global Programme on Evidence for Health Policy Discussion Paper No. 6)
- Streiner, D. L. & Norman, G. R.** 1995, Health measurement scales: a practical guide to their development and use, Second edn, Oxford University Press, Oxford.
- Ware JE, Snyder MK, Wright WR, Davies AR.** Defining and measuring patient satisfaction with medical care. *Evaluation and Program Planning* 6: 247-263, 1983.
- Wouters A.** Essential national health research in developing countries: health care financing and the quality of care. *International Journal of Health Planning and Management* 6: 253-271, 1991.
- Zaslavsky A.** *Issues in Case-Mix Adjustment of Measures of the Quality of Health Plans*. Unpublished manuscript.

APPENDIX 1

Identifying the Components of Individuals' Satisfaction: Review of the Literature

Aspect	Studies
Dignity	Ali and Mahmoud (1993) Avis, Bond and Arthur (1997) Bassett, Bijlmakers and Sanders (1997) Cleary et al (1991) Collins (1996) Etter et al (1996) Grol et al (1999) Gross et al (1998) Kenagy, Berwick and Shore (1999) Lim, Tan, Goh and Ling (1998) Morris (1997) Rurnbull and Hembree (1996) Rylance (1999) Wensing et al (1998)
Autonomy	Avis, Bond and Arthur (1997) Charles, Gafni and Whelan (1997) Cleary et al (1991) Coulter, Entwistle and Gilbert(1999) Meredith et al (1993)
Confidentiality	Grol et al (1999) Rylance (1999) Denley and Smith (1999)
Prompt attention	Collins (1996) Etter et al (1996) Grol et al (1999) Lim, Tan, Goh and Ling (1998) McIver (1991) Pascoe and Attkisson (1983) Ware et al (1983)
Access to social support networks during care	Cleary et al (1991)
Quality of basic amenities	Abramowitz et al (1987) Baker (1991) Collins (1996) McIver (1991) Ware et al (1983) Minnick et al (1997)
Choice of Care provider	Collins (1996) Campbell (1994) Hall et al (1994) Meredith et al (1993)

APPENDIX 2

Elaboration on the Understanding of the Responsiveness Domains

1. Dignity

- Safeguarding human rights of patients with communicable diseases
- Respecting reproductive rights of women
- Treating with respect in health care settings
- Treating with care in health care settings
- Safeguarding(physical) privacy of persons during treatment and examination

2. Autonomy

- Right to make decisions regarding type of treatment in consultation with the health care provider
- “Informed consent” given before testing, surgery or research undertaken
- Right to refuse treatment

3. Confidentiality of Information

- Consultations conducted in a manner that safeguards privacy
- Confidentiality of information provided by the patient (‘privileged communication’)
- Confidentiality of medical reports

4. Communication

- Individuals should be able to communicate freely about their disease with their health care provider
- Individuals should be able to ask health care providers questions freely about their diseases, symptoms, treatment and care
- Patients should be encouraged to be involved in deciding about their treatment or care, if they wish to be
- Information about the health system, its financing and how to access it should be well distributed
- Health care facilities should be well sign-posted

5. Prompt Attention

- Knowledge that access to fast emergency care exists
- Having geographically accessible health care facilities
- Having short waiting times for consultations
- Having short waiting lists for non-emergency surgery
- Having short waiting times on telephones when making appointments (where this is the practice)

6. Support for Social Integration/Social Consideration/Access to Social Support

- The health system facilitates the integration of patients into the care process and then back into society following an illness episode and considers the impact that the process of care has on the patient’s family and social networks
- Health services facilitate the support of the patients through visits by relatives and friends being permitted during care

Religious practices or traditional observances that do not prove a hindrance to hospital activities, adversely affect patient health or hurt the sensibilities of other individuals are permitted

7. Quality of basic amenities

- Clean surroundings
- Adequate furniture
- Healthy and edible food
- Sufficient ventilation
- Clean water
- Clean toilets
- Clean linen

8. Choice and Continuity of Health Care Provider

- Right to choose a health care provider within an institution (if more than one provider is available)
- Choice between health care institutions
- Able to consult a specialist if so desired