Implementation Protocol for Testing the Washington Group (WG) General Measure on Disability

Background on the rationale for and development of the WG questions

Disability as an umbrella term refers to problems, such as impairment, activity limitation or participation restrictions that indicate the negative aspects of functioning. While it is important to collect information on all aspects of the disablement process, it is not possible to do so in censuses or surveys not dedicated to disability. However, important information on selected aspects of disability can be obtained from censuses.

In their ongoing deliberations, the Washington Group (WG) has agreed that measurement of disability is associated with a variety of purposes which relate to different dimensions of disability or different conceptual components of disability models. A fundamental agreement of the WG was the need for a clear link between the purpose of measurement and the operationalization of indicators of disability. Equalization of opportunities was agreed upon and selected as the purpose for the development of an internationally comparable general disability measure. This purpose was chosen because:

1) It was relevant (of high importance across countries with respect to policy), and;
2) It was feasible (it is possible to collect the proposed information using a comparable general disability measure that includes a small set (1-4) of census-like questions).

In order to address this purpose, we begin by identifying persons who are at greater risk than the general population of experiencing restrictions in performing tasks (such as activities of daily living) or participating in roles (such as working). Measurements intended to identify this 'at risk' population represent the most basic end of the spectrum of activities (i.e. functional activities such as walking, remembering, seeing, hearing). This 'at risk' group would include persons with limitations in basic activities who may or may not also experience limitations in more complex activities and/or restrictions in participation depending in some instances on whether or not they use assistive devices, have a supportive environment or have plentiful resources.

Based on these decisions, the Washington Group has developed this question set for use on national censuses for gathering information about limitations in basic activity functioning among national populations. The questions were designed to provide comparable data cross-nationally for populations living in a great variety of cultures with varying economic resources. The objective was to identify persons with similar types and levels of limitations in basic activity functioning regardless of nationality or culture. It was not our purpose to identify every person with a disability within every community. We recognize that this may not meet all the needs for disability statistics, nor will it replicate a population evaluated across a wider range of domains that would be possible in other forms of data collection or in administrative data.

The census format requires that a limited number of questions be devoted to any one statistic that needs to be produced. For the reasons of simplicity, brevity and comparability, the choice was made to identify limitations in domains of basic activity functioning that are found universally, which are most closely associated with social exclusion, and which occur most frequently. The information that results from the use of these questions is expected to:

1. Represent the majority, but not all persons with limitation in basic activity functioning in any one nation.
2. Represent the most commonly occurring limitations in basic activity functioning within any country.
3. Capture persons with similar problems across countries.

The proposed questions identify the population with functional limitations that have the potential to limit independent participation in society. The intended use of this data would compare levels of participation in employment, education, or family life for those with disability versus those without disability to see if persons with disability have achieved social inclusion. In addition the data could be used to monitor prevalence trends for persons with limitations in the particular basic activity domains. It would not represent the total population with limitations nor would it necessarily represent the 'true' population with disability which would require measuring limitation in all domains and which would require a much more extensive set of questions.

Question batteries and question by question specifications (see Appendix 1 for detailed plan)

These specifications provide detailed explanations of objectives, conceptual definitions, and specific instructions related to each question that is asked of respondents.

Objectives of the testing program

1) Purpose of testing

The question set being proposed by the Washington Group for use on Censuses or similar surveys was designed to produce comparable data cross-nationally. To do that and to keep within the limitations imposed by the Census format, the Group felt that the best option was to focus only on selected functioning, basic domains and to ask just one question per domain. It was recognized that this approach would not meet all the needs for disability statistics and that the population identified as disabled by these measures (difficulty functioning in any of the domains) would not replicate a population evaluated across a wider range of domains. In addition, it was recognized that one non-specific question per domain would not identify all persons with difficulty in that domain.

In addition to developing this question set, the Washington Group also included in its work agenda the development of a plan for testing the proposed question set. The objectives of this plan are to determine if: 1) the questions are being interpreted as intended by the developers in that they are capturing the important aspects of the functional domains selected and 2) the questions are interpreted consistently across countries. The first objective includes determining whether the single question per domain provides a reasonable representation of those with functioning difficulties in that domain.

Two types of tests, cognitive and field, are proposed for determining if the questions are comparable cross nationally and if they are capturing the information they are intended to capture. In most instances, cognitive testing precedes field testing and modifications are made to the questions to be field tested based on the results of the cognitive tests. This will not be possible in this instance as testing will be taking place at different times in different places and it will be necessary to review all of the results before making changes to the questions. If possible, another round of testing will be considered in the future. The testing has been designed so that it can be administered as consistently as possible across countries. Ideally, countries participating in the pre-testing activity will be able to conduct both elements of the testing protocol. In that way, more extensive evaluations can be done within and across countries.

Field tests can take various forms. We anticipate two types of field tests. In some cases, countries will be in the process of testing a Census or survey and will add the Washington Group questions to that test. In situations like this, the testing of the WG questions will have to be done within the context of the larger test. The resulting reduction in flexibility might affect how much of the test can be included. In these cases, a core module will be identified that should be used. Other countries might be able to mount a test just of the WG questions. In this situation, the entire testing protocol should be used.

The objectives of the test are described below.

a) Determination of whether the single question per domain is representative of that domain: For each of the domains included in the Washington Group question set, there exist longer batteries that tap various aspects of the domain. For example, in the case of vision, it is possible to ask questions on near vision, far vision, peripheral vision, etc. A set of question that taps these aspects of vision will be included in the field and cognitive tests and the responses to the general question about 'difficulty seeing' will be compared to the responses to the more detailed set. From this comparison it will be possible to determine, for each domain, which aspects of the domain are captured or missed by the Washington Group question and it will be possible to see if these relationships hold across countries (see objective 2). The Washington Group questions will be evaluated as to whether responses to the single question per domain are consistent with responses to the detailed questions, that is, if difficulty is reported for the detailed items, is difficulty reported for the Washington Group item or the opposite, is difficulty reported on the single item but not on the detailed items? The validity of this test is dependent on how well the detailed questions work. Attempts will be made to choose questions that have been used previously and have some degree of validity but there is likely to be some disagreement on whether the detailed questions are more effective in eliciting the desired information. There is likely no easy way to evaluate the goodness of the Washington Group questions on this testing objective but the information generated will be extremely useful for understanding the properties of the short set. See attachment A for a listing of the detailed questions that correspond to the short set.

b) Determination of whether the questions produce comparable data across countries: A major source of non-comparability of data across countries is that the cultural context introduces differences in how questions are interpreted. The Washington Group questions were designed to reduce the possibility of differences in interpretation by focusing on aspects of functioning that would be the most independent of culture and place. A major objective of the testing is to see if this in fact was achieved. The cognitive tests addresses the interpretation issue more directly by eliciting information using less structured techniques on the processes the respondent went through to answer the questions. The cognitive test needs to address not only the question stem but the response categories as well as this aspect of the question might be the most non-comparable across cultures. While the richness of the information obtained from the cognitive process is much greater than that from a field test, the process is less standardized and is more dependent on the skill of the interviewer in eliciting information and recording it. It will be harder to analyze the results of the cognitive interviews cross nationally and the results will not be as definitive but important information will be obtained. Among the criteria to be used to determine that a question has been interpreted in the same way is that similar types of responses are provided to the cognitive probes such as "what were thinking about when answered". The more detailed set of questions for each domain described above provides also provides some information on how the questions are interpreted cross nationally by evaluating if the relationships between the single question and the more detailed questions are the same across countries but this kind of test is limited by how comparable the more detailed questions are (see objective 1).

c) Determination of how the Washington Group questions work as a set in comparisons with other questions used by the country: While not an overall objective of the current test, countries where information on disability is already being collected might want to include these questions in their test to see how the results from the Washington Group questions compare with questions already in use.

Ideally, the questions should work for the population as a whole and for important subgroups. It will not be possible to undertake a test on all potentially important subgroups and each country that participates in the test should determine which population groups need to included in the test. At a minimum, testing information should be available for the total population over age 5 as the

questions being tested are not appropriate for young children. Given that disability varies significantly with age, the test should include samples (see section on sample characteristics) of sufficient size so that evaluations can be made for large age groupings such as those 45 and over. Other age groups would be of interest but this will result in increases in cost as sample sizes within each subgroup need to be large enough to do the evaluation.

2) Evaluations

In order to establish "fitness for purpose", assessments of validity and reliability need to be conducted. The following evaluations are proposed.

a) Validity

Content validity

Test how well the WG question set compares with an expanded disability measure(s) (the short set of WHS/WHODAS or other questions) also collected in the test instruments.

Conduct sensitivity and specificity analysis to provide information about the suitability of the combined WG question set to be used as a general disability measure, and if it can be used, the extent to which it can be used and caveats to such use.

Use of the Kappa Statistic for pairwise comparison can elicit useful information about the relatability of sets of questions that appear to be collecting data about a similar concept e.g. derived disability status as a 'yes/no' output response from combining scaled responses from the proposed questions. The scale response categories can be combined in a number of ways for analysis purposes.

Criterion related validity

Test individual WG questions, e.g. sight loss, against the relevant similar concept in a comparison measure (that is, the selected WHS/WHODAS or other questions included in the test instrument).

See comments under 'content validity' in relation to use of the kappa Statistic.

The sensitivity and appropriateness of questions including cultural sensitivity need to be considered in this context. Cultural reasons for variation between countries, or between groups within countries, might be best examined via 'expert discussion'. This aspect fits more into 'face validity' below.

Face validity

Assess whether the measure 'looks to be valid'. Interviewer feedback will inform this issue, as will considerations as to whether test results can be compared favorably/logically with local knowledge of disability. Are the data as would be expected? Can the results be explained? Are the data considered to be useable?

Interviewer feedback should be assisted by completion of a debriefing form after workloads have been completed, as well as facilitated group discussion between interviewer and office staff. Identification of questions/wording which were not understood, confused respondents, or appeared to produce incongruous answers should be identified. Respondents need for clarification of meaning should also be identified.

A comparison of prevalence rates (where countries have been able to incorporate the proposed questions into an existing survey with sufficient representative sample to generate prevalence estimates) against previous output data should be made, with expert discussion to examine possible reasons for difference in measured rates. How do the rates obtained

compare with 'what would be expected' and against other countries, both developed and non-developed.

Can the data be examined in a meaningful way for planning purposes in relation to existing administrative data? With some existing disability counts there are fewer people identified from the survey/census than there are receiving services targeted specifically at people with a disability.

b) Reliability

Need to assess the repeatability of the measures. Conduct a test/retest analysis in countries where the instrument can be retested. Calculate the kappa statistics for individual questions, as well as the question set.

Test/retest

Where resources allow a retest will provide valuable information on the stability and repeatability of the questions being asked.

At the completion of initial interview, it should be explained to the respondents that a follow-up interview should also be conducted. This follow-up interview should take place between 2 to 4 weeks after the initial interview. Too soon and respondents will simply be 'remembering' their responses, and too long risks higher number of changes off residence and real change in disability status for individuals.

Where participation in the interviews is voluntary, agreement from respondents for the call-back could be obtained at the initial interview completion.

For the retest interviews:

1. Repeat the initial interview with exactly the same procedures and wording.
2. At each selected household identify whether any respondents from the initial interview have been omitted, or whether any resident has been added since initial contact.
3. Identify whether the same person supplied the information for both the initial and follow-up interview.
4. Where possible, identify where responses differ between initial interview and follow-up interview. Ask respondents for their understanding of the reasons for these differences - stressing to them that there is no right or wrong answer to the questions.

Use of the Kappa Statistic to examine relatability between responses will determine the repeatability/stability of the question sets. It will be important where possible to take into account the 'same/different respondent' effect.

An issue to be considered for retesting is whether the same or different interviewers should be used. Where interviewers are experienced and trained to ask questions exactly as worded, and not to 'lead' or influence responses the effect on the data of using different interviewers for the initial and follow-up interviews can be minimized. Where this is not feasible, consideration should be given to using the same interviewers where possible. The output dataset should contain an indicator item to show if the same or different interviewer was used.

Issues for the group to consider and provide comment: We need to establish who will undertake the analysis, and in what format the data will need to be sent for collation.

Testing Protocols

Translation: adaptation of Euro-Reves method (see Appendix 2 for detailed plan)
Points covered in the translation protocol:

1. A finalized questionnaire
2. Prepare translation 'cards' in English
3. Each card represents a question and explains why certain words are used and what we are trying to measure.
4. Individuals working in the field of health and who have an understanding of what we are trying to accomplish are chosen as translators.
5. Characteristics of translators are as follows:
   a. Target language as mother tongue
   b. English as working language
   c. Understanding of health concepts used
6. The translators are briefed to translate first the cards explaining the concepts and then the health module itself.
7. Once the translations are returned we send them out again to another person to check.
8. The 'checker' is given instructions not to provide another translation but to answer a questionnaire on whether each question had been properly translated to tap the concepts and if not, why not.
9. The checker provides reasons for alternative wording.
10. The comments from the translation checkers is reviewed and agreed on a final version of the questionnaire.

Sample design (see Appendix 3 for sample size calculator)
The sampling design for a particular study depends on the purpose of the research. In order to test the small set of census questions on disability, three separate procedures are being carried out: pre-testing (do the questions make sense), internal validity (do the questions measure what they are supposed to be measuring) and pilot testing (how do the questions work in a survey or census context). Each procedure requires a different sampling design. It is preferable that all these stages are carried out in sequence but logistical and economic restraints may mean that this may not always occur.

Pre-testing including expert review
The research team in each country should answer the questions themselves to see if they can answer all of them without any problems. If the team members do not find any problems with the questions, a number of colleagues or family members or friends should be asked to answer the questions and any difficulties recorded. Usually, the biggest problems with questions relate to clarity, comprehensiveness and acceptability and these are picked up during this process. This process does not cost much and can be done speedily.

The questionnaire should then be administered to a larger sample, around 50 subjects. This more extensive test is the pre-testing phase. Persons selected for the pre-test should have the same background profile as the target population of the survey. However, since the sample for the pre-test is still aiming to test clarity and comprehension it is important that people of different educational levels are included.

In summary, the sampling at this stage is based on informational, not statistical, considerations. Its purpose is to maximize information, not to facilitate generalization. The criterion invoked to determine when to stop sampling is informational redundancy, no further problems are being identified, not a statistical confidence level.

Internal validation
The purpose of internal validation is not to generate data for generalization (the same as the pre-testing stage) but to assess whether the census questions work, i.e. comparing the census responses with those from a separate set of questions – a greater number and more specific – yet covering the same domains.

In this case the sample needs to include people with and without the disabilities covered by the census question. Asking the census questions and the more elaborate set of questions of people who have no disabilities at all will not help as everyone will respond, no difficulty to all, to all the questions. Such a correspondence would be fallacious and misleading. Asking the questions of everyone who has a severe disability in the domains included in the census would also give a similar spurious correspondence.

Hence, it is desirable to include in this sample as far as possible people with and without disabilities, preferably those who have difficulties in the domains under consideration. Usually, one does not know the answer to the disability questions before they are asked so more transparent variables can act as proxy. For example, people in older age groups are more likely to have disabilities than those in younger age groups. It is often useful to have an urban and rural split and include both men and women as they tend to respond to health questions somewhat differently. Individuals from disability organizations should be invited to be involved at this stage.

In essence this strategy is known as quota sampling but it is not being used, in the market research sense, for the purpose of statistical inference. In fact, the sample should not be representative of the population; people with disabilities are being over-sampled for the purpose of testing the census questions. The technique is called "quota sampling" because a quota is set for different sections of the population according to sex, age, income, social class, occupation and so on. Quota sampling does not require a sampling frame.

A minimum of 200 interviews are required, but the more that can be carried out the better as this gives a greater chance of identifying subgroups where the test questions are particularly problematic. It is important to recognize that this figure represents the number of people who have agreed to participate and far more people will have to be approached to arrive at this number.

Another way in which we are carrying out this field-test is to include the proposed census questions in a large survey which is already taking place as distinct from specifically setting up a study. In this case, the test incorporates the advantages and limitations of the sampling design of the larger survey. The main advantage is that it will generate a large sample.

Pilot testing

So far we have been looking at pre-testing and validation and have focused on individual questions or instruments. Pilot testing, on the other hand, concerns the complete questionnaire to be used in the final survey or census. Pilot testing is not concerned with how individual questions are understood by respondents and what kind of probes or clarifications for questions is needed. These issues should have been tested and settled during the pre-testing and validation stages.

During the pilot testing, the following items are assessed:

- order and location of questions in the questionnaire;
- how well potential jump rules in the complete questionnaire work;
- length of complete questionnaire and the time taken to fill it in;
- respondent burden.

The pilot testing is intended to be a test of overall survey process, covering the questions, and the entire survey logistics and organization. For pilot testing, a sample of 100 to 200 should be selected. The detailed methodology of the planned census or survey should be applied and the sample of individuals chosen should resemble the final sample as closely as possible following the same sample design if possible or of appropriate.

The data collection method used during the pilot has to be the same as in the planned survey or census: mail administration, face to face interview, computer assisted interview or telephone assisted interview.

It is not possible to describe in detail here the often complex sampling designs used in large national surveys. The most commonly used are clustered, multi-stage, random probability samples. Those planning to use such designs should seek advice from sampling experts or consult the literature available on this subject.

Cognitive test (see Appendix 4 for detailed plan)

The objective of the cognitive test is to determine if the questions are being interpreted as intended by the developers and if this interpretation is consistent across countries. Cognitive testing obtains in-depth information about the respondent's understanding of the questions and the processes they went through in determining their answers. The cognitive test proposed here are more structured than is often the case. This is done to ensure a greater level of standardization across test sites. The test is composed of several components: questions asking the interviewer to report on problems the respondent had with the questions (e.g., needing the questions repeated), traditional cognitive probes designed to obtain information on the respondent's thought process, questions derived form previous cognitive tests about specific factors related to how respondents answer these questions and questions on specific aspects of the functioning domains addressed by the core questions. The aim is to understand how the response mechanisms operate in the different countries in which the questions will be tested.

Field Test instrument and instructions (see Appendix 5 for detailed plan)

In a field test, the questions are administered under conditions that closely approximate how the final study will be done. Although many Censuses use multiple modes, a common practice is to have an enumerator visit the household where the questionnaire is administered. Field tests of Census questions often involve administering the questionnaire to a randomly selected sample. Such a test is possible for the disability questions but since disability is a relatively rare characteristic, large samples are needed to fully test the questions. An alternative is to test the questions on a sample that is selected based on the probability of responding yes to one or more of the Census questions. Different methods are used for each of these approaches and the protocol addresses the requirements of each.

One option for conducting a field test is to design a special study that will focus primarily on the disability questions. As field tests can be expensive, this is not always an option. Another option is to tie the testing of the disability questions to another test. The content of the field test will vary by which of these options is chosen. A special study allows the test to include a larger number of additional questions that will shed light on how the disability questions are being answered. If the test of the disability questions is added to another study, the opportunity to add extra questions maybe more limited.

The Field Testing protocol discusses recommended content including different options depending on how the test will be conducted and sample strategies for different design options.

The main test is to see how the main core questions as proposed by the WG function in different countries. However, in order to have a better understanding of these core questions, it is useful to compare that set to a larger set of more detailed questions. The point of this would be to determine whether the same population is identified by each set. The larger more detailed set would use questions in the same domains of functioning as covered in the core set with subdomains being covered in some detail.

In addition, there are two other domains of functioning which could be added as a further set of questions. These are domains of learning (Do you have difficulty learning a new task?) and the domain of interpersonal relationships (e.g. Do you have difficult getting on with others in your community?). These additional domains can be added to determine if they identify the same or a different population to that identified using the core sets or more detailed questions for the core

domains. The combination of the core domain questions together with these additional two domains provide a good coverage of the more important domains of human functioning.

Lastly, the questions used by an individual country in a recent census or survey could be added to provide a further comparison between the population identified with those questions and that identified by the core set.

While the core set and the detailed questions would be a basic minimum to be administered, the other sets can be added to the test.

The question set to be used will depend to a large extent on the nature of the test to be conducted and whether more detailed questions can be added other than the core set.

The final instrument to be tested with the minimum set of questions is presented in Appendix 3. The instructions to be given to the interviewers are set out in the training manual described above.

Form design / question wording / and introduction to respondents at initial contact (i.e. prototype)

To be developed

Enumerator training (see Appendix 6 for detailed plan)

In this section, interviewers are provided with instructions on survey administration, description and handling of the questionnaires, best practice in interviewing people with disabilities, and soliciting feedback from respondents about their impressions of the survey.

Data entry / analysis plan, compilation of results, and report writing (see Appendix 7 for detailed plan)

The objective of the proposed analysis plan is to test the consistency of the census questions drafted by the WG in regards to how their interpretation may differ across different core domains, countries, and subpopulations. The analysis plan is meant to complement the cognitive testing being undertaken to gain deeper insight into how these core questions are understood by respondents.