

## **Overview of Implementation Protocols for Testing the Washington Group Short Set of Questions on Disability**

### **Rationale for and development of the WG questions**

Disability serves as an umbrella term for impairments, activity limitations or participation restrictions (ICF, 2005). While it is desirable to collect information on all aspects of disability, this aim can not be achieved in censuses or in surveys not dedicated to disability. However, censuses can be used to obtain data on selected aspects of disability.

Disability can be measured for a variety of purposes. Each purpose can be related to different dimensions of disability or different conceptual components of disability. Hence, there is a need for a clear link between the purpose of measurement and the operationalization of the indicators of disability. The Washington Group selected equalization of opportunities as the primary purpose for an internationally comparable short set of disability questions. 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 small set of questions (4-6) that could be included in censuses, comparable across populations).

In order to address this purpose, the questions identify persons who are at greater risk than the general population of experiencing restrictions in performing complex activities

(such as activities of daily living) or participating in roles (such as working) if no accommodations were made. Measurements intended to identify this 'at risk' population represent the most basic end of the spectrum of activities (i.e. 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. This in turn may depend on whether or not they use assistive devices or have a supportive environment.

Based on these premises, the Washington Group has developed a short question set for use on national censuses for gathering information about limitations in basic activities in national populations. The questions were designed to provide comparable data cross-nationally for populations regardless of culture or economic resources. The objective is to identify persons with similar types and levels of limitations in basic activities regardless of nationality or culture.

The questions will not identify every person with a disability in every community. The choice of a limited number of disability domains means that all the needs for disability statistics can not be realized in the census. A comprehensive picture of disability can only come from large, national, sample surveys or administrative data.

Within a census format, only a limited number of questions can be devoted to health. For the reasons of simplicity, brevity, and comparability, questions on limitations in basic activity domains were chosen; they are most closely associated with social exclusion,

and occur most frequently. The information that results from the use of these questions is expected to:

1. Represent the majority, but not all disabled persons with limitations in basic activities;
2. Represent the most commonly occurring disability domains in any country;
3. Capture persons with similar problems across countries.

The proposed questions identify a population with functional limitations that have the potential to limit independent participation in society. The data can be used to compare levels of participation in activities covered in the census such as employment, education, or family life for those with and without disability. The data could also be used to monitor prevalence of limitations among in each of the four chosen domains. It would not be possible to estimate the total population of persons with limitations. This would require measuring limitation in all domains and would require a much more extensive set of questions.

### **Proposed Washington Group Short Measurement Set on Disability**

#### **Introductory phrase:**

The next questions ask about difficulties you may have doing certain activities because of a HEALTH PROBLEM.

#### **Core Questions:**

1. Do you have difficulty seeing, even if wearing glasses?
  - a. No - no difficulty
  - b. Yes – some difficulty

- c. Yes – a lot of difficulty
- d. Cannot do at all

2. Do you have difficulty hearing, even if using a hearing aid?

- a. No- no difficulty
- b. Yes – some difficulty
- c. Yes – a lot of difficulty
- d. Cannot do at all

3. Do you have difficulty walking or climbing steps?

- a. No- no difficulty
- b. Yes – some difficulty
- c. Yes – a lot of difficulty
- d. Cannot do at all

4. Do you have difficulty remembering or concentrating?

- a. No – no difficulty
- b. Yes – some difficulty
- c. Yes – a lot of difficulty
- d. Cannot do at all

**Additional Questions:**

5. Do you have difficulty (with self-care such as) washing all over or dressing?

- a. No – no difficulty
  - b. Yes – some difficulty
  - c. Yes – a lot of difficulty
  - d. Cannot do at all
6. Because of a physical, mental or emotional health condition, do you have difficulty communicating, (for example understanding others or others understanding you)?
- a. No – no difficulty
  - b. Yes – some difficulty
  - c. Yes – a lot of difficulty
  - d. Cannot do at all

**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**

In addition to developing this short 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 test are to determine 1) whether the questions are being interpreted as intended by the developers in that they are capturing the most relevant key aspects of

the functional domains selected and 2) whether the questions are interpreted consistently across countries. The first objective includes determining whether the single question per domain captures a reasonable proportion 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 valid, i.e. 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 test has been designed so that it can be administered as consistently as possible across countries. Ideally, countries participating in the pre-testing activity would conduct both elements of the testing protocol. In that way, more extensive evaluations can be done within and across countries. However, this may not be possible and countries should conduct as much of the test as possible.

Cognitive tests are intended to identify potential response errors related to question design and to understand how and why these errors might occur. Through cognitive testing, we also hope to identify socio-cultural factors that might influence question response. Cognitive tests usually involve in-depth, face to face interviews with a small sample of respondents representing the group of interest.

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 other surveys and will add the Washington Group questions to the ongoing tests. In situations like this, the testing of the Washington Group 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 should be used. Other countries might be able to mount a test for the sole purpose of evaluating the Washington Group 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 supplementary questions on specific 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 questions that covers these aspects of vision will be included in the field and cognitive tests and the responses to the Washington Group 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 also be possible to see if these relationships exist across all countries (see objective 2). 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 been found to 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. Even if it is not possible to reach a definite conclusion about the validity of the Washington Group questions, 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 is the case. The cognitive test addresses the interpretation issue more directly by using less structured techniques to elicit information on the processes the respondent went through to answer the questions. The cognitive test needs to address the question stem and the response categories since this aspect of the question (the response categories) might be the greatest source of differences between 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. Among the criteria to be used to determine whether a question has been interpreted in the same way is

the extent to which similar types of responses are provided to the cognitive probes (such as “what were you thinking about when you answered”?) The more detailed set of questions for each domain also provides some information on how the questions are interpreted cross nationally by examining similarities or differences in the relationships between the single question and the more detailed questions across countries. The effectiveness of cognitive testing is enhanced by having the same detailed questions in as many countries as possible (see objective 1).

- c) Determination of how the Washington Group questions work as a set in comparison with other questions used by the country: While not an overall objective of the current test, countries for which 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 subgroups. It will not be possible to undertake a test on all potential subgroups. Each country that participates in the test should determine which population subgroups need to be included in the test. At a minimum, testing information should be available for the total population aged 5 or over as the questions being tested are not appropriate for younger 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 aged 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 have confidence in the conclusions from the test.

## **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**

#### Face validity

Assess whether the measure 'looks right '. Interviewer feedback will inform this issue, as will considerations as to whether test results can be compared favorably / logically with local knowledge of disability. If someone who is blind says they have no difficulty seeing or if someone in a wheelchair says they have no difficulty walking then something is not right.

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 instances that 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 a sufficiently representative sample to generate prevalence estimates) against previous output data should be made, with expert discussion to examine possible reasons for differences in measured rates. How do the rates obtained compare with 'what would be expected' and against other developed and developing countries.

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.

#### Content validity

Test how well the Washington Group 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 Washington Group question set to be used as a short 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 reliability 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 by 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. visual 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'.

### **b) Reliability**

The repeatability of the measures needs to be assessed. 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

A retest, if resources are available, 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 might also be conducted. This follow-up interview should take place between 2 to 4 weeks after the initial interview. If re-tested too soon, respondents will simply be 'remembering' their responses; if too long, there is a higher risk of change of residence and real change in disability status for individuals.

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.

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, (i.e. 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.

### **Testing Protocols**

#### **Translation: adaptation of Euro-Reves method** (see Appendix 2 for detailed plan)

In order for the short set of Washington Group questions and the associated test questions to be understood, in a way that is comparable within and across countries that rely on different languages and dialects, it is necessary to have a translation procedure that yields equivalent versions of the test questions across a variety of settings and cultures. The translation protocol is intended to outline such a procedure for use with the short set of Washington Group questions. Different methods have been developed to standardize translation processes across countries. The two most prominent methods are forward-backward translation used by the World Health Organization and the conceptual translation method used by Euro-Reves in the development of the European Health Status module. The forward-backward method begins with a version the question set in the language in which it was originally

developed, for example, English. This version is given to professional translators who translate the module into another language, for example, German. Then, another professional translates the German version back into English (back translation) and the two English versions are compared. However, strict translation does not necessarily capture the underlying concept to be measured. Therefore, an alternate method (which is referred to as the conceptual method) was developed by Euro-Reves and applied to development of the European Health Status module. Although both methods are intended to yield translations that are conceptually equivalent between countries and cultures, the conceptual method does so by relying on detailed explanations of the terms used in each survey question as well as the underlying concepts that the questions were intended to measure. This approach differs from the forward-backward method in the “backward” step when, rather than translating the question back into the original language, a checker determines whether or not each question was properly translated such that the intended concepts were actually captured. For Washington Group purposes, it was agreed to adopt the conceptual translation method since a key aspect of our endeavor was to develop question sets and a process of implementing question use cross-nationally that yielded internationally comparable results. Thus, the Washington Group believed that it was critical for respondents in different countries and cultures to have the same understanding of the questions, regardless of whether the terms were the same.

**Cognitive test** (see Appendix 3 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 is 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 processes, questions derived from previous cognitive tests about specific factors related to how respondents answer these questions, and questions on specific aspects of the activity 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.

The research team in each country should answer the questions themselves to see if they can answer all of them without any problems (i.e. conduct an expert review). 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 results of this step should be recorded but the wording of the questions should not be altered in any way based on this review.

The questionnaire should then be administered to a larger sample, around 50 subjects. The sample should include people with and without the types of limitations included in the short set of questions, but the number with limitations needs to be large enough to obtain information on the range of questions. Persons selected for the cognitive test should have the same background profile as the target population of the survey. However, since the aim for the pre-test is 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.

**Field Test instrument and instructions** (see Appendix 4 for detailed plan)

The purpose of field testing is to assess whether the Washington Group questions work in a setting that is similar to that which will be encountered in the census and survey. Quantitative data are collected on how the Washington Group questions compare to a longer battery of more specific questions that cover the same domains.

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. It

is not necessary and possibly not desirable for the sample to be representative of the population; people with disabilities should be over-sampled for the purpose of testing the census questions. 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. Another 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. Individuals from disability organizations should be invited to be involved at this stage.

200 interviews are recommended for a field test, but the more that can be interviewed 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 this field-test can be carried out 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.

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.

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 difficulty getting on with others in your community?). These additional domains can be added to determine the extent to which they overlap with the population identified using the core set or more detailed questions reflecting the core domains. The combination of the core domain questions together with these additional two domains provides a good coverage of the principal domains of human functioning.

Lastly, the questions used by an individual country in a recent census or survey could be added to the field test to provide an additional 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 4. The instructions to be given to the interviewers are set out in the training manual described in Appendix 5.

**Enumerator training** (see Appendix 5 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 6 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.