A Delphi exercise to refine the WHO three-point Disability Grading system for leprosy, and to develop guidelines to promote greater accuracy and reliability of WHO Disability recording.

HUGH CROSS*, FIRMANSYAH ARIEF**, KERSTIN BEISE**, WIM BRANDSMA***, JOSEPH NGOZI CHUKWU****, JANNINE EBENSO******, MARIA APARECIDA DE FARIA GROSSI*******, HERMAN-JOSEPH KAWUMA*******, LINDA LEHMAN********, SUERESH MANI*********, SATHISH KUMAR PAUL**********, ANGELIKA PIEFER***********, WIM VAN BRAKEL************ & JINLAN LI*************

*American Leprosy Missions, GPO 8975 EPC 385, Kathmandu, Nepal
**Netherlands Leprosy Relief, Graha Aktiva Building, 7th floor, Suite No. 701 Jl. H.R. Rasuna Said Blok X - 1 Kav 3, Jakarta Selatan 12950, Indonesia
***Free University Amsterdam (assistant professor, Athena institute) and Royal Tropical Institute Amsterdam, Koolmeeslaan 18, 3871 HG Hoeveleken Netherlands
****German Leprosy & Tuberculosis Relief Association, 35 Hill View, Independence Layout, Enugu, Nigeria
*****The Leprosy Mission International, 80 Windmill Road, Brentford, TW8 0QH, UK
******Secretariat of Health of Minas Gerais State – Brazil, Rua Monte Alegre, 373/202 Serra - Belo Horizonte - Minas Gerais – Brasil, 30240230
*******German Leprosy and TB Relief Association (GLRA), P.O. Box 3017 Kampala, Uganda
**********American Leprosy Missions, R. Castelo de Alenquer 390 Apt. 302, Belo Horizonte MG 31330-050 Brazil

Correspondence to: Hugh Cross, GPO 8975 EPC 385, Kathmandu, Nepal (e-mail: hcross@leprosy.org)
**Summary** A Delphi Exercise was undertaken with two objectives:
(1) To ascertain whether operational definitions for WHO Disability Grading could be formulated by consensus.
(2) To develop a set of simple guidelines based on those definitions for the guidance of health workers.

Fifteen people with general expertise in prevention of disability due to leprosy were invited to participate as Delphi panel members, twelve responded positively. Eight issues that commonly cause confusion provided foci for the exercise. Operational definitions aimed at resolving those issues were developed by consensus. Simple guidelines for health workers, based on those definitions were also created and supported by consensus. This paper presents the process followed and the outcomes gained from the endeavour.

**Introduction**

Before 1960 there was no standard conceptual framework for classifying leprosy-related impairments or disabilities. The Second Expert Committee on Leprosy tried to address this need by devising a “Scheme of Classification for Grading of Physical Disability Resulting from Leprosy”.1 The system was offered as a means by which “the extent of deformity in different countries” might be estimated. It was, however, rather complex involving different five-point scales for disabilities of the hand, foot and face with another two-point scale for “Miscellaneous” impairments (gynaecomastia and involvement of the larynx).

Six years later the Third WHO Expert Committee had some discussion on disability which resulted in the following comment, “The committee emphasized the importance of producing a simple and practical classification of disabilities for use in field projects”.2

When the Fourth Expert Committee was convened in 1970 it was reported that the 1966 classification had been used and that useful data had been collected. It was acknowledged, however, that for programme application it was too complex. It was in the 1970 report that focused interventions for the prevention of disability were recognised as important, not only for the well-being of leprosy-affected people but that they also had a “favourable influence on attendance at treatment clinics and thus on leprosy control”.3 It was specifically for those reasons that a revised classification system was presented by the 1970 Expert Committee.
The 1970 Classification was presented as a system “so simple and practical that it could be used by auxiliary health workers”. The Fifth Expert Committee which was convened in 1977, however, reported that, “It was understood by well trained workers, but appeared to be rather beyond the comprehension of primary health workers”. Although criticising the classification system, no changes were proposed.

The Sixth Expert Committee (1988) made another concerted effort, and introduced the three-point system of classification, which was devised by that committee and is still in use. The committee conceded that no single system will meet all requirements but they offered the revised system which they suggested would be appropriate for “the collection of general data regarding disabilities and/or impairments”.

The Seventh Expert Committee addressed the issue of disability quite critically (disability is an umbrella term, but what is actually graded are impairments). The pursuance of the elimination goal was of primary importance at that time, but the Expert Committee did acknowledge that the number of people living with leprosy-related disability in 1997 far outnumbered the number of active cases. It was presumably in response to that reality that the committee stressed the importance of disability grading and the desirability of a plan to assess the scale of disabilities and impairments using a revised International Classification of Impairment and Disability framework (as yet such a plan has not been developed). The committee did, however, suggest three specific objectives for disability grading amongst which was the first suggestion that disability should be used as an indicator for programme performance:

- to assess the disability burden attributable to leprosy in the community so as to plan necessary actions;
- to use it as an indicator for assessing the performance of the elimination programme;
- to assess the potential for preventing disabilities in individual patients.

WHO disability grading is of limited use for individual patient care as it is not sufficiently sensitive to monitor changes in impairment. The EHF score as developed by van Brakel et al. is more useful, but even that has limitations as van Brakel has indicated. WHO Grade 2 disability is, however, a very significant epidemiological indicator. It is especially important to strive for the accuracy and reliability of disability assessments at this point in time, because in 2012 the Eighth WHO Expert Committee on Leprosy confirmed a target to reach 35% reduction in G2 disability in new cases per million population from 2011 to 2015 and a longer term goal of reducing G2 in new cases to less than 1 per 1,000,000 at global level by 2020. Confidence that these can be achieved will be undermined unless competent assessment and recording of disabilities can be assured.

Two groups of researchers independently established that either after training or with experience, very high levels of agreement could be attained between small groups of people applying the WHO three-point system for disability assessment. The conditions under which those two studies were conducted, however, are not a reflection of the realities of field situations where most assessments are conducted. Establishing the reliability of assessment and recording among health staff in field situations is problematic for operational and logistical reasons, which is perhaps why such studies are not undertaken; or if they are, they are not published.

In 2012, Professor W.C.S. Smith, Chair of the ILEP Technical Commission, led a survey through the Leprosy Mailing List. He wanted to find out whether assessments of disability due to leprosy are conducted and recorded in the same way everywhere. He suggested that
there is over-reporting of WHO G2 disability and action should be taken on classifying conditions that commonly cause confusion among people recording disabilities.

He suggested that simple, clear guidance was needed so that the consistency of disability assessment could be improved. It was in response to this call that a Delphi Exercise was conducted, with the support of the ILEP Technical Commission and in the knowledge of the Leader of the Global Leprosy Programme.

DELPHI EXERCISE

A Delphi exercise or investigation is a well-established method for developing and securing a consensus. Since its conception in the 1950s when it was used by the Rand Corporation for use in defence-related problems, Delphi has been applied extensively to clarify issues that have required sharper definition.

Delbecq and Van de Ven were instrumental in the early development of the Delphi Method. They suggested that the sample size for a Delphi exercise should be dictated by the homogeneity of the group and the nature of the exercise, but that a homogenous group of 10 to 15 participants is generally adequate. Increasing the size of a homogenous group beyond 15 will, apparently, not result in more information and is likely only to increases administration difficulties. As with other consensus methods, for example, the Nominal Group Technique, the selection and definition of experts is cited as being the most potentially confounding effect on a Delphi outcome. Panelists are usually (though not exclusively) recruited by merit of an intimate academic or experiential association with the topic under investigation. Acknowledged expertise or influence may validate a choice of participants; however, Delbecq and Van de Ven suggested that such attributes per se are insufficient for the inclusion of participants. They cautioned that commitment to the exercise, motivation to comply with the demands of procedure and the acceptance of the consensus (even though it may be at variance with personal inclination) are fundamental.

Methodological details vary according to the requirements of individual Delphi projects. Delbecq and Van de Ven explained, however, that the general approach in any consensus exercise, including a Delphi exercise, is that sequential, structured questionnaires are used for participants to rank, or rate, responses to indicate their priorities related to the topics of interest. On receipt of returned questionnaires, information is collated and analysed before being redistributed for further refinement and for final comments.

Method

Fifteen people with general expertise in prevention of disability due to leprosy were invited to participate as Delphi panel members; 12 responded positively (see Appendix 1). Having recruited the Delphi panel, a list of eight questions was emailed to each participant. These related to issues of criteria and the precise definition of terms for those criteria.

After receiving the panel’s replies, the facilitator subjected the participants’ responses to those questions to a simple content analysis, i.e. the essential details from each response were noted and then the frequencies of common expressions were recorded. The expressions most frequently suggested were then used to form definitions for the terms by which the criteria could be explained. None of the definitions that the facilitator formulated were direct quotes from any individual response. They were inclusive of the most commonly occurring expressions describing each term as submitted by all the individuals in the Delphi panel. In
the event that any response was at complete variance with other responses that response was not reflected.

The definition of terms that the facilitator formulated, along with all the responses that had been used to formulate them, were subsequently resubmitted to the Delphi panel. Panel members were requested to consider the definitions and the material from which they had been formulated and were asked to rate the facilitator’s interpretation of ‘the group mind’. Each definition was followed by a scale on which panel members could rate the facilitator’s interpretation of the definition and the simple guideline that had been composed on the basis of each definition. Interpretations could be rated on a scale from ‘inaccurate or inappropriate’ to ‘very precise or wholly acceptable’. To reduce bias, the Delphi participants returned their responses to an independent collator who sent on anonymous results to the facilitator. The facilitator then attributed numerical values to the ordinal data.

(N.B. On assuming the role, the facilitator relinquished any right to express or contribute his own opinions.)

Another scale was then constructed to reflect the strength of consensus. For each task (definition) the median score was found. The decision was that definitions would only be acceptable for inclusion in the guidelines if the median score for that definition was greater than 2. If a score was greater than 2 but less than 3 it was categorised as ‘Robust’, a score of 3 was categorised as ‘Strong’, greater than 3 but less than 4 was categorised as ‘Very Strong’ and a score of 4 was categorised as ‘Complete Consensus’ (Please see Appendix 1).

Had a median score been 2 or less then further rounds of consensus building would have been required. All the definitions, however, were found to be acceptable under these criteria for inclusion (i.e. scored greater than 2).

A document was then compiled with a set of simple guidelines for health workers based on the panel’s inclusive definitions. Those guidelines were then submitted to the panel for review and refinement. After some refinement a draft set of guidelines was submitted to the ILEP Technical Commission for review. The ITC recommendation was that a joint ILEP/WHO endorsement should be sought so that the guidelines might be introduced.

**Results**

*Question 1:* When can a crack on an insensate hand or foot be described as a wound and thereby justify the hand or foot to be graded as ‘2’ (what signs indicate that a crack is a wound)?

*Response:* The size (length and breadth) of the fissure is irrelevant. If any of the following features are noticed the fissure is a ‘wound’:

- it can be clearly seen and felt by an examiner
- it will have two clear edges
- it breaches the full thickness of the skin to subcutaneous tissue (a depth of approximately 5 mm)
- it will have a base of granulation tissue (red spongy tissue)

Other features that may be associated with a crack that justify a classification of Grade 2 are:

- it may present with macerated (white) tissue at the base and edges
it may exude body fluids (exudates or blood)
if it exudes fluid, the fluid may smell foul (depending on infection)

“If a fissure with any such features is observed on a hand or foot that does not have sensory feedback it is considered to be a wound and the hand or foot should be described as presenting WHO G2 disability.”
Level of consensus - very strong

Simple guideline for health workers:
If you have recorded that the patient does not feel a ball pen or filaments on the palms of hands or soles of feet, check the skin on the patient’s hands and feet and look for cracks (long thin openings in the skin). Examine all parts of the hands and feet including parts between and under the toes and fingers.

- if there is any wetness or blood coming from the crack, the hand or foot is graded as WHO G2
- if there are cracks in the skin, but there is no wetness or blood, gently pull the two edges apart with your thumbs. If you can see any red tissue in the bottom of the crack, the hand or foot is graded as WHO Grade 2 (G2)
Level of consensus - very strong

Question 2: Should a hand or foot with scars only be classified as WHO Grade 2?
Response: A scar is an indication that the integrity of the skin has been restored. Although the presence of a scar does suggest that the skin is compromised (less elastic and more prone to breakdown due to increased tissue bulk and density) it is none the less an indication of satisfactory resolution of trauma and should not be graded as an impairment. If scarring is so extensive that it causes a contracture of the hand, foot, or parts thereof, it is the visible disfigurement of the body part per se that warrants a WHO G2 classification, not the scarring.
Level of consensus - very strong

Simple guideline for health workers:
If you have recorded that a patient does not feel a ball pen or filaments on the palms of hands or soles of feet, you still need to consider if the shape of the hands or feet are normal. When you are examining a hand or a foot you may see scars from old wounds or burns. Such scars do not mean that the hand or foot should be graded as WHO G2, but if a scar has caused a part of the hand or foot to be twisted or bent, the abnormally shaped hand or foot should be graded as WHO G2.
Level of consensus - very strong

Question 3: Should muscle weakness in the hands (without clawing or obvious muscle atrophy) be graded > 0?
Response: Early nerve function impairment is comparable in aetiology with sensory nerve damage and should, therefore, be graded. It is indicated by a loss of muscle function, but without visible signs of muscle atrophy or imbalance. It is detected through voluntary muscle testing. Any hand that is found to have less than optimal strength on voluntary muscle testing, but has no visible sign of nerve function impairment should be classified as WHO G1
Level of consensus - robust
Simple guideline for health workers:
In early stages of nerve damage parts of the hand may be weak, but there may not be signs of muscle wasting or clawing. In such cases the hand should be classified as WHO G1. When you do a voluntary muscle test of the hand, if you find that the patient is not able to hold any one or more of the three test positions against moderate force (wrist up, small finger out or thumb up) the hand should also be classified as WHO G1. If muscle paralysis or weakness has caused muscle wasting or finger clawing the hand should be classified as WHO G2.
Level of consensus – robust

Question 4: When can foot drop justify a Disability Grade and what should that grade be? 
Response: A foot that clearly demonstrates weakness of the lower leg anterior muscle group should be graded as follows:

- if on voluntary muscle testing a foot is found to be weak against resistance to dorsiflexion (MRC 3 or 4) or if it only demonstrates residual dorsiflexion power (MRC 1 or 2) it should be classified as WHO G1
- if on voluntary muscle testing a foot is found to be without any sign of muscle activity for dorsiflexion it should be classified as WHO G2
- a feature that instantly justifies a classification of Grade 2 is an obvious high stepping gait.
Level of consensus - robust

Simple guideline for health workers:

- Watch the patient walk a short distance. If the patient lifts his leg high and his foot points downward when he walks you should grade the hanging foot as WHO G2.
- The patient may not show any sign of foot-drop when he walks but he may show weakness when you do a muscle strength test. If the patient can lift the front part of his foot up when his heel is on the ground, but cannot hold his foot up against pressure from your hand you should grade the foot as WHO G1.
Level of consensus - very strong

Question 5: Should the eyes of a person affected by leprosy, with severe visual impairment (visual acuity < 6:60) be classified as WHO G2 regardless of cause?
Response: The eyes of a person affected by leprosy, with severe visual impairment (visual acuity < 6:60) should be classified as WHO G2 regardless of cause.
This task only required that panel members should register their agreement or disagreement with the statement. The strength of agreement on the issue was very strong.

Question 6: Should corneal anesthesia be recorded as Grade 1 and if so how should it be tested?
Response: It is the absence of blink reflex that is the principle indicator of corneal anaesthesia. If in the absence of any visible damage to an eye it is noticed (through unobtrusive observation) that an eye does not blink spontaneously, that eye should be classified as WHO G1. The patient should be referred to a specialist facility for confirmation (and possible treatment).
Level of consensus - very strong
Simple guideline for health workers:

Loss of blink

- Take note of whether the patient’s eyes are damaged or not.
- If the eyes do not look damaged, carefully watch the patient’s face whilst taking the patient’s history.
- Do not explain that you are looking to see whether or not the patient blinks but watch carefully to notice whether the patient is blinking or not.
- If an eye does not blink, that eye should be classified as WHO G1.
- The patient should be referred to a referral unit for confirmation and possibly for treatment.

Level of consensus - very strong

Question 7: Should wounds on hands or feet of people affected by leprosy be graded as WHO G2 if the affected body part still has sensory feedback.

Response: Hands or feet of people affected by leprosy who have sensory feedback intact should not be classified as WHO G2 if they present with wounds.

This task only required that panel members should register their agreement or disagreement with the statement. The strength of agreement on the issue was very strong.

Simple guideline for health workers:

If you are examining a patient’s hands or feet and you notice that there is a wound, do not immediately classify the hand or foot as WHO G2. Test for sensory loss first. If the patient has sensory loss it is correct to classify the hand or foot with a wound as WHO G2. If the hand does not have sensory loss it should be classified as WHO G0 even if it does have a wound.

Level of consensus - very strong

Question 8: If hands or feet are visibly impaired due to an event that occurred before nerve function impairment was detected, should the damaged body part be classified as Grade 2?

Response: When the principal reason for disability grading is that it should be an indicator of detection delay only, leprosy-related impairment is relevant. If the purpose is to grade impairment and/or to consider the risk of further impairment then the cause of pre-existing impairment is irrelevant. Since reporting systems do not provide a structure to differentiate between leprosy related impairment and pre-existing impairments caused by other factors, pre-existing impairments should not influence the grading of the body part. (In practice, since the data recorded will often be used as a detection delay indicator and for impairment assessment too, health workers should be required to record the presentation of pre-existing impairments but they should also note that the impairments are not due to leprosy.)

Level of consensus - strong

Simple guideline for health workers:

When you examine a patient’s hands and feet you may notice things that are not normal: e.g. the loss of a finger, contractures etc. Do not immediately grade the body part as WHO G2. First, carefully ask the patient what the cause of the injury was and when it happened.

- If the impairment to the hand or foot happened before leprosy was diagnosed the hand or foot should not be classified as a WHO G2.
- If the hand or foot does not have sensory loss it should be classified as WHO G0.
(Make a note in the patient’s record that the damage happened before leprosy was diagnosed)
Level of consensus - robust

Discussion

Consensus is the result of a group decision-making process that seeks the consent, not necessarily the agreement, of participants in the resolution of issues. It captures a sense of unity and togetherness of thought. The central question of consensus reliability was investigated in early studies by Van de Ven and Delbecq14 who ascertained that judgmental accuracy may be achieved where the following features are encapsulated in the method of exercise:

- Individuals make independent judgments.
- Individual judgments are expressed through mathematical rank-ordering and/or rating of items.
- The mean value of independent judgments are accepted as indicating group decision.

The Delphi methodology used for this study satisfied those criteria (except that the median was used as the measure of central tendency). The power of a Delphi consensus is that it is not easily biased by views of dominant individuals; it recognises and respects the value of each participant equally. An objective of a Delphi exercise is that normative group behavior (which basically favors the performance of dominant or aggressive characters) will be controlled by nonconformance tactics so that performance and outcome are maximised whilst hidden agendas and negative group dynamics are suppressed. Delphi aims to draw out minority opinions and to promote the tolerance of conflicting ideas.

The outcome of this Delphi exercise will not satisfy all the critics of the WHO three-point system of disability assessment. Although clearer instructions should result in more accurate assessment of individual patient’s conditions (a knock-on effect of which will be better EHF scoring) the system per se remains inadequate for optimal management of individual patients with impairments and disabilities. The main defense of the Delphi outcome is that it should help to resolve common issues that may have resulted in inaccurate recording of Grade 2 disabilities. If it does, the objective of eliciting more reliable indications of the burden of leprosy may be achievable and greater confidence that WHO strategy is affecting impact may be gained.

Conclusion

The accuracy and reliability of disability grading has been an issue of concern to many; a concern compounded by reports in the WHO’s Weekly Epidemiological Record. Those reports have consistently shown improbably wide variations between countries in the proportion of new cases with G2 Disability.

A Delphi exercise has resulted in some refinement of the WHO Disability Grading system and has clarified some of the ambiguities found in the definitions as set in the 1988 WHO Classification of Disability.
The suggested simple guidelines for health workers, which are a product of the Delphi exercise, will be incorporated in a manual for WHO disability grading which will be submitted for publication.

References

1. WHO Expert Committee on Leprosy, Second Report. *World Health Organization Technical Report Series.* 189. 1960. (p 24).
2. WHO Expert Committee on Leprosy, Third Report. *World Health Organization Technical Report Series.* 319. 1966. (p 18).
3. WHO Expert Committee on Leprosy, Fourth Report. *World Health Organization Technical Report Series.* 459. 1970. (p 26).
4. WHO Expert Committee on Leprosy, Fourth Report. *World Health Organization Technical Report Series.* 459. 1970. (p 27).
5. WHO Expert Committee on Leprosy, Fifth Report. *World Health Organization Technical Report Series.* 607. 1977. (p 18).
6. WHO Expert Committee on Leprosy, Sixth Report. *World Health Organization Technical Report Series.* 789. 1988. (p 35).
7. WHO Expert Committee on Leprosy, Seventh Report. *World Health Organization Technical Report Series.* 874. 1998.
8. Van Brakel WH, Reed NK, Reed DS. Grading impairment in leprosy. *Lepr Rev.* 1999; 70: 180–188.
9. WHO Expert Committee on Leprosy, Eighth Report. *World Health Organization Technical Report Series.* 968 2012.
10. Nienhuis WA, Van Brakel WH, Butlin CR, van der Werf TS. Measuring impairment caused by leprosy: inter-tester reliability of the WHO disability grading system. *Lepr Rev,* 2004; 75: 221–232.
11. Brandsma W, Larsen M, Richard C, Ebenezer M. Inter-rater reliability of WHO 'disability' grading. *Lepr Rev,* 2004; 75: 131–134.
12. Smith, C. Disability Grading Survey Results http://www.aifo.it/english/leprosy/mailing_list/2012/250212.htm. 2-25-2012.
13. Delbecq AL, Van de Ven AH, Gustafson DH. *Group Techniques For Programme Planning: A Guide To Nominal Group and Delphi Processes.* Scott Forresman & Company, Illinois, 1975.
14. Van de Ven AH, Delbecq AL. The nominal group technique as a research instrument for exploratory health studies. *Exploratory Health Studies,* 1972; 337–342.
Appendix 1

| Delphi Facilitator | Hugh Cross |
|--------------------|------------|
| Delphi Panel       |            |
| Angelika Piefer    | Cida Grossi|
| Joseph Chukwu      | Herman J Kawuma |
| Linda Lehman       | Satish Paul |
| Wim van Brakel     |             |
| Delphi Collator    | Paul Saunderson |

Appendix 2

Delphi panel members had been requested to rate the accuracy of the facilitator’s interpretations (a simple tick box rating procedure was offered). Options ranged from ‘Inappropriate or Inaccurate’ (interpreted as complete rejection) to ‘Wholly Appropriate or Very Precise’ (interpreted as complete acceptance). On return of the results from the collator the ordinal scores were given numerical values.

| Inaccurate or Inappropriate | low | moderate | high | Very precise or Wholly Acceptable |
|-----------------------------|-----|----------|------|-----------------------------------|
|                             | 0   | 1        | 2    | 3                                 |

For each task the median score ($n = 12$) was calculated and the level of consensus was described as follows:

| Median Score | 0 | 1 | >1.1 $< 2$ | 2 | >2.1 $< 3$ | 3 | >3.1 $< 4$ | 4 |
|--------------|---|---|------------|---|------------|---|------------|---|
| Strength of Consensus | No consensus | very weak | weak | moderate | robust | strong | Very strong | Complete consensus |

If suggested definitions and guidelines scored higher than 2 (robust consensus) they were considered to have sufficient support to be recommended. If two or more members submitted comments concerning the same issue, that issue would be addressed.