Green Pastures Complex
PO Box 28, Pokhara,
Nepal

RELEASE

Rehabilitation Leprosy control AIDS prevention

Statistics and Research Department

Participation scale development programme

Phase 3– Psychometric testing
Final scale
Revised scoring and with dynamicity data

Alison M Anderson
on behalf of the PSP team

Document information

Version No: 5 17/08/2016

Filename: Anderson et al 2004 P-scale Phase 3 Report - psychometric validation.docx

Creation date: New version with dynamicity data and rescoring 30th January 2004

Created by: Alison Anderson

This document is an update of an earlier version which was completed before the dynamicity aspects of the scale were addressed. It therefore includes extra data in many sections. The scoring was revised in the final workshop.

The information in this document should be used in conjunction with the P Scale questionnaire versions 2 [22 questions] and 3 [18 questions].

The P-Scale has been developed by the P Scale team, with funding from ILEP partners – GLRA, TLMI, ALM and resource from centres worldwide.

Table of contents

Summary

The first and second parts of the psychometric testing for the P Scale have been completed.

The original scale showed some evidence of redundancy. A second review in terms of coverage, endorsement and internal consistency removed four questions to give an 18-question scale.

Inter-interviewer agreement is good at 0.80, [n=296], although some questions in some centres have poorer reliability co-efficients. Short-term stability is also good at 0.83 [n=210].

The scale discriminates between people with disability and their unaffected peers. A cut-off for the diagnosis of 'participation restriction' is suggested at a score of 12, a score of 13 or above indicating participation restriction. This cut-off is set at the 95%ile of the observed data; local studies will be necessary to determine appropriate cut-offs in particular target populations and resource settings.

The scale score is significantly associated with the score given by experts. There is also an association with the self-assessment, and with the impairment status as defined by the EHF score although the trends are not perfect. There is no obvious association with BMI; the correlation coefficient was statistically significant but very small.

There were other associations with age, social status and location. These may be true effects – that is, greater participation restriction associated with disability in these sub-groups.

The scale shows differences in score after a life change which mirror the differences expressed by the clients themselves. In many cases these differences are likely to be insignificant; this finding appears to be consistent with reality and not a performance limitation in the scale itself. The scale can be considered to be dynamic, but recommendations should be made for limits on the time frame over which genuine change could be expected and the interpretation of the magnitude of changes.

Recommendations

The technical properties of the scale are adequate to currently enable recommendation of use of the scale in cross sectional studies and in long term follow up of both individuals and groups.

It is possible that the scale may be shortened, with associated loss of technical properties; this may be useful as a population screening tool.

Background

Improvement in social participation should be a key outcome of interventions in socio-economic rehabilitation. Up to now, no objective measure of participation has been available for use. Different individuals visiting a client/patient see the situation differently, and simple comparison between people or measurement of change within a situation has been difficult. Based on observed and spoken indicators of participation from observational studies, a scale has been developed to simplify and standardise measurement of participation [or restrictions in participation], particularly in the context of clients who previously had leprosy. The scale will be primarily for use in assessment of socio-economic rehabilitation and therefore will emphasise domains of participation that reflect this aspect and will be tested amongst this population of people.

Phases 1 and 2 of the scale development study comprised item generation and item selection. This phase, phase 3 covered the psychometric testing of the draft scale. The results are given in this document.

The draft scale

The draft scale was produced from the questionnaire by classical scale development techniques. Endorsement was checked and a selection made of the questions with highest endorsement over all centres and both genders. For the shortlist, discrimination and homogeneity were checked and the items subjected to factor analysis. The approximate appropriate weighting for the responses was determined using multivariate techniques. The vision at this stage was to produce a draft scale that could be administered in 20-30 minutes, has good mathematical properties and is culture and gender free. The draft scale contained 22 items, with six potential responses [same as everyone else, not relevant, no problem, small problem, medium problem, large problem]

Psychometric testing

The psychometric testing programme looked at four features, which are desirable for a scale used for routine monitoring and evaluation.

Two features are based on the concept of reliability, which is defined as the ratio between true subject variability and the observed subject variability, which includes measurement error.
In an experimental situation where clients are interviewed by different staff members, the measurement error will come from the interview technique etc, as long as the time interval between interviews does not include a real change in the client's life. This is termed *inter-observer reliability.*
If the clients are interviewed by the same staff member after a period where no intervention/change has taken place, then the measurement error will reflect only short-term changes. This measurement is termed s*tability.*

In addition, this phase has looked at *dynamicity,* defined as sensitivity to change, and will make some estimate of *discrimination* [the ability to detect a difference between 'Control' and 'leprosy & disability' situations], and of *validity* in comparison with other measures.

The psychometric testing analysis suggested good properties of the 22-question scale, with the possibility of some redundancy. Expert review highlighted four questions for removal. This document gives the results of reanalysis of the final 18-item scale.

Data handling

Data were generated in eight centres over three countries. Each centre entered data into an identical database; databases were merged for analysis. Each centre was responsible for its own training of interviewers, organisation of interviews and data entry. Each centre was requested to make a 100% check of data entry.

A checklist of questions was provided to check whether the person being interviewed had undergone a significant life change since the previous interview. Significant life changes included medical concepts such as a hospital admission as well as social concepts such as a death in the family or a change in employment. Life changes between interviews make the interview ineligible for inter-rater and stability [since the data are *expected* to change] but are required for eligibility for dynamicity, since the test is measuring responsiveness to change. Centres were responsible for monitoring the inclusion of interviews into each part of the study.

A randomisation table was provided for each centre, to ensure that there was a balance between interviewers and first, second and third interviews.

Cleaning in centres

Centres were asked to make a 100% manual check, before the data were passed through for analysis.

Computerised checks

All possible cross tabulations were made, and anomalies highlighted. These included missing data, especially where one part of the question was coded missing and the other had an appropriate answer recorded. Range checks were made, including for derived variables such as time taken to complete the interview. Queries were referred to the centres for review and comment or correction.

Systematic changes

The computerised checks highlighted four systematic problems:

- Controls cannot logically answer question J27, which refers to talking about the health condition. Interviewers had solved this issue is a variety of ways. All results for this question for controls were recoded to 8, 9 [not answered, missing].
- People who live alone cannot logically answer question f8, about sharing food *in the home.* All results for this question, for people who have answered ‘alone’ to the question about living conditions, have been recoded to 8,1[not answered, no problem]
- People who live on a pension probably do not need to work, and find difficulty answering questions g9 and g10. Where the interviewer has made this note, the answers have been recoded to 3, 1 [irrelevant, no problem]. A note as to how to code this option must be added to the q by q.
- People have occasionally not answered the first part of the question, but given a valid answer for the second part, or have answered ‘irrelevant’ to the first part, but gone on to answer the second part indicating a problem with this ‘irrelevancy’. This will require discussion at the review workshop, to determine if better instructions can give a more consistent use of coding.

Demography

Over the eight centres there are 1382 interviews which appear to have valid data.

Of these interviews, 682 are first interviews, either for cases or for controls. A further 335 inter-rater interviews, coded 2, have been performed [limited time interval, different interviewer]. For stability, 218 interviews [3, short time interval, same interviewer] are now complete. There are 147 interviews coded for dynamicity, this includes some interviews where there was a change in the situation, which makes these interviews ineligible for inter-rater or stability, but the timescale may or may not be suitable for dynamicity.

	INTERVIEWN				
PLACE	1	2	3	4	Total
BRAZIL	106	33	15	26	180
DAYA	75	41	48	0	164
KARIGIRI	55	32	29	14	130
KOLKOTA	100	46	34	19	199
NAINI	150	47	30	23	250
NEPAL	67	31	15	45	158
SALUR	78	77	33	10	198
VADA	51	28	14	10	103
Total	682	335	218	147	1382

General and medical information

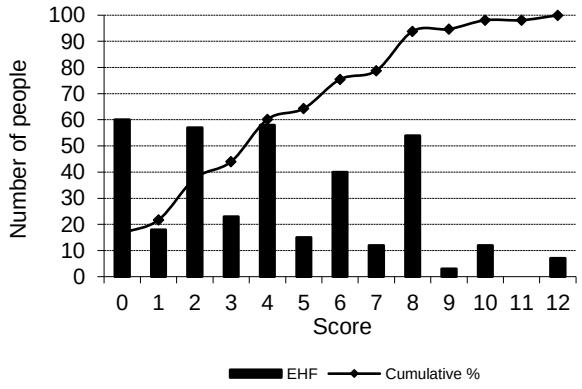
Of the first interviews, 185 were for controls, people without a disabling or stigmatising health condition whose interviews will be used for the tests of discrimination.

There were 497 interviews of people with disability, 496 have gender recorded. The male: female ratio was 1.5:1, with 295 males in the sample. The majority of the people interviewed were affected by leprosy, as expected. Polio was the next most common condition. People classified as 'Other' include amputees and people with post burn contractures, other traumatic injury, blindness, deafness and congenital deformity, amongst others.

A disability grading using the EHF score was recorded for 359 people. Over 80% of people had impairment in at least one eye, hand or foot, and over 50% of people had impairment in more than one eye, hand or foot.

	GENDER	
AGEGRP	F	M	Total
10 to 19	22	19	41
20 to 29	37	52	89
30 to 39	30	66	96
40 to 49	38	60	98
50 to 59	43	55	98
60 to 69	31	37	68
70 to 79	0	6	6
Total	201	295	496

	Health condition					
PLACE	CP	LEPROSY	OTHER	POLIO	SPINAL	STROKE	Total
BRAZIL	5	50	10	5	1	0	71
DAYA	0	34	14	10	0	0	58
KARIGIRI	0	25	13	1	1	0	40
KOLKOTA	0	60	7	3	0	0	70
NAINI	0	99	6	15	0	4	124
NEPAL	0	45	8	0	0	0	53
SALUR	0	30	4	11	1	2	48
VADA	0	24	5	3	1	0	33
Total	5	367	67	48	4	6	497



Social information

Information was collected about work and living situation only. Of the 497 people interviewed who had disease or disability, 184 [37%] were unemployed; the remainder were split between full employment [39%] and part time employment [24%]. Unsurprisingly it was common for people to be working around the house and their land [25%], although over 22% were involved in labouring. This reflects the rural urban split of the sample – 61% of the people came from rural communities.

An estimate of caste or social status was made in all countries. High social status groups accounted for 12%, 35% were from the lowest strata, and 53% were recorded as mid groups. Over half of the people were recorded as living in a nuclear family, with a further 36% living in a joint /extended family. The balance between these two groups is unexpected, since the culture in India and Nepal is to live in a joint family setting. This may be a reflection of the participation restrictions associated with leprosy, or the trends towards migration and urbanisation. Over half of the people are currently married, with a further 20% as yet unmarried. The norm for these people however, is to be married and living in a family setting. Alternatives, such as leprosy communes and nursing homes, along with being unmarried because marriage was not an available option are the situation about 10% of the people.

Scale performance & internal consistency

The scale is designed to be simple to administer, culture free, universally applicable and able to be completed in less than 20 minutes. As such, there should be few if any questions which are not answered or recorded properly. The time taken to administer, when the additional work such as consent and change questions are removed, should be as short as possible. Scale performance was assessed using all data – all levels of interview plus controls.

Questions validity and endorsement

Question	Valid	Irrelevant	Cases=0	Controls=0
a2	98.8	1.2	70.8	94.5
a4	98.7	1.2	51.2	81.7
aa	99.7	0.3	73.9	96.3
bb	100.0	0.0	75.0	95.9
c3	97.3	2.5	68.8	88.5
e1	97.0	2.7	79.5	95.4
f12	99.4	0.6	72.3	90.9
f21	99.0	1.0	60.7	94.1
f8	88.8	8.1	60.7	83.3
g10	97.1	2.6	26.2	80.4
g16	94.4	4.9	35.4	83.9
g24	93.9	5.6	44.4	84.5
h10	91.8	8.0	47.5	64.1
h8	97.6	2.2	47.4	72.4
hh	97.9	2.0	56.3	84.8
j15	98.6	1.4	57.0	83.0
k11	99.9	0.1	74.4	98.2
k9	99.1	0.9	66.6	91.3

Individual questions were recorded as valid, if all the appropriate data was recorded. There remain four questions with a validity of less than 95% [f8, g16, g24, h10]. In all cases, the problem is due to 'irrelevancy' rather than missing data. There are a further four questions which have more than 2% of the people suggesting that the question is irrelevant to them [c3, e1, g10, h8]. Six questions have more than 70% of the people with disability and expected participation restriction saying that they participate normally in this area [a2, aa, bb,e1, f12, k11] There are two questions where more than 20% of the 219 controls [people without a disabling or stigmatising health condition] are recorded as having some restriction [h10,h8]. These results for endorsement and validity of individual questions are deemed acceptable.

Timing

Timing was also reviewed using all data. The median time for administration of the scale was 20 minutes, with a 75th percentile of 25 minutes. Realistically, in order to meet the requirement of a 20-minute administration time, the 75th percentile would need to be 20 minutes or less. There are significant difference between centre [Kruskal Wallis P <0.001], with Brazil and Kolkota recording the lowest median times [9, 13 minutes], and Daya and Salur recording the highest [25, 30 minutes]. This suggests that interviewer technique and training may be a factor in the time taken. The manual to accompany the scale could include hints and tips. There is also a significant difference with number of interview, with the median time for interview 4 being fastest at a median of 15 minutes ,and a 75th percentile of 20 minutes. Again, this supports the suggestion that a 20 minute scale is achievable, but that interview technique can have a significant effect on the time taken. The scale fits into an interview with starting information, consent and closing information. The median time for the whole interview is 35 minutes, but the differences between centres are highly significant.

Scoring and full scale use

A scoring pattern with a small amount of weighting towards the highest response for the question ["it is a big problem"] has been chosen. This choice was made on the basis not only of the felt need not only to recognise that the differences between "no problem", "small problem", "medium problem" and "large problem" are unlikely to be linear in the minds of the people being interviewed, but also to maximise the chance of people highlighting large problems in two or more areas of being graded as 'with participation restriction'.

There are 18 questions in the scale, with a maximum score of 5 for each question. Full participation would therefore have an expected score of 0 and complete restriction has a score of 90. The maximum score recorded in the interviews of cases so far is 90 [100% of the scale length] with a median of 17 [19% of scale length] and an approximate 75th percentile of 39 [43% of scale length]. Although the full scale of the questionnaire is completely used, the data are

skewed and only 5% of the scores lie in the upper 25% of the scale length. This is probably acceptable; it would be expected that the people so far interviewed for the scale development do not yet include many of the very highest level of participation restriction. However if the issue persists in routine use, adaptations to the scoring method could be considered.

Based on the current performance, the following would be recommended as grading based on the scale:

Score	**Description**	
Less than 13	No significant participation restriction	Close to 40% of data below 12
13- 22	Mild restriction	Close to 55% of data below 22
23-33	Moderate restriction	Close to 70% of data below 32
34- 53	Severe restriction	Close to 85% of data below 52
More than 53	Extreme restriction	Less than 15% of data

Internal consistency

For the scale to be functional, all the items in the scale should be measuring the same trait [participation]. Scale consistency is measured by factor analysis, by the use of Cronbach's alpha, and through item-total correlation.

Factor analysis

Factor analysis of the full model [18 questions] suggests 8 factors. However, the first factor accounts for 90% of the variation. A second factor may be represented by questions F8, e1, and k9, although only f8 loads more strongly on the second factor than the first. The second factor may describe a concept such as 'status or 'self image'. The model is therefore adequate.

Internal consistency – Cronbach's alpha

The alpha value for the 18 question scale is 0.92, indicating high internal consistency. The target value for alpha remains debatable; an alpha over 0.9 may indicate question redundancy. It is possible that the high value for alpha indicates too high an internal consistency, that is, possibilities not only of redundancy, but also of a lack of breadth in the focus of the instrument.

Internal consistency – item total correlation

All the questions were correlated with the total score minus the score for that question, to obtain the item total correlation for that question.

All the correlations showed a significant correlation, with F8, E1 and K9 showing the weakest correlation.

Question	**n**	**Correlation**	**Question**	**n**	**Correlation**
A2	**494**	**0.715**	**A4**	**496**	**0.695**
AA	**497**	**0.701**	**BB**	**497**	**0.569**
C3	**489**	**0.554**	**E1**	**491**	**0.400***
F12	**497**	**0.620**	**F21**	**492**	**0.710**
F8	**477**	**0.327***	**G10**	**493**	**0.579**
G16	**483**	**0.520**	**G24**	**485**	**0.569**
H10	**492**	**0.725**	**H8**	**493**	**0.673**
HH	**495**	**0.739**	**J15**	**497**	**0.711**
K11	**495**	**0.522**	**K9**	**496**	**0.428***

Inter-rater

Of the 335 second interviews done, 296 have a first interview which can be matched by the computer. There are 177 men [60%]. Leprosy related disability accounts for 75% of the paired data. All age groups and all EHF scores are represented.

PLACE	Freq	Percent	Cum.
BRAZIL	33	11.1%	11.1%
DAYA	38	12.8%	24.0%
KARIGIRI	32	10.8%	34.8%
KOLKOTA	46	15.5%	50.3%
NAINI	46	15.5%	65.9%
NEPAL	31	10.5%	76.4%
SALUR	47	15.9%	92.2%
VADA	23	7.8%	100.0%
Total	296	100.0%	

Scale score

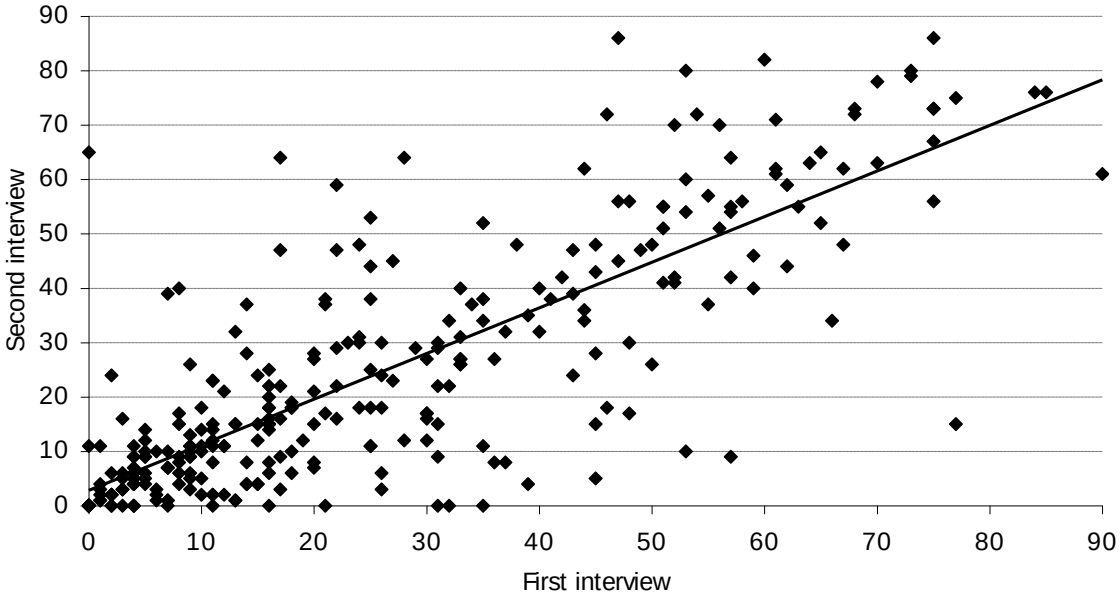
The scale score was calculated for each interview separately. Some scale responses have missing information for one or more questions. Arbitrarily this has been coded zero, that is, a question with a missing response contributes nothing to the overall score, but does not stop the score being calculated. The totals scores have been compared by the method of ICC [using a two-way analysis of variation without replication], assuming that the paired interviewers are a random sample of all possible pairings. There are 44 second interviewers in eight centres, with 1-21 interviews recorded for them, so the assumption that the interviewer pairs are representative is likely to be valid.

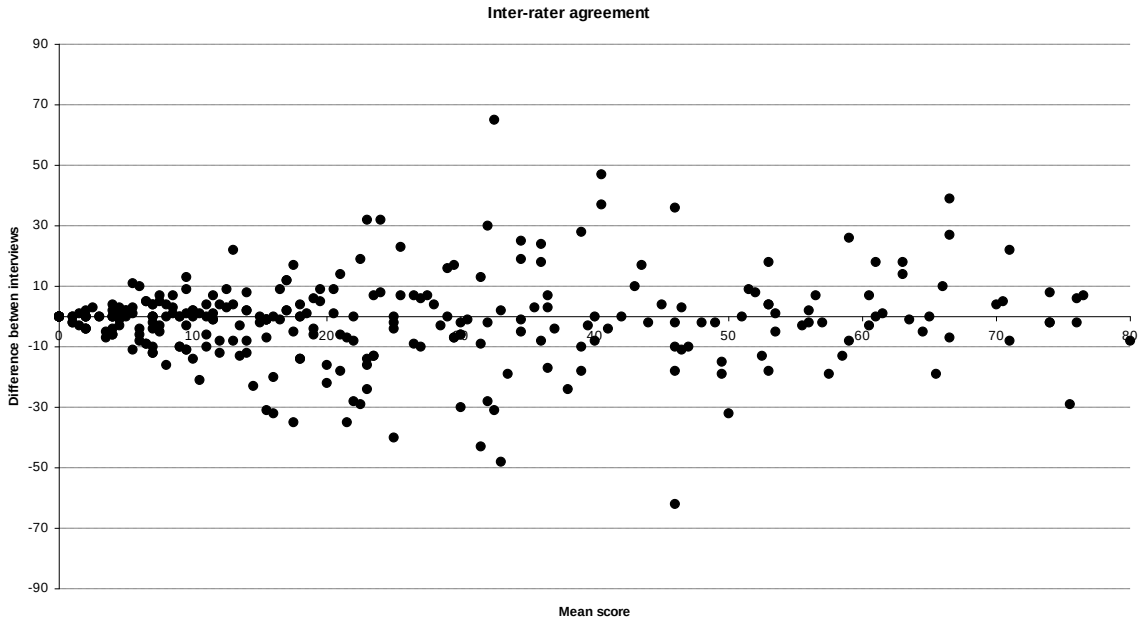
For the first interview, the median score was 20 [out of a possible 72] with a range of 0 to 90. For the second interview, the median was 16.5 with a range of 0 to 86. The reliability quotient is 0.80, which can be interpreted as ‘81% of the variance in the score results from ‘true’ variance among clients’. This suggests good reliability despite the fact that some individual differences are large.

Extreme observations

There are 17 paired observations with a difference in the score, between interviews, of 30 points or more

REC	PLACE	STUDYNUMBE	CLINICNUMB	DISEASE	ITOT	TOTAL
62	BRAZIL	31009	11587	LEPROSY	34	66
83	BRAZIL	31081	11569	LEPROSY	59	22
208	DAYA	23039	039	LEPROSY	40	8
205	DAYA	23043	043	LEPROSY	4	39
461	KARIGIRI	32005	0001	LEPROSY	0	31
477	KARIGIRI	32021	0007	LEPROSY	65	0
324	KOLKOTA	2091	NIL	POLIO	64	17
353	NAINI	24020	1638/02	LEPROSY	15	77
368	NAINI	24035	1252/01	LEPROSY	10	53
386	NAINI	24053	2083/01	LEPROSY	39	7
9	NEPAL	15009	PFR	LEPROSY	0	32
32	NEPAL	15032	PFR	LEPROSY	64	28
136	SALUR	26012	1503	LEPROSY	9	57
140	SALUR	26016	1621	LEPROSY	17	48
164	SALUR	26045	5610	LEPROSY	5	45
240	VADA	28012	28012	LEPROSY	86	47
254	VADA	28026	28026	LEPROSY	0	35



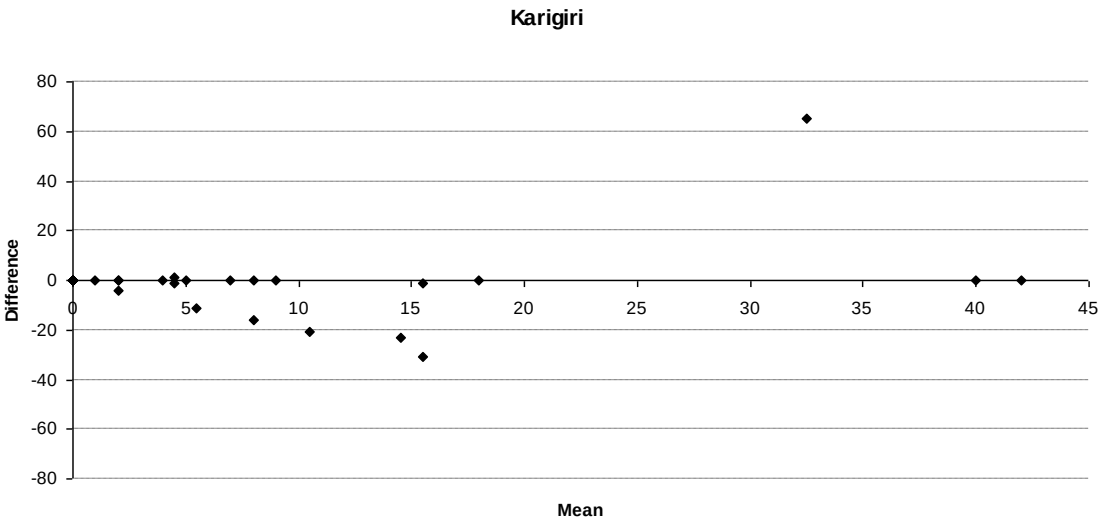


By centre

Dividing the data by centre reduces the confidence in the estimate of the reliability quotient, but shows differences between the centres.

Centre	Brazil	Daya	Karigiri	Kolkota	Naini	Nepal	Salur	Vada
Number of pairs	33	38	32	46	46	31	47	23
Reliability	0.71	0.67	0.42	0.73	0.79	0.83	0.72	0.22

The comparably poor robustness to changes in interviewer in Karigiri, and the unusual spread of differences, which do not appear random, needs further investigation.



Individual questions

Since the result for an individual question is categorical – with just five potential categories, the reliability of individual questions can be checked using Cohen's Kappa [weighted]. Missing data are treated as missing, hence the difference in numbers of paired observations for each question. The observed reliability coefficients for each question are moderate to good [range 0.49-0.73] according to the categorisation of Altman. This is to be expected, the reliability of the whole scale should be better than the reliability of the individual items.

Question	N	Kappa	SE of Kappa	Actual agreement
G16	283	0.734	0.059	89%
G10	291	0.614	0.058	87%
G24	284	0.616	0.059	84%
A4	295	0.687	0.058	89%
F21	295	0.581	0.058	87%
H8	293	0.601	0.058	88%
HH	294	0.563	0.058	87%
J15	296	0.586	0.058	88%
K9	295	0.536	0.058	86%
K11	296	0.584	0.058	87%
AA	296	0.617	0.058	89%
A2	296	0.674	0.058	90%
BB	296	0.508	0.058	85%
F12	295	0.646	0.058	91%
E1	287	0.487	0.059	88%
F8	270	0.575	0.060	90%
H10	291	0.624	0.059	86%
C3	292	0.569	0.058	89%

By centre

The reliability of individual questions by centre is affected by the relatively small numbers of paired observations for each centre and the prior distribution of results. Accepting this, there is some poor reliability [Kappa <0.4] of some questions, particularly in some centres.

	Overall Kappa	Brazil	Nepal	Naini	Salur	Kolkota	Karigiri	Vada	Daya	Number of low reliability centres
G16	0.734	0.498	0.651	0.724	0.822	0.514	0.366	0.793	0.726	1
G10	0.614	0.449	0.584	0.228	0.211	0.638	0.585	0.732	0.615	2
G24	0.616	0.128	0.180	0.662	0.507	0.500	0.742	0.861	0.715	2
A4	0.687	0.205	0.551	0.727	0.802	0.529	0.565	0.714	0.761	1

F21	**0.581**	**-0.116**	**0.644**	**0.667**	**0.550**	**0.286**	**0.494**	**0.409**	**0.523**	**2**
`H8	**0.601**	**0.363**	**0.538**	**0.676**	**0.364**	**0.406**	**0.389**	**0.353**	**0.752**	**4**
HH	**0.563**	**0.339**	**0.359**	**0.774**	**0.315**	**0.382**	**0.560**	**0.244**	**0.644**	**5**
J15	**0.586**	**0.267**	**0.589**	**0.667**	**0.687**	**0.208**	**0.745**	**0.395**	**0.381**	**4**
K9	**0.536**	**0.458**	**0.505**	**0.595**	**0.441**	**0.171**	**0.884**	**0.259**	**0.483**	**2**
K11	**0.584**	**0.228**	**0.506**	**0.721**	**0.455**	**0.612**	**0.694**	**0.632**	**0.240**	**2**
AA	**0.617**	**0.589**	**0.188**	**0.664**	**0.643**	**0.521**	**-0.032**	**0.614**	**0.732**	**2**
A2	**0.674**	**0.626**	**0.589**	**0.636**	**0.662**	**0.514**	**0.513**	**0.651**	**0.866**	**0**
BB	**0.508**	**0.142**	**0.336**	**0.597**	**0.262**	**0.020**	**0.311**	**0.905**	**0.972**	**5**
F12	**0.646**	**0.105**	**-0.080**	**0.702**	**0.511**	**0.474**	**0.837**	**0.875**	**0.777**	**2**
E1	**0.487**	**0.746**	**0.402**	**0.617**	**0.116**	**0.343**	**0.152**	**0.469**	**0.854**	**3**
F8	**0.575**	**0.602**	**0.542**	**0.606**	**0.477**	**0.302**	**0.647**	**0.823**	**0.860**	**1**
H10	**0.624**	**0.840**	**0.686**	**0.508**	**0.390**	**0.599**	**0.414**	**0.490**	**0.439**	**1**
C3	**0.569**	**0.351**	**0.538**	**0.722**	**0.386**	**0.573**	**0.475**	**0.372**	**0.663**	**3**
Number of low reliability questions		**10**	**5**	**1**	**7**	**7**	**5**	**5**	**2**	

Stability

PLACE	Freq	Percent	Cum.
BRAZIL	15	7.1%	7.1%
DAYA	45	21.4%	28.6%
KARIGIRI	29	13.8%	42.4%
KOLKOTA	34	16.2%	58.6%
NAINI	29	13.8%	72.4%
NEPAL	15	7.1%	79.5%
SALUR	29	13.8%	93.3%
VADA	14	6.7%	100.0%
Total	210	100.0%	

For assessment of stability, using the same interviewer there are 210 matching paired interviews. There are 84 men [52%]. Leprosy related disability accounts for 72% of the paired data. All age groups and all EHF scores are represented.

Scale score

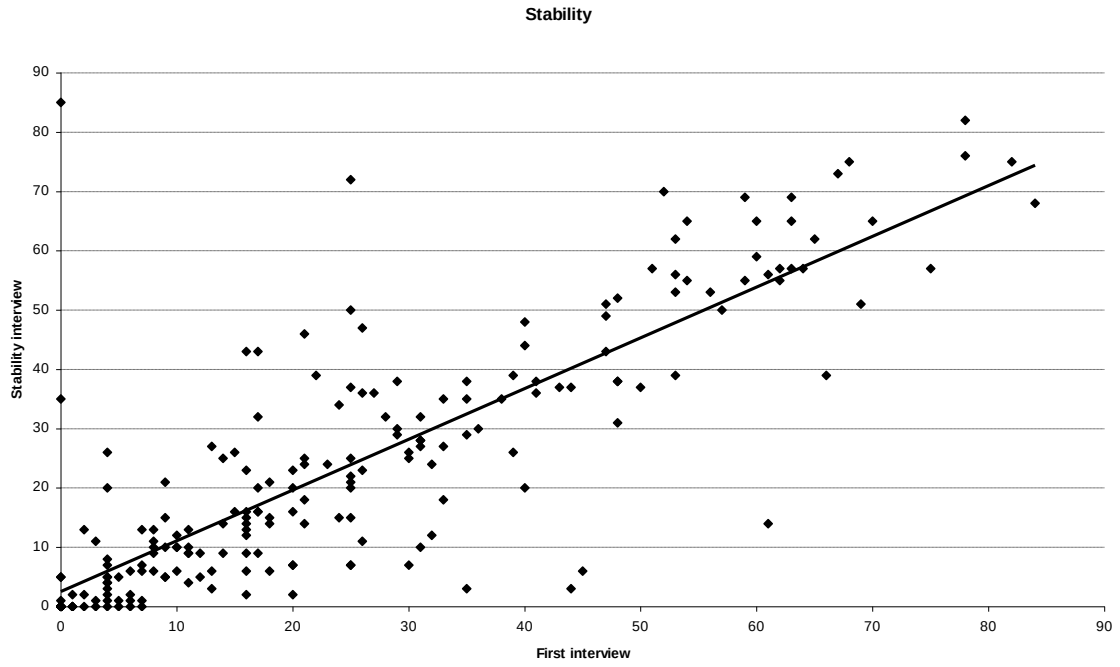
As for the inter-rater, the scale score was calculated for each interview separately and the scale responses with missing information have been coded zero. The total scores have been compared by the method of ICC [using a two-way analysis of variation without replication].

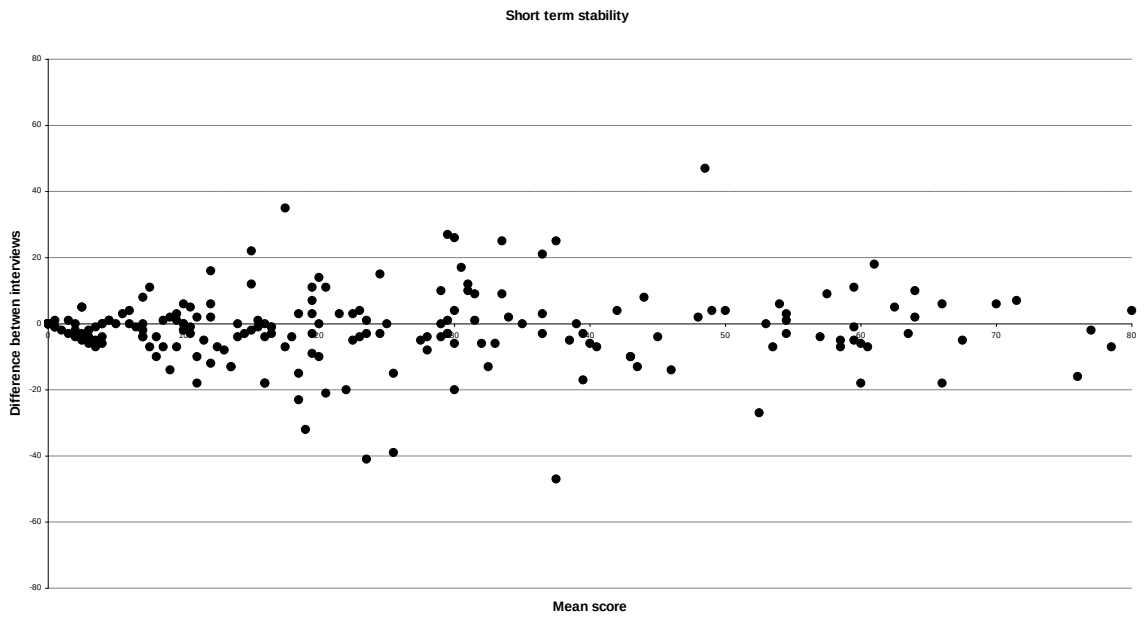
For the first interview, the median score was 19 [out of a possible 90] with a range of 0 to 84. For the stability interview, the median was 16 with a range of 0 to 85. The reliability quotient is 0.83, suggesting good reliability despite the fact that some individual differences are large.

Extreme observations

There are seven paired observations with a difference in the score, between interviews, of 30 points or more

REC	PLACE	STUDYNUMBE	CLINICNUMB	DISEASE	STOT	TOTAL
59	BRAZIL	31006	11811	LEPROSY	72	25
477	KARIGIRI	32021	0007	LEPROSY	85	0
484	KARIGIRI	32027	0011	OTHER	35	0
145	SALUR	26021	1976	LEPROSY	6	45
233	VADA	28003	02/5330	LEPROSY	3	44
254	VADA	28026	28026	LEPROSY	3	35
255	VADA	28027	28027	LEPROSY	14	61





By centre

Dividing the data by centre reduces the confidence in the estimate of the reliability quotient substantially, especially in Nepal and Brazil where only a few interviews were completed, but shows differences between the centres.

Centre	Brazil	Daya	Karigiri	Kolkota	Naini	Nepal	Salur	Vada
Numberof pairs	15	45	29	34	29	15	29	14
Reliability	0.62	0.83	0.20	0.86	0.95	0.87	0.82	0.54

Question	N	Kappa	SE of Kappa	Actual agreement
G16	201	0.682	0.071	87%
G10	208	0.675	0.069	88%
G24	201	0.657	0.071	86%
A4	209	0.712	0.069	90%
F21	208	0.688	0.069	91%
H8	206	0.626	0.069	89%
HH	209	0.575	0.069	88%
J15	210	0.537	0.069	87%
K9	208	0.592	0.069	89%
K11	209	0.570	0.069	86%
AA	210	0.606	0.068	91%
A2	209	0.713	0.069	92%
BB	210	0.726	0.067	92%
F12	210	0.679	0.069	92%
E1	205	0.571	0.070	91%
F8	200	0.623	0.070	93%
H10	208	0.645	0.069	86%
C3	203	0.638	0.070	90%

The comparably poor robustness in Karigiri and Vada needs further investigation as, without significant life-change between occasions, it would be expected that the robustness within interviewer [stability] would be better than seen between interviewers [inter-rater]. Although the reliability in Brazil is less than the inter-rater agreement, the difference is not great.

Individual questions

Since the result for an individual question is categorical – with just 5 potential categories, the reliability of individual questions can be checked using Cohen's Kappa [weighted]. Missing data are treated as missing, hence the difference in numbers of paired observations for each question. The observed reliability coefficients for each question are moderate to good [range 0.54-0.73] according to the categorisation of Altman. This is to be expected, the reliability of the whole scale should be better than the reliability of the individual items, and the reliability within interviewers should be better than that between interviewers, unless the scale is measuring short-term change.

By centre

The reliability of individual questions by centre is affected by the relatively small numbers of paired observations for each centre and the prior distribution of results, especially in Nepal where there were only five paired interviews. Accepting this, there is some poor reliability [Kappa<0.4] of some questions, particularly in some centres.

	Overall Kappa	Brazil	Nepal	Naini	Salur	Kolkota	Vada	Daya	Number of low reliability centres
G16	0.682	0.263	0.237	0.966	0.618	0.644	0.472	0.683	1
G10	0.675	0.257	0.636	0.550	0.550	0.773	-0.085	0.574	2
G24	0.657	0.492	0.208	0.862	0.576	0.602	0.604	0.662	1
A4	0.712	0.400	0.495	0.908	0.727	0.810	0.510	0.692	0
F21	0.688	-0.119	0.585	0.930	0.661	0.508	0.302	0.854	2
H8	0.626	0.590	0.602	0.720	0.256	0.410	0.024	0.756	2
HH	0.575	0.340	0.328	0.749	0.505	0.462	0.027	0.664	3
J15	0.537	0.207	0.158	0.765	0.692	0.075	0.407	0.727	3
K9	0.592	0.276	0.767	0.583	0.492	0.522	0.237	0.135	3
K11	0.570	0.579	0.429	0.444	0.445	0.841	0.308	0.600	1
AA	0.606	0.368	0.281	0.727	0.439	0.342	0.408	0.742	3
A2	0.713	0.651	0.762	0.775	0.689	0.801	0.625	0.589	0
BB	0.726	0.459	0.526	0.841	0.808	0.560	0.413	0.900	0
F12	0.679	0.000	-0.071	0.810	0.685	0.337	0.607	0.850	3
E1	0.571	0.100	0.822	0.663	0.506	0.399	0.654	0.804	1
F8	0.623	0.309	0.563	0.504	0.680	0.749	0.387	0.702	1
H10	0.645	0.683	0.619	0.390	0.504	0.392	0.307	0.602	3
C3	0.638	0.214	0.549	0.721	0.511	0.733	0.768	0.358	2
Number of low reliability questions		9	6	1	1	4	8	2	

Discrimination

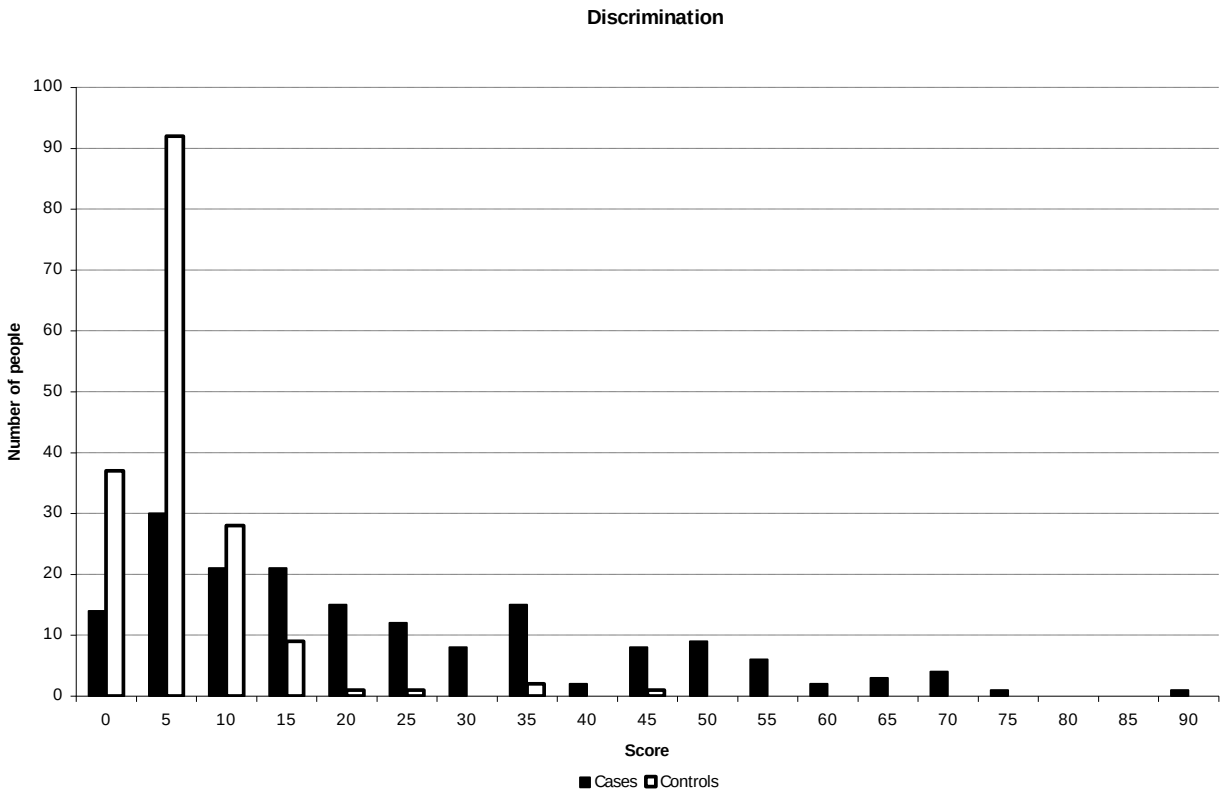
For the scale to be functional, it must be able to discriminate between people with and without participation restriction, and to discriminate between people with different levels of participation restriction. Since there is no 'gold standard' assessment method for participation, discrimination is mainly assessed through external validity tests [see below]. However, people without disability are by definition unable to have participation restrictions *due to their disability.*

As part of the study, 185 people without disability or stigmatising disease, but part of the communities of the other interviewees, were interviewed on one occasion. The scores for these people were compared with the scores for the first interview for the people with disability.

Equivalence of groups

Demographic indicators were compared between the groups, to ensure that the control group had been appropriately selected to be similar to the people with disability. There was no association of case/control with gender [p=0.187], location [p=0.09].There was an association with age [Controls younger, p=0.03], being in work [Controls more likely, p<0.001], Caste/social status [controls higher, p<0.001], Living situation [Cases living alone, p<0.001], Marital status, [cases unmarried, p<0.001].

None of these associations is unexpected; however, they are all potential confounding variables in any assessment of discrimination. Accordingly, matched sample of cases and controls were selected for the test of discrimination. In each centre a case was selected for every control, matching on the seven parameters above. Where a case matched on a least five of the seven, a single case was selected using random numbers. Controls where no match could be found on at least five parameters were rejected. A matched case was found for 171 controls; testing for associations showed that all except the association with work [p<0.001] had been removed.



Comparison of scores

The median score in the controls was 2 [range 0-44], compared with a median for the cases of 15 [range 0-90].This difference is highly significant [p<0.001]. The approximate 95th percentile of the scores for the controls is 12. It would be logical to use this as a cut-off for the score that indicates disability related participation restriction. The effect of this would be to have to assume that people with disability, who do not score over 12 on the scale, have minimal if any disability related participation restriction. This cut-off and assumption will be reviewed during the tests for validity.

Using the cut off on this data set, 44% of cases score of 12 or less, and would therefore be classed as without participation restriction. On the full [unmatched] dataset, this is 40% [n=497]

Stratifying the data by the three categories of work, shows a similar pattern.

	Cases		**Controls**	
	N	**Median**	**n**	**Median**
OVERALL	**171**	**15**	**171**	**2**
NONE	**53**	**21**	**23**	**4**
PART	**30**	**10**	**23**	**1**
FULL	**88**	**16**	**125**	**2**

In addition, there are differences in median scores between centres:

	Case		**Control**	
	N	**Median**	**N**	**Median**
Brazil	**32**	**12**	**32**	**7**
Daya	**16**	**9.5**	**16**	**1**
Karigiri	**14**	**2**	**14**	**0**
Kolkota	**30**	**8**	**30**	**4**
Naini	**26**	**33**	**26**	**1**
Nepal	**12**	**30**	**12**	**2**
Salur	**26**	**24**	**26**	**1**
Vada	**15**	**23**	**15**	**1**

It would be recommended that centres check the validity of the cut-off in their own target population.

External validity

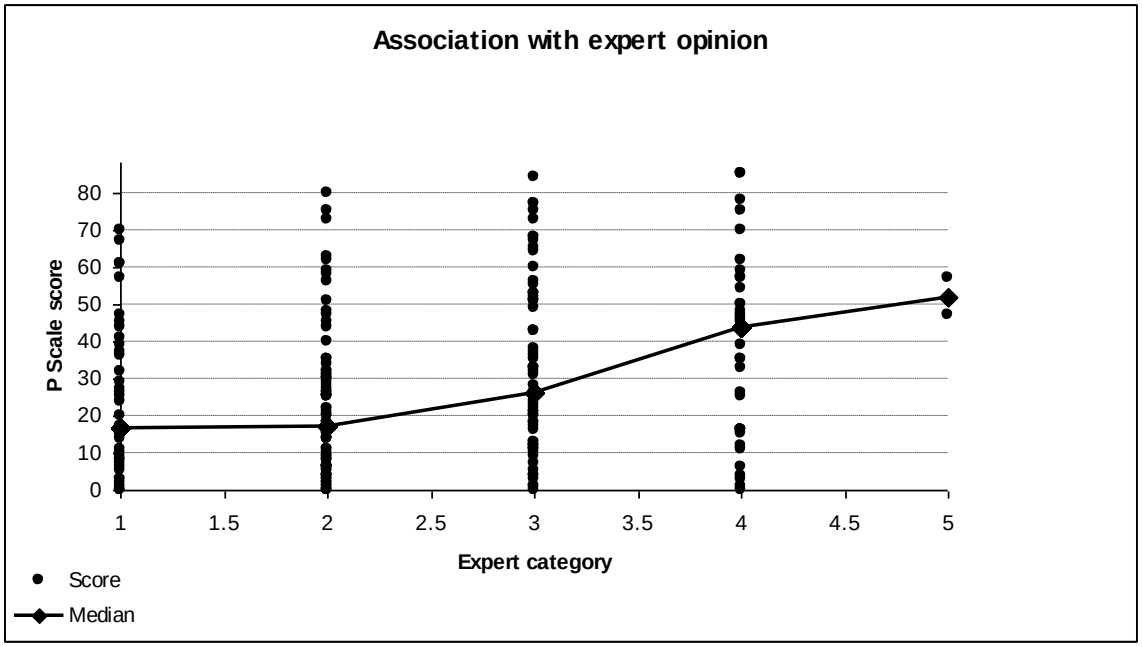
The scale has been designed to measure participation restriction. There is no 'gold standard' method that can be used in this context to check whether the scale is actually measuring participation. In order to check the validity, the following hypotheses were proposed:

- The scale score should be significantly associated with scores given by individual social workers / interviewers/ socio economic rehabilitation experts who know about the concept of participation but are using their own, varied, methods of assessment.
- The scale score may be associated with a self-assessment of 'how my life is?' based on a Likehart scale, but failure to correlate will not mean that the P Scale is invalid.
- The scale score may be associated with Body mass index [BMI]. BMI has been proposed as a simple indicator of need for socio-economic rehabilitation, but this may not reflect participation fully.
- The scale score will be associated with impairment status, although this will not be a perfect association. A perfect association would suggest that the scale is measuring impairment.
- The scale should not correlate strongly with age, gender, urban/rural status or social status [caste], as a strong association might indicate that the scale was measuring these rather than participation.

Expert assessment

The expert assessment gave an output in five categories 1-5 [none, mild, moderate, severe, complete participation restriction]. For 227 baseline interviews there was a summary opinion from an expert assessor. Just two were graded 5. Non-parametric methods were used to compare the median score for each category.

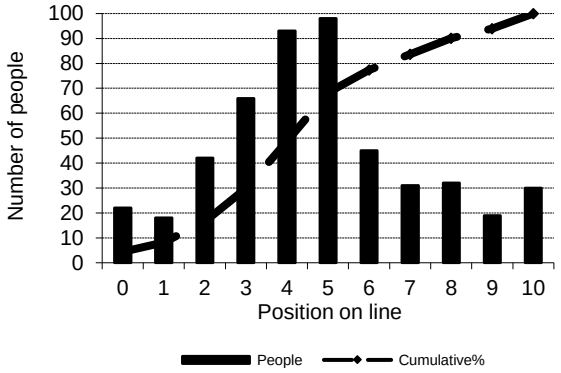
The median score was significantly different between the groups, with medians ranging from 16.5 in the group with 'no



participation restriction' to 52 in the group with 'Severe participation restriction' [p=0.001, Kruskal Wallis test]. The data suggest an association between score and expert opinion grouping, as hypothesised.

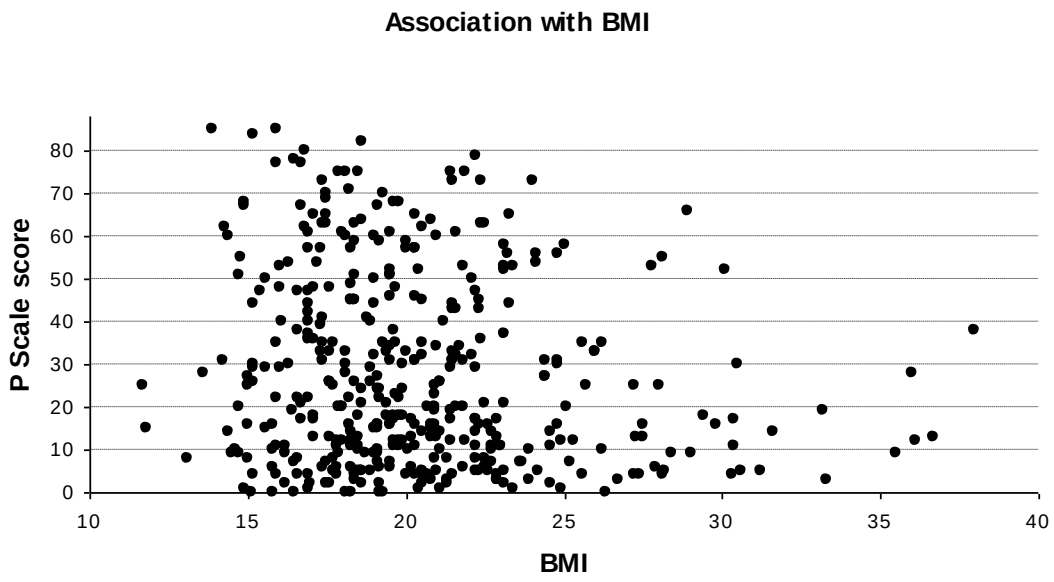
Association with self-assessment

The interviewees were asked to make a mark on a line to show their own opinion of how their life was at the time of interview, with 0 being the negative end of the scale and 10 being the positive end. Despite the use of numbers as indicators, the intervals between the numbers have no validated meaning, and therefore the data cannot be assumed to be continuous. This concept of self-assessment has not been tested in this context; the data generated could be analysed to check the robustness of this method. During baseline interviews, 496 ‘cases’ recorded their opinion. There is some evidence of overuse of the end-points [0,10],and anecdotal evidence that people found the concept very difficult. However, there is evidence of a significant [negative] association with the total score from the P Scale [p<0.001 Kruskal Wallis].



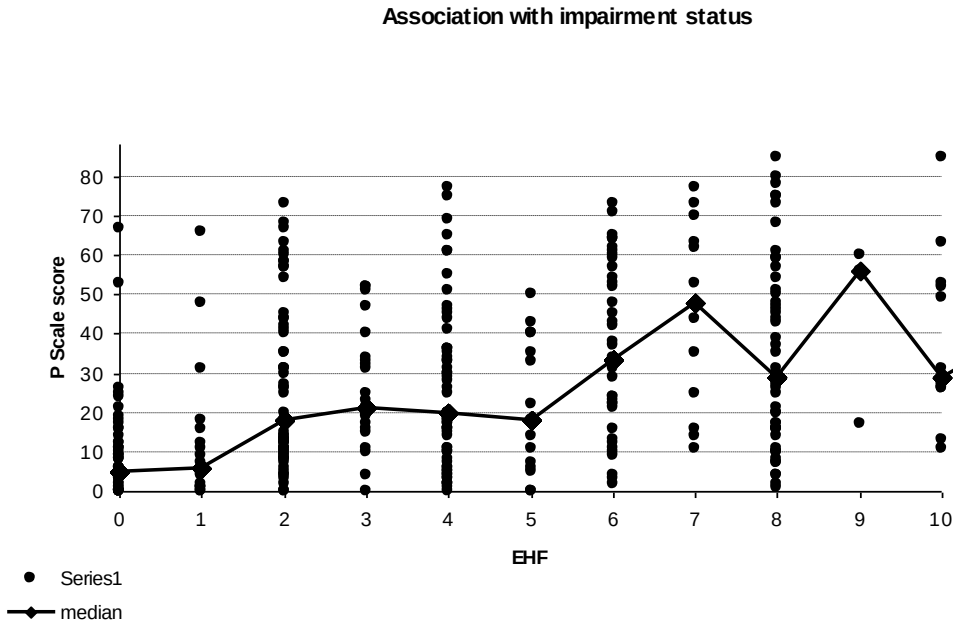
Association with BMI

There has been a recent report in the literature that Body Mass Index [BMI] might be a suitable, simple indicator of the need for socio-economic rehabilitation. Height and weight of the clients were recorded for the baseline interviews of cases and the BMI calculated. The correlation with the total score from the P scale was significant, given the number of observations [458], but very small at -0.09. It is probably a reflection of the small number of very high BMI values; these are not currently verifiable.



Association with impairment status

For all people affected by leprosy in the study, an impairment grading was made using the Eye-hand-foot system. This gives a score in the range 0-12. This score was further grouped to four categories – none [EHF=0], mild [EHF=1-3] Moderate [EHF=3-6] or severe [EHF = 7 or more]. Association with both the raw score and the categorised score was strong [p<0.001], and there is a clear, if imperfect, trend to higher P scale scores with increasing impairment.



Association with age and gender

Age and gender were recorded for each person with disability giving a baseline interview. The median P scale score for males was 19, whilst for females it was 18. This difference is not significant at the 95% level [p=0.58, Kruskal Wallis].

AGEGRP	Minimum	Median	Maximum
10 to 19	0.000	13.000	55.000
20 to 29	0.000	18.500	79.000
30 to 39	0.000	13.000	75.000
40 to 49	0.000	20.000	90.000
50 to 59	0.000	23.000	82.000
60 to 69	1.000	27.500	85.000
70 to 79	2.000	43.000	73.000

For age group, the median scores different between the decades with the scores rising as people age [p=0.001 Kruskal Wallis]. This effect persists even when the small group of people over 70 years, who were originally to be excluded from the study, are excluded from the analysis.

Association with social conditions

An assessment of social status [caste in India and Nepal, 3 categories] and location [urban/rural] shows a significant association at the 95% level with status [p=0.02]. The median P scale score in the high status group is higher than that in the mid and low status groups, [30 compared with 17.5 & 17]. The numbers of people in the high status group is comparatively small [57 compared with 266 & 174]. The median score in the rural communities was 25.5 compared with 14 in the urban centres. This difference is again, significant [p<0.001].

Dynamicity

For assessment of dynamicity, clients were re-interviewed after an elapsed time in which a significant life change had occurred. The underlying hypothesis was that the life change would affect participation, and that the scale score should show a significant positive or negative change. A total of 121 people had paired interviews before and after a life change, and including an elapsed time of at least 3 months. A further 6 [Karigiri] were interviewed without a known life change.

There are 79 men [62%]. Leprosy related disability accounts for 78% of the paired data. All age groups are represented.

For only 67 people is the nature of the life change currently known. Of these the majority are either family issues [births, marriages and deaths] or participation in income generation schemes. This is expected given the population from which participating centres were drawing interviewees. The median time between interviews was 270 days [approx 9 months] with a range of 82-422 days. Interviews collected less than 80 days from the first interview were excluded.

PLACE	Freq	Percent	Cum.
BRAZIL	21	16.5%	16.5%
KARIGIRI	13	10.2%	26.8%
KOLKOTA	19	15.0%	41.7%
NAINI	22	17.3%	59.1%
NEPAL	32	25.2%	84.3%
SALUR	10	7.9%	92.1%
VADA	10	7.9%	100.0%
Total	127	100.0%	

CHANGE	Freq	Percent	Cum.
FAMILY	14	20.9%	20.9%
HOSPITAL	5	7.5%	28.4%
HOUSING	9	13.4%	41.8%
IGP	13	19.4%	61.2%
PROSTHESIS	2	3.0%	64.2%
SHG	4	6.0%	70.1%
STIGMA	1	1.5%	71.6%
SURGERY	7	10.4%	82.1%
TRAINING	9	13.4%	95.5%
WORK	3	4.5%	100.0%
Total	67	100.0%	

Scale score

The scale score was calculated for each interview separately and the scale responses with missing information have been coded zero. The difference between the paired scores was calculated, ignoring the direction of the change.

Score differences

The mean difference between scores was 12.6 points [ignoring the direction of change]. This difference is highly statistically significant [p <0.001]. However, the measure of agreement from inter-rater and stability testing is such that 54% of paired interviews which are expected to give the same score give scores differing by up to +/-10 points. Only differences above this level should be considered a significant change in terms of dynamicity. On this basis, 58% of the paired interviews for dynamicity do not show a difference which could be assumed to be significant.

Change		
0-5 points	78	15%
5-10	193	54%
10-15	105	74%
15-20	43	83%
20-25	35	90%
25-30	16	93%
30-35	12	95%
35-40	10	97%
40-45	5	98%
45-50	2	99%
50-55	4	99%
55-60	0	99%
60-65	0	99%
65-70	2	100%
70-75	0	100%
75-80	0	100%
80-85	0	100%
85-90	1	100%

Score categories

Based on the categorisation into five bands [no significant restriction, mild, moderate, severe and extreme restriction], 51 people [40%] show a significant change which results in a change of category.

Direction of change

For the 67 people where some information about the life change is known, and using a broad categorisation of life changes into ‘positive’, ‘negative’ and ‘both positive and negative’, the appropriateness of the direction of change was investigated.

Clients with life changes which appeared to include positive and negative elements [n=11] showed a median difference in score of 10 points [worse participation] [range -22-+27]. In 7 [64%] people there was change substantial [more than 10 points difference and a category change]. Of these four showed worse participation and three, better participation. For clients with a broadly positive experience [eg income generation schemes, housing and self help groups], [n=43], the median difference was -8 points [increased participation] [range -51- +32]. In 20 [47%] the change was substantial. In all but four cases, the change was positive. For people with an experience expected to be negative [death in the family,

stigmatisation, hospital admission], [n=13] the median change was -7 points [increased participation] [range -19-+29]. Of these, eight show a substantial change, but and six of these are in a positive direction.

Conclusion

There is a statistically significant difference in scores between baseline data and post life-change data. In 42% of people this difference is greater than would be expected from variability in the interview. However, the direction of change does not consistently reflect the expectation from a gross categorisation of life changes.

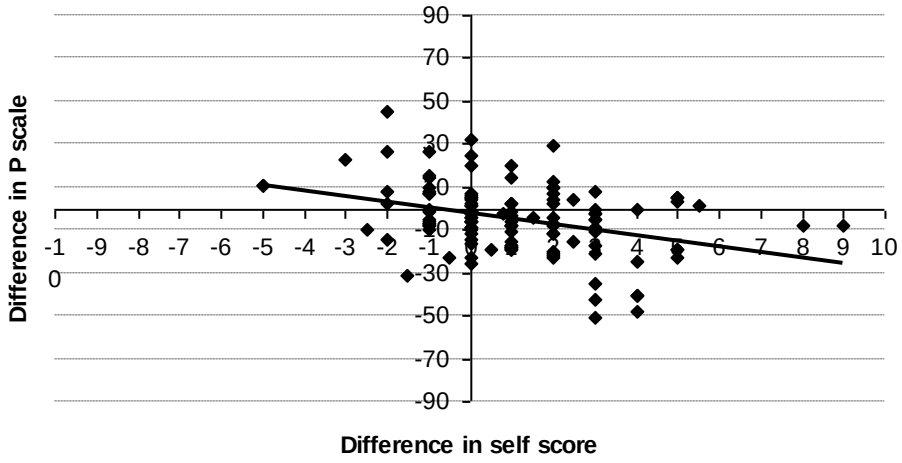
Comparison with self assessment

Each interview concluded with a self assessment of ‘how life is now?’ on a line marked from 0-10 with 1`0 as the most positive score. As a measure of external validity, the self assessment showed a significant negative association with the score from the P scale.

In self assessment 54 % of people indicated a change of -1 to +1. An analysis of the expected occasion to occasion variation has not been performed, but it is probable that +/-1 lies within these limits. These observations support the conclusion that the failure to see more people with a significant score change is not a reflection of the scale but of the life-change experienced.

Of these 70 people, 45 [64%] show no substantial change in P scale score [change less than 10 points or not across a category]

The difference in self-assessment was compared with the difference in scale score. Again, the data show a significant negative association, although the scatter in the data is wide [slope -2.5]



Recommendations for monitoring use

The data suggest that there is an appropriate association of change in scale score with life change. The observation that over 50% of the people experience a change is score which unlikely to be greater than the expected interview to interview variation is consistent with the self assessment where 54% of people indicated a change of -1 to +1. It is possible that the 9 month median duration of the dynamicity assessment is too short to successfully measure many participation changes. It is possible too, that the life changes that are broadly categorised as positive or negative do not have the same clear-cut impact on participation restriction as the categorisation would suggest. In addition, it is possible that we have an expectation of a greater impact of an intervention on a person’s social participation than is realistic.

Screening tool

A screening tool for population surveys has been requested. A screening tool should have properties such that it detects participation restriction in the majority of the people for whom the full scale would have detected restriction [high sensitivity]. It is generally less important whether people without participation restriction are also shown as having restriction [specificity], as these people will generally be excluded at a later stage in an intervention.

Along with the characteristics of sensitivity and specificity are 'positive and negative predictive value' [ppv, npv]. If a majority of the people assessed by the tool to have participation restriction in screening are found to have restriction with subsequent testing, this is a high positive predictive value.

The balance between sensitivity and specificity, PPV and NPV is normally a local issue, depending on the resource available for second line judgements, and the sensitivity of making a judgement about a person which later proves to be false amongst other issues. However, in all circumstances, a screening tool should probably aim at a minimum performance of a sensitivity of around 90%, to ensure that people are not being missed, and a PPV of around 80% to ensure that resources are not wasted on people who do not need them. These limits are, of course, arbitrary.

Yes/No configuration

The full scale has 5 possible answers [0-4 score] and a supplementary question for each score over 0. A possible simplification resulting in significant time savings is to convert the scale to Yes/No. The questions remain the same, "Do you…". An answer of Yes scores 0, and answer of No scores 1. The cut-off is 'more than X questions answered "No"'. The interview can terminate when the limit has been exceeded.

Based on this configuration, the screening tool can be compared with the full scale [at a cut-off of 12 points for participation restriction] for performance

Cut off	PPV%	NPV%	Sensitivity%	Specificity%	Wrong judgement
More than 4	82.1	92.1	92.3	81.8	90/682
More than 5	87.2	87.6	86.1	88.5	86/682
More than 6	89.3	81.8	77.5	91.6	103/682
More than 7	93.0	77.7	69.8	95.3	115/682

Four cut-offs are close to meeting the needs of a screening tool, in the range 4-7 points on a YES/NO Scale. The choice between them probably depends on the balance of need in individual circumstances.

Question reduction

Question reduction without a major loss of sensitivity and specificity is likely to be possible. However the choice has been made to keep the screening tool the same as the full scale.