The five item Barthel index

J C Hobart, A J Thompson

Abstract

***Objectives*—Routine data collection is now considered mandatory. Therefore, staff rated clinical scales that consist of multiple items should have the minimum number of items necessary for rigorous measurement. This study explores the possibility of developing a short form Barthel index, suitable for use in clinical trials, epidemiological studies, and audit, that satisfies criteria for rigorous measurement and is psychometrically equivalent to the 10 item instrument.**

***Methods*—Data were analysed from 844 consecutive admissions to a neurological rehabilitation unit in London. Random half samples were generated. Short forms were developed in one sample (n=419), by selecting items with the best measurement properties, and tested in the other (n=418). For each of the 10 items of the BI, item total correlations and effect sizes were computed and rank ordered. The best items were defined as those with the lowest cross product of these rank orderings. The acceptability, reliability, validity, and responsiveness of three short form BIs (five, four, and three item) were determined and compared with the 10 item BI. Agreement between scores generated by short forms and 10 item BI was determined using intraclass correlation coefficients and the method of Bland and Altman.**

***Results*—The five best items in this sample were transfers, bathing, toilet use, stairs, and mobility. Of the three short forms examined, the five item BI had the best measurement properties and was psychometrically equivalent to the 10 item BI. Agreement between scores generated by the two measures for individual patients was excellent (ICC=0.90) but not identical (limits of agreement=1.84±3.84).**

***Conclusions*—The five item short form BI may be a suitable outcome measure for group comparison studies in comparable samples. Further evaluations are needed. Results demonstrate a fundamental difference between assessment and measurement and the importance of incorporating psychometric methods in the development and evaluation of health measures.**

(*J Neurol Neurosurg Psychiatry* 2001;**71**:225–230)

Keywords: five item Barthel Index; psychometric methods; health measurement; item reduction

Department of Clinical Neurology and Neurorehabilitation, Neurological Outcome Measures Unit, Institute of Neurology, Queen Square, London WC1N 3BG, UK
J C Hobart
A J Thompson

Correspondence to:
Dr J Hobart
J.Hobart@ion.ucl.ac.uk

Received 21 September 2000
and in final form
22 March 2001
Accepted 28 March 2001

Routine data collection for audit purposes is now considered mandatory. In addition, effectiveness studies and large multicentre trials are dependent on outcomes data collection being integrated into daily clinical practice. This formidable task is compounded by a requirement to supplement traditional health indicators that can be collected easily (for example, mortality rates and duration of stay), with measures of patient oriented outcomes that are commonly multi-item scales rated by clinicians (for example, disability levels). These measures, which generate total scores by combining the scores of many items, must be simple, easy to use, and rigorous (reliable, valid, responsive) if they are to be administered routinely, and used to influence patient welfare and guide the expenditure of public funds. Therefore, they should have the minimum number of items necessary for rigorous measurement.

In the development of multi-item measures, the balance between item number and scientific rigor can be achieved using psychometric methods. Briefly, a large pool of items is generated to ensure that all important variables are considered for inclusion in the final instrument,[1] and then reduced to its quintessential number on the basis of item performance in empirical field tests.[2] Although psychometric methods have been used extensively in the social sciences,[3] they have been slow to transfer to medicine. Consequently, many widely used health measures—for example, the Barthel index (BI),[4] which is a 10 item measure of physical dependence in personal activities of daily living (PADL)—were developed by choosing items on the basis of their clinical relevance. Whereas this clinical approach to scale development is intuitively sound, it assumes that the items chosen have adequate measurement properties and that all these items are required to measure a construct rigorously.

The fact that the BI was developed clinically raises the question of whether its number of items can be reduced using psychometric methods. Although it has relatively few items, takes only a few minutes to score, and is already recommended for use in elderly populations,[5] rehabilitation,[6] and patients with stroke,[7] there is evidence that a short form BI might be a valuable measure. The 1998 and 1999 Royal College of Physicians National Sentinel Audits of Stroke (n=6894 and 5823) are only able to report BI scores for 59% and 61% of survivors respectively.[8 9] Therefore, the objective of this study was to explore the possibility of developing a short form BI that is psychometrically equivalent to the 10 item measure.

Methods

PARTICIPANTS AND DATA COLLECTION

All admissions to the neurorehabilitation unit of the National Hospital for Neurology and Neurosurgery in London were studied between May 1993 and March 1999. Data routinely collected were diagnostic and demographic

information; admission and discharge disability level measured by the BI (the version of Collin *et al*[10]; recommended by McDowell and Newell[11]) and functional independence measure (FIM[12]) rated by staff from observation. Also, as part of an ethically approved multicentre study conducted between 1994 and 1996, the London handicap scale (LHS[13]) and medical outcomes study 36 item short form health survey (SF-36[14]) were administered to predetermined participants (first two admissions each week).

ANALYSES

The database was randomly divided into two samples. In one sample, short forms were developed by performing an item analysis and selecting those items with the best measurement properties. In the other sample, five measurement properties of the short forms were examined and compared with the 10 item BI. We hypothesised that to improve clinical usefulness significantly while maintaining scientific soundness a short form BI should have a minimum of three and a maximum of five items. Therefore, three short forms (five, four, and three items) were developed and tested.

Development of short forms

The goal was to develop a short form BI that maximised concurrent validity (correlation with the 10 item BI) and responsiveness (ability to detect change in disability). Therefore, items were evaluated on the basis of corrected item total correlations computed from admission scores, and effect sizes computed from change scores (discharge minus admission). The best items were then selected.

Corrected item total correlations are correlations between each item and the sum of the remaining items in the scale. For example, the corrected item total correlation for the transfer item is the correlation between this item and the total score generated by summing the item scores of the other nine Barthel items. Correcting the total score by removing the item of interest prevents spuriously high values due to item overlap. Product-moment correlations were computed for items with polychotomous (three or more) response options, and its equivalent, point biserial correlations, were computed for items with dichotomous (two) response options.[15] Corrected item total correlations indicate the extent to which each item relates to the construct measured by the total score. Consequently, higher values indicate better items.[2]

Effect sizes are standardised change scores.[16] There are many types of effect size calculation.[17] Here they are calculated as the mean change score divided by the SD of admission scores.[18] Effect sizes indicate the extent to which each item changes due to rehabilitation. Therefore, higher values indicate better items.

An index of overall item superiority was determined by rank ordering item total correlations and effect sizes (1=best), and then computing the cross product of these rank orderings. Lower values indicate better items. Short forms with five, four, and three items were generated by selecting the best five, four, and three items respectively.

Psychometric evaluation of short forms

Standard methods were used to examine five psychometric properties: acceptability, reliability, validity, responsiveness, and agreement between scores generated by short form and 10 item BIs.[2 19–23] To aid comparison of different versions of the BI which have different numbers of items and therefore different score ranges, the scores for all scales were transformed to have a range of 0–20. This was achieved using the following formula[24]:

$$\text{Transformed score} = 20 \times \frac{(\text{observed score} - \text{minimum score})}{(\text{maximum possible score} - \text{minimum possible score})}$$

Acceptability is the extent to which the range of health measured by a scale matches the distribution of health in the study sample. It is determined by examining score distributions.[20] Ideally, the observed scores from a sample should span the entire range of the scale, the mean score should be near the scale midpoint, and floor and ceiling effects (% of the sample having the minimum and maximum score respectively) should be small. McHorney and Tarlov recommend that floor and ceiling effects should be<15%.[25]

Reliability is defined as the extent to which random (measurement) error is associated with a measurement instrument (high reliability=low error).[2 20] Reliability is a generic term. Multiple types of reliability (and therefore many reliability coefficients) exist for each instrument, each addresses a different source (or sources) of random error.[26] Although clinicians are most familiar with interrater and intrarater reproducibility, internal consistency is considered a superior indicator of reliability for multi-item measures.[24] Some of the reasons for this are discussed later. Internal consistency reliability is calculated from the intercorrelations among the items using Cronbach's α coefficients.[27] Confidence intervals for α coefficients can be calculated using the formula suggested by Nunnally and Bernstein.[28] It is recommended that reliability estimates should exceed 0.80 for group comparison studies, and 0.95 for individual patient clinical decision making.[2] Confidence intervals for individual patient scores can be computed from reliability estimates by calculating the standard error of measurement (SEM).[2] The SEM is an estimate of the dispersion of scores that would be obtained if a measure was administered to a given individual multiple times.[15] The following formulae are used :

$$\text{SEM} = \text{standard deviation of sample scores} \times \sqrt{(1 - \text{reliability})}$$

$$95\%\ \text{confidence intervals for individual patient scores} = \pm 1.96 \times \text{SEM}$$

Validity is the extent to which a rating scale measures what it purports to measure.[23] In this study, the aim was to determine the extent to which the validity of short forms and original BIs were similar. Three methods were used. Firstly, the extent to which each short form BI predicted the original 10 item BI (concurrent validity) was determined by examining their

Table 1 Characteristics of participants in random half samples

Variable	*Sample 1**	*Sample 2†*
Age (y) (mean (SD))	46.8 (15.1)	47.0 (14.5)
Sex (% female)	55.3	58.4
Length of stay (days) (mean (SD))	31.2 (25.4)	32.5 (23.9)
Diagnosis (%):		
Multiple sclerosis (MS)	45.6	50.8
Stroke	14.2	15.4
Spinal cord disorder	16.5	12.6
Cerebral tumour	5.2	4.8
Peripheral nerve/muscle disorder	5.2	3.1
Other	13.3	13.3
Admission disability level (mean (SD)):		
Barthel index	12.0 (5.4)	11.8 (5.4)
FIM‡	89.0 (23.7)	88.3 (23.0)

*Sample in which short form Barthel index developed (n=419).
†Sample in which short form Barthel index tested (n=418).
‡Functional independence measure total score. Sample 1, n=406; Sample 2, n=396.

intercorrelations. Secondly, the extent to which different forms of the BI related to measures of similar and dissimilar constructs (convergent and discriminant validity[28]) was determined by comparing the magnitude and pattern of their correlations with four other health measures (FIM, LHS, SF-36 PCS, and SF-36 MCS) and two demographic variables (age and sex). Furthermore, we examined the extent to which these correlations conformed with a priori predictions. We expected BIs to correlate highly (r>0.80) with other measures of dependency (FIM), low to moderately (r=0.10 to 0.50) with measures of handicap (LHS) and health status (SF-36), and be uncorrelated (r<0.10) with age and sex. Thirdly, the extent to which short forms and the 10 item BI are interchangeable was determined by examining the agreement between the admission scores they generated using a random effects model intraclass correlation coefficient (ICC[19 29]) and the method proposed by Bland and Altman.[22] Responsiveness is the ability of an instrument to detect change in the construct being measured.[30] This was determined by calculating effect sizes from admission and discharge total scores.[18 20 23] Larger values indicate greater responsiveness. Effect sizes for the different forms of the BI were compared.

Table 2 Item analysis of Barthel index admission scores (n=419)

	Item analyses		
Item	*Item-total correlation* (RO†)*	*Effect size‡ (RO†)*	*Overall index§ (RO†)*
Transfer	0.83 (1=)	0.53 (5)	5 (1)
Bathing	0.57 (6)	0.68 (1)	6 (2)
Toilet use	0.83 (1=)	0.42 (7)	7 (3)
Stairs	0.68 (5)	0.64 (2)	10 (4)
Mobility	0.77 (4)	0.61 (3)	12 (5)
Dressing	0.81 (3)	0.51 (6)	18 (6)
Feeding	0.53 (8)	0.55 (4)	32 (7)
Grooming	0.56 (7)	0.41 (8)	56 (8)
Bladder	0.42 (9)	0.30 (9)	81 (9)
Bowels	0.34 (10)	0.17 (10)	100 (10)

*Calculated as the correlation between the item score and the total score generated by summing the other 9 items.
†Rank order: 1=highest value, 10=lowest value.
‡Mean change score (discharge minus admission) divided by SD of admission scores.
§Cross product of rank order for item total correlation and effect size—for example, for transfer item=1×5=5.

Table 3 Comparison of acceptability, reliability, and agreement (n=418)

	Barthel index (BI)			
*Psychometric property**	*10 item*	*5 item†*	*4 item‡*	*3 item§*
Acceptability:				
Mean score (SD)	11.8 (5.4)	10.0 (6.4)	10.0 (6.3)	11.6 (6.6)
% Floor/ceiling effect¶	1.0/5.0	4.5/8.9	6.5/10.0	6.5/20.1
Reliability:				
α (LL 95% CI**)	0.89 (0.83)	0.88 (0.80)	0.84 (0.73)	0.80 (0.55)
SEM††	1.8	2.2	2.5	2.9
95% confidence interval‡‡	±3.5	±4.3	±4.9	±5.7
Agreement with 10-item BI:				
ICC¶¶	N/A	0.90	0.89	0.92
Mean difference (SD)***	N/A	1.8 (2.0)	1.8 (2.1)	0.2 (2.5)
Limits of agreement†††	N/A	−2.1 to +5.7	−2.3 to +5.9	−4.7 to +5.1

*These psychometric properties were undertaken on admission scores.
†Transfer, bathing, toilet use, stairs, mobility.
‡Transfer, bathing, toilet use, stairs.
§Transfer, bathing, toilet use.
¶Per cent of sample scoring 0 (floor effect) and 20 (ceiling effect).
**Lower limit 95% confidence interval calculated as (α−1.96 SE), where SE = √(SD rii)/ √(k/2 (k−1)−1). SD rii = standard deviation of item intercorrelations; k=number of items in scale.
††Standard error of measurement calculated as SD√(1−α).
‡‡Calculated as 1.96×SEM.
§§Intraclass correlation coefficient (random effects model).
***Admission 10 item BI score minus admission short form BI transformed score.
†††Mean difference ± 1.96 SD.

Results

A total of 844 patients were admitted to the rehabilitation unit between 1993 and 1999. Barthel index scores could not be computed for seven patients (0.8%) due to missing data. The characteristics of those people from whom the short forms were developed and those in whom short forms were evaluated were similar (table 1).

ITEM ANALYSIS AND DEVELOPMENT OF SHORT FORMS

Corrected item-total correlations ranged from 0.83 (toilet use and transfers) to 0.34 (bowels), and effect sizes ranged from 0.68 (bathing) to 0.17 (bowels). The five best items were transfers, bathing, toilet use, stairs, and mobility (table 2).

PSYCHOMETRIC EVALUATION OF SHORT FORMS

All short forms showed good variability as scores spanned the full scale range. Mean scores were situated near the scale midpoint and floor and ceiling effects were small (table 3). Only the three item short form failed to satisfy all acceptability criteria as its ceiling effect exceeded the suggested maximum of 15%.

All α coefficients exceeded the suggested minimum criterion of 0.80, but lower limit confidence intervals for the four and three item short forms fell below this standard (table 3). Confidence intervals around individual patient scores were wide and inversely related to the number of items.

Short forms correlated highly (range 0.93 to 0.96; table 4) with the 10 item BI indicating they were equivalent measures of the same construct. The direction, magnitude, and pattern of correlations with other measures and variables was consistent with predictions and near identical across the four instruments indicating that they had equivalent convergent and discriminant validity. Intraclass correlation coefficients between the 10 item BI and all short forms were high (range 0.89 to 0.92) and exceeded the standard of 0.75 for "excellent" agreement.[21] However, the limits of agreement indicated that scores for individual patients were not identical and inversely related to the number of items (table 3).

Effect sizes for the 10, five, and four item versions of the BI were similar indicating

Table 4 Comparison of validity and responsiveness

Psychometric property	Barthel index (BI) 10 item	5 item*	4 item†	3 item‡
Validity:§				
Concurrent				
10 item Barthel index (%v¶)	1.0 (100)	0.96 (92)	0.95 (90)	0.93 (86)
Convergent and discriminant				
FIM total score**	0.93	0.87	0.87	0.87
LHS††	0.37	0.32	0.34	0.37
SF-36 PCS‡‡	0.22	0.22	0.24	0.23
SF-36 MCS§§	0.14	0.09	0.09	0.11
Age¶¶	−0.04	−0.04	−0.05	−0.07
Sex	−0.09	−0.09	−0.08	−0.05
Responsiveness:***				
Change score (mean (SD))†††	3.8 (3.8)	4.5 (4.9)	4.4 (4.9)	4.2 (4.9)
Effect size‡‡‡	0.71	0.71	0.70	0.64

*Transfer, bathing, toilet use, stairs, mobility.
†Transfer, bathing, toilet use, stairs.
‡Transfer, bathing, toilet use.
§Product-moment correlations between admission scores.
¶Per cent variance of 10 item BI scores explained.
**Functional independence measure (n=396).
††London handicap scale (n=98).
‡‡SF-36 Physical component summary score (n=99).
§§Sf-36 Mental component summary score (n=99).
¶¶n=418 for age and sex.
***n=404 to 406.
†††Discharge minus admission score. All change scores were statistically significant (p<0.001).
‡‡‡Calculated as mean change score divided by SD of admission scores.

equivalent responsiveness (table 4). The effect size for the three item BI was a little smaller.

Discussion

The goal of this study was to develop a short form BI that satisfies criteria for rigorous measurement and is psychometrically equivalent to the 10 item instrument. Of the three short forms developed, the five item BI (table 5) best meets this goal. Reducing the number of items from 10 to five could decrease the time taken to administer the measure and enter data, and lessen the potential for incomplete data collection. Further studies are required to consider these empirical questions. More importantly, selecting items on their performance has resulted in no significant loss of acceptability, reliability, validity, or responsiveness.

Results from this study suggest that the five item BI could replace the original measure in clinical trials, epidemiological studies, and audit. However, can the two instruments be used interchangeably? As different measurement methods are not expected to generate identical results, the essential question is whether the difference between scores is large enough to affect clinical interpretation.[22] The ICC is very high indicating "excellent" agreement between scores.[21] Nevertheless, the sample mean scores differ, signifying a small relative bias.[22] This is a predictable finding (we have selected items with high item total correlations and, therefore, more symmetric item response distributions[31]) and if consistent across samples can be adjusted for by adding 1.84 to mean scores generated by the five item measure. The limits of agreement between scores for individual patients may at first sight seem large (±3.84). However, they are smaller than others have reported for the test-retest reproducibility of the 10 item BI, which is widely accepted to be adequate (±4.2[32]). More importantly, health measures such as the BI are recommended for group comparison studies and not individual patient clinical decision making. This is because confidence intervals around individual scores, as demonstrated here, are too wide to be able to make reliable and valid judgements at the level of the individual patient.[25]

Results from this study underline a fundamental difference between assessment and measurement. When assessing a health construct—for example, a person's dependence in personal activities of daily living—clinicians needs to gather as much relevant information as possible. By contrast, measurement requires that this construct be quantified rigorously. There is no doubt that the 10 item BI provides a more comprehensive assessment of physical dependence in personal activities of daily living than the five item short form. Therefore, the 10 item BI is a superior assessment tool. However, this study demonstrates that the two instruments generate equivalent quantitative estimates (measurements) of this construct in this sample. Consequently, the two instruments are equivalent measures. This finding shows that the entire range of clinically relevant items is not required to measure a construct rigorously. In fact, surprisingly few items are needed provided they are chosen on the basis of their empirical performance as measures. Interestingly, previous investigators have generally added clinically chosen items to the BI thinking that its content was too limited and that longer instruments would be superior measures (see McDowell and Newell[11] for review of 12, 14, 15, 16, and 17 item BIs).

This difference between assessment and measurement emphasises the importance of a psychometric approach to scale development. That is, a large pool of items should be generated and reduced to form rating scales on the basis of their performance in empirical field

Table 5 The five item Barthel index

Item	Response options 0	1	2	3
Transfer	Unable, no sitting balance*	Major help (1 or 2 people, physical), can sit	Minor help (verbal or physical)	Independent
Bathing	Dependent	Independent (or in shower)	—†	—
Toilet use	Dependent	Needs some help, but can do something alone	Independent (on and off, dressing, wiping)	—
Stairs	Unable	Needs help (verbal, physical, carrying aid)	Independent	—
Mobility	Immobile	Wheelchair independent, including corners	Walks with help of one person (verbal or physical)	Independent (but may use an aid—for example, stick)

*From Wade[6]
†No response option.

tests,[3] and not of the sole basis of clinical criteria. However, several methods exist for reducing an item pool to its quintessential number. Our criteria were chosen specifically to develop a measure that predicted the 10 item BI and maximised responsiveness. Other studies have selected items on the basis of linear regression,[33] factor analysis,[34] interitem correlations,[35] equidiscriminatory item-total correlations,[36] Rasch item analysis,[37] item response theory modelling,[38] and patient ratings of item importance and frequency.[35] Although some of these methods have been compared,[35 36] the impact of different item reduction techniques on the development of multi-item measures has yet to be adequately determined.

One previous study developed a short form version of the BI by selecting the items that best predicted function 6 months after stroke.[39] The measurement properties of this four item BI (feeding, grooming, bladder, bowels) have never been reported. In our sample they are limited. The ceiling effect (27.5%) and reliability (α=0.60) fail to satisfy recommended criteria. The correlation (r=0.82) and agreement (ICC=0.69; limits of agreement ±6.24) with the 10 item BI, and responsiveness (effect size=0.52) are notably less than for all of our short forms. Therefore, these four items reflect the 10 item BI to a limited extent and do not constitute a reliable and valid measure of physical independence in personal activities of daily living.

Our study has two limitations. Firstly, test-retest and interrater reproducibility were not examined. Although these data are important, high levels of agreement between the five and 10 item BI indicate good reliability for both measures.[2] In addition, previous studies have consistently demonstrated high test-retest and interrater reproducibility for BI items suggesting that the five item short form total score will be reliable.[11] Moreover, internal consistency is recognised to be the most important type of reliability for multi-item measures because α coefficients are conservative estimates and the test-retest method generates spuriously high values due to memory effect.[2]

The second limitation of this study is its generalisability. We have only studied a sample of people with neurological disability undergoing inpatient rehabilitation. Although subgroup analyses show that results are generalisable to stroke (n=125) and multiple sclerosis (n=407), work is needed to determine the applicability to other samples and to define whether these five items are consistently the most superior. It is also important to note that we have merely shortened an existing instrument and not examined the extent to which these scales are effective outcome measures. Any inherent limitations of the BI remain—for example, its restricted applicability to people with moderate and severe disability, and its failure to measure directly the cognitive and communication impact of disease.

Conclusions

A psychometrically equivalent five item short form BI has been developed. Future studies are now required to determine the generalisability of these results and to establish the limitations and understand fully the trade off of this instrument. Results highlight a fundamental difference between assessment and measurement and the value of a psychometric approach to health measurement.

We thank all the people who participated in this study, the staff of the neurorehabilitation unit who routinely collect data, and Dr Barney Reeves (Royal College of Surgeons, London) for an important discussion. Dr Hobart was funded by a Wellcome Training Fellowship in Health Services Research and a grant obtained by AJT from the NHS Central Audit Fund. The multicentre study was funded by a grant from the North Thames Regional Health Authority Research and Development Responsive Funding Programme (JH PI). There are no conflicts of interest.

1 Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indices and rating scales. *J Clin Epidemiol* 1992;**45**:1201–8.
2 Nunnally JC, Bernstein IH. *Psychometric theory*. 3rd ed. New York: McGraw-Hill, 1994.
3 Spector PE. *Summated rating scale construction: an introduction*. Newbury Park, California: Sage, 1992.
4 Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Maryland State Medical Journal* 1965;**14**:61–5.
5 Royal College of Physicians. *Standardised assessment scales for elderly people: report of joint workshops of the Research Unit of the Royal College of Physicians and the British Geriatrics Society*. London: Royal College of Physicians, 1992.
6 Wade DT. Measurement in neurological rehabilitation. *Curr Opin Neurol* 1993;**6**:778–84.
7 Rudd A, Goldache M, Amess M, *et al*, eds. *Health outcome indicators: stroke. Report of a working group to the department of health*. Oxford: National Centre for Health Outcomes Development, 1999.
8 Irwin P, Rutledge Z, Lowe D, *et al*. *A report on the national sentinel audit of stroke 1998*. London: Royal College of Physicians; 1999.
9 Rutledge Z, Lowe D, Rudd A, *et al*. *National sentinel audit of stroke 1999*. London: Royal College of Physicians, 2000.
10 Collin C, Davis S, Horne V, *et al*. Reliability of the Barthel ADL index. *Int J Rehabil Res* 1987;**10**:356–7.
11 McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. 2nd ed. Oxford: Oxford University Press, 1996.
12 Granger CV, Hamilton BB, Keith RA, *et al*. Advances in functional assessment for medical rehabilitation. *Topics on Geriatric Rehabilitation* 1986;**1**:59–74.
13 Harwood RH, Ebrahim S. *Manual of the London handicap scale*. Nottingham: Department of Health Care of the Elderly, University of Nottingham, 1995.
14 Ware JE Jr, Kosinski MA, Keller SD. *SF-36 physical and mental health summary scales: a user's manual*. Boston, Massachusetts: The Health Institute, New England Medical Centre, 1994.
15 Guilford JP. *Psychometric methods*. 2nd ed. New York: McGraw-Hill, 1954.
16 Cohen J. *Statistical power analysis for the behavioural sciences*. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.
17 Norman GR. Issues in the use of change scores in randomized trials. *J Clin Epidemiol* 1989;**42**:1097–105.
18 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**: S178–89.
19 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1995.
20 Lohr KN, Aaronson NK, Alonso J, *et al*. Evaluating quality of life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;**18**:979–92.
21 Rosner B. *Fundamental of biostatistics*. Toronto: Duxbury Press, 1995.
22 Bland JM, Altman DG. Statistical methods for assessing the agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
23 Fitzpatrick R, Davey C, Buxton MJ, *et al*. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**:14.
24 Stewart AL, Ware JE Jr, eds. *Measuring functioning and well-being: the medical outcomes study approach*. Durham, North Carolina: Duke University Press, 1992.
25 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995;**4**:293–307.
26 Stanley JC. Reliability. In: Thorndike RL, ed. *Educational measurement*. 2nd ed. Washington DC: American Council on Education, 1971.
27 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
28 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
29 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;**1**: 30–46.

30 Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;**40**:171–8.
31 Ware JE Jr, Harris WJ, Gandek B, *et al. MAP-R for windows: multitrait/multi-item analysis program: revised user's guide*. Boston, MA: Health Assessment Lab, 1997.
32 Gompertz P, Pound P, Ebrahim S. The reliability of stroke outcome measures. *Clin Rehabil* 1993;7:290–6.
33 Hays RD, Hayashi T, Stewart AL. A five-item measure of socially desirable response set. *Educational and Psychological Measurements* 1989;**49**:629–36.
34 Peto V, Jenkinson C, Fitzpatrick R, *et al.* The development and validation of a short measure of functioning and well-being for individuals with Parkinson's Disease. *Qual Life Res* 1995;**4**:241–8.
35 Juniper EF, Guyatt GH, Streiner DL, *et al.* Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* 1997;**50**:233–8.
36 Marx RG, Bombardier C, Hogg-Johnson S, *et al.* Clinimetric and psychometric strategies for development of a health measurement scale. *J Clin Epidemiol* 1999;**52**:105–11.
37 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1960.
38 Lord FM, Novick MR. *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley, 1968.
39 Granger CV, Hamilton BB, Gresham GE. The stroke rehabilitation outcome study: part II. Relative merits of the total Barthel index and a 4-item subscore in predicting patient outcomes. *Arch Phys Med Rehabil* 1989;**70**:100–3.

NEUROLOGICAL STAMP

Christjaan Eijkman (1858–1930)

Eijkman was awarded the Nobel Prize for his discovery of the role polished rice played in causing polyneuritis in chickens. This work led to the first real understanding of a possible cure for beriberi and was the starting point of the field of vitamin research.

While a student in Amsterdam he served for 2 years as assistant to the professor of physiology, Thomas Place, under whose guidance he wrote his thesis on *polarisation of nerves*. Immediately after graduation in 1883 he went to the Dutch East Indies and worked in Java and Sumatra as medical officer of health. There he developed malaria, and was so weakened by it, that in 1885 he returned to Holland on sick leave. His young wife died shortly afterwards. In 1886 he returned again to the East Indies with two Dutch physicians CA Pekelharing and C Winkler. They had been appointed by the Dutch Government to study beriberi, an increasingly serious problem, particularly it seemed when people lived closely together such as in army barracks, labour camps and in prisons.

The Pekelharing and Winkler mission demonstrated that beriberi caused a polyneuritis. After a year Pekelharing and Winkler returned to the Netherlands and Eijkman became director of a new laboratory in Batavia (now Jakarta). In July 1889 or 1890 he noted a disease very similar to human beriberi in the chickens in his laboratory. They became restless and unsteady and when a bird descended from its perch it seemed to have to make an effort not to fall and an ascending paralysis occurred over the next few days.

Eijkman noticed that the disease had developed over 5 months when the diet was changed from raw unpolished rice to polished rice. From 10 June, for about a month before the outbreak of the disease the laboratory attendant had been feeding the chickens on polished rice from the kitchen. Five and half months later a new cook refused to supply rice for the chickens, and soon after their return to ordinary chicken food containing raw unpolished rice their disease disappeared.

At Eijkman's request, the medical inspector for Java studied the rice diets in prisons where outbreaks of beriberi had occurred. He found that in prisons where there was beriberi the stable diet was polished rice, whereas in those prisons free of the disorder the normal diet was unpolished rice.

Eijkman, however, failed to recognise that beriberi was a deficiency disease. He argued that the endosperm produced a toxin that was neutralised by the outer hull. He concluded by eating polished rice the toxin would be released in its unneutralised form.

Although Eijkman had clearly demonstrated how to cure and prevent beriberi it was left to Hopkins to identify its cause as a vitamin deficiency. It was not until the early 1930s that Robert Williams identified the vitamin as vitamin B1 (thiamine). Eijkman was honoured philatelically by Grenada in 1978 (Scott 827, Stanley Gibbons 900).

LF HAAS