Age and Ageing 2005; **34:** 228–232
doi:10.1093/ageing/afi063

Reliability of the Barthel Index when used with older people

ANITA SAINSBURY[1], GUDRUN SEEBASS[2], ARUNA BANSAL[1], JOHN B. YOUNG[1]

[1]St Luke's Hospital, Little Horton Road, Bradford BD5 0NA, UK
[2]General and Geriatric Medicine, St James's Hospital, Leeds LS7 9TF, UK

Address correspondence to: G. Seebass, General and Geriatric Medicine, St James's Hospital (c/o Prof Mulley's Secretary), Beckett Wing, Beckett Street, Leeds LS7 9TF, UK. Fax: (+44) 0113 2429195. Email: Gudrun.Seebass@leedsth.nhs.uk

Abstract

Objective: the Barthel Index (BI) has been recommended for the functional assessment of older people but the reliability of the measure for this patient group is uncertain. To investigate this issue we undertook a systematic review to identify relevant studies from which an overview is presented.
Method: studies investigating the reliability of the BI were obtained by searching Medline, Cinahl and Embase to January 2003. Screening for potentially relevant papers and data extraction of the studies meeting the inclusion criteria were carried out independently by two researchers.
Results: the scope of the 12 studies identified included all the common clinical settings relevant to older people. No study investigated test–retest reliability. Inter-rater reliability was reported as 'fair' to 'moderate' agreement for individual BI items, and a high percentage agreement for the total BI score. However, these findings were difficult to interpret as few studies reported the prevalence of the disability categories for the study populations. There may be considerable inter-observer disagreement (95% CI of ±4 points). There was evidence that the BI might be less reliable in patients with cognitive impairment and when scores obtained by patient interview are compared with patient testing. The role of assessor training and/or guidelines on the reliability of the BI has not been investigated.
Conclusions: although the BI is highly recommended, there remain important uncertainties concerning its reliability when used with older people. Further studies are justified to investigate this issue.

Keywords: *Barthel Index, functional assessment, reliability, multiple diagnoses, older people, elderly*

Introduction

The Barthel Index (BI) was developed as a measure to assess disability in patients with neuromuscular and musculoskeletal conditions receiving inpatient rehabilitation [1] and has been recommended by the Royal College of Physicians for routine use in the assessment of older people [2]. The index is an ordinal scale comprising ten activities of daily living. The original BI was scored in steps of five points to give a maximum total score of 100. A widely adopted modification to the index by Collin and Wade [3] includes a revised score range of 0–20.

Clinical measurement scales such as the BI need to be reliable: that is scores should be consistent on serial testing in the absence of real change. Several types of reliability should be investigated. Test–retest (intra-rater) reliability evaluates a single assessor's consistency on repeated testing of the same patient. Inter-rater reliability examines the concordance of responses between independent observers assessing the same patient. In clinical practice the BI can be administered in several ways, such as by interview, by observation of the patient's performance in a care setting, or by asking the patient to demonstrate an activity (testing). If scores obtained in different ways are to be compared, their equivalence needs to be demonstrated.

The reliability of the BI has been well documented for stroke patients [4, 5] but less so for older people with other medical or multiple conditions. The objective of this review was to summarise the available evidence for BI reliability for this common group of older people.

Methods

Studies of BI reliability were identified by searching MEDLINE (1966 to January 2003), EMBASE (1980 to January 2003) and CINAHL (1992 to January 2003) using 'Barthel' and 'Barthel Index' as search terms. The resulting abstracts were inspected for relevance by two researchers and discrepancies resolved by discussion.

Studies examining the reliability of the commonly used version of the BI [1, 3] with a comparison of at least two ratings for the same group of subjects were included. Studies exclusively involving stroke patients or patients aged less than 65 years, and those using other versions of the BI were excluded. Study characteristics and results were extracted from each paper by two independent researchers. Discrepancies were discussed to reach agreement.

Results

Of the 1,857 abstracts identified, 12 studies involving 923 subjects met the inclusion criteria [3, 6–16]. Five of the 12 studies were based on less than 50 subjects [3, 7, 8, 14, 15]. The studies encompassed different care settings, assessment methods and assessors (Table 1), and used different statistical methods (Tables 1 and 2). The study heterogeneity precluded a meta-analysis and the results are therefore presented as a descriptive summary.

The BI was completed by personal interview in seven studies [3, 10, 11, 13–16], by occupational therapist testing in four studies [3, 7, 8, 15], by observation of day-to-day performance in four studies [6, 9, 11, 12], by telephone interviews in two studies [10, 16] and by postal self-report in one study [16]. Five publications included details on training of assessors prior to the reliability study. Training varied from none [16] to pre-study training sessions and provision of additional scoring guidelines [7, 8, 10, 13]. None of the studies compared reliability with and without prior training or guidelines. Two investigators using interviews explicitly excluded patients unable to understand English or answer questions [7, 14]. Two others used proxy responses in place of the patient when necessary [10, 16]. No study has investigated the reliability of proxy responses compared to self-reported scores.

We did not find any study investigating test–retest reliability. Inter-rater reliability was examined by simultaneous observed testing [7, 8] and by sequential BI administration using assessors of different professional backgrounds [6, 9, 12, 13]. Three studies compared different methods of administration [10, 15, 16]. The remaining three studies compared assessments performed at different times by assessors with different professional backgrounds using different administration methods [3, 11, 14] and are hence testing three types of reliability simultaneously.

Six studies investigated the reliability of individual BI item scores and reported kappa values in the 'fair' to 'moderate' agreement range (Table 2). Bowel control appeared to be the most unreliable item with the lowest kappa values in four studies [6, 11–13]. However, only one study reported the prevalence of bowel continence with the majority (93%) of patients fully continent [13]. This would lead to high chance agreement and consequently a lower kappa value.

All 12 studies examined the reliability of the total BI scores but used various statistical methods (Table 1). Percentage agreement within two points varied between 70 and 100% [3, 10, 11, 14]. The largest study used the kappa statistic and found 'good' agreement on total scores lying above or below the threshold of 12 (or 18) points [10].

Three studies [7, 8, 10] used an intraclass correlation coefficient (ICC) that represents a type of weighted kappa with a range from zero ('no agreement') to 1.0 ('perfect agreement') [17]. These studies reported agreement well above the 'accepted minimum' of 0.75 (indicating disagreement in 25% of cases [17]). Some analysis of variance (ANOVA) methods are equivalent to the ICC but the studies using ANOVA [6, 16] only report 'no significant difference', which makes it difficult to interpret the degree of agreement found.

Three studies [12, 13, 16] used the method recommended by Bland and Altman [18] for assessing reliability. This involves calculating the mean difference of the two BI scores for each patient to identify any systematic bias between the two raters and calculating the 95% confidence interval for these mean differences to indicate the magnitude of random measurement error. Both the mean difference and the 95% confidence interval are expressed in the same units as the scale (0–20 in the case of the BI), which aids interpretation. The three studies using the Bland and Altman method reported mean differences between paired measurements close to zero (implying no systematic bias between the two raters), but the 95% confidence limits were wide (indicating considerable random measurement error).

Pearson's correlation coefficient was used in two studies [9, 14] but merely shows linear correlation, which can be perfect even in the presence of a large systematic bias between raters. Its use as a reliability measure is therefore inappropriate [17].

The highest agreement for individual items and for total scores was found when comparing raters from a similar background using the same method of administration such as two occupational therapists during simultaneous testing [8] or the observations of nursing assistants and experienced nurses [12]. Conversely agreement was lower when scores from interview were compared with testing [11, 14, 15]. Skruppy [15] found agreement of scores from interview and testing only for those patients who were independent.

The presence of cognitive impairment has an inconsistent effect on reliability, with studies reporting poor agreement of individual BI items [6, 11] and total scores [11], and one study demonstrating no effect [13].

Discussion

The BI has become a widely adopted clinical measure of disability. The modest number of studies investigating BI reliability in older people with mixed medical conditions is therefore disappointing. Notable gaps in the literature include the absence of any study investigating test–retest reliability, self-report versus carer-report and the effect of training or guidelines on reliability. Our search criteria were broad and identified a large number of potentially relevant

Table 1. Study characteristics and reliability of total BI scores

Author [reference] sample size, setting (age range)	Diagnoses	Comparison	Statistical methods and results
Richards [13] $n = 94$, acute hospital (inter-quartile range 73–84)	FNOF, stroke, others	Interview by nurse and researcher	B&A = −0.14 (−3.75 to +3.47)
Edwards [7] $n = 15$, geriatric unit (>64)	Unspecified	Testing by OT and trained lay persons	ICC = 0.97
Fricke [8] $n = 25$, rehabilitation inpatients (52–87)	Orthopaedic, neurological	Testing by 3 OTs	ICC = 0.957
Skruppy [15] $n = 30$, rehabilitation inpatients (63–89)	Orthopaedic, neurological	Interview and testing by one OT	SC = 0.74 overall; SC = 0.15 for dependent patients
Collin [3]$n = 25$, rehabilitation inpatients (3 patients over 65)	Neurological	Interview of patient and nurse by researcher, testing by nurse and OT	72% agreement for all 4 comparisons KC = 0.93
Roy [14] $n = 20$, rehabilitation inpatients (18–86)	Neurological	a) Physician interview and OT testing; b) 2 OTs testing	a) 70% agreement; b) 100% agreement Pearson's correlation $r = 0.884 - 0.991$; t-test: No significant difference
Ranhoff 1993 [11] $n = 59$, rehabilitation care home (60–99)	Stroke, FNOF, dementia, cancer	Physician interview and nurse observation	71% agreement (if MMSE 20+), 33% agreement (if MMSE < 20)
Ranhoff 1997 [12] $n = 60$, rehabilitation care home (60–99)	Stroke, FNOF, dementia, cancer	Observation by nurse and care assistant	B&A = 0.15 (−4.25 to +4.55)
Artaso [6] $n = 53$, psychogeriatric day centre (>64)	Dementia	Observation by nurse and informal carer	ANOVA: no difference in mean scores
Korner-Bitensky [10] $n = 366$, community (discharged from rehabilitation) (25–98)	Orthopaedic, neurological	Interview by telephone and visit (by OTs and lay persons)	ICC = 0.89
Yeo [16] $n = 80$, community(discharged from acute hospital) (61–95)	Unspecified	Interview a) Post and telephone; b) Post and face-to-face; c) Telephone and face-to-face (by several researchers)	a) B&A = 0.11(−5 to +4), SC = 0.88; b) B&A = 0.05 (−7 to +9), SC = 0.88; c) B&A = −0.28 (−7 to +7), SC = 0.80 ANOVA: no significant difference in scores obtained by three methods
Hartig [9] $n = 96$, nursing home (59–103)	Unspecified	Observation by nurse and nursing assistant	Pearson's correlation $r = 0.9$; paired t-test: $t = 3.15$ ($P = 0.002$)

OT, occupational therapist; FNOF, fractured neck of femur; ICC, intraclass correlation coefficient; B&A, Bland and Altman; SC, Spearman rank correlation; KC, Kendall rank correlation; MMSE, Mini-Mental State Examination [20].

Table 2. Inter-rater reliability of the BI: Cohen' s kappa results for individual activities

			Ranhoff [11]				Yeo [16]		
Activity	Richards [13]	Fricke [8]	MMSE 20+	MMSE <20	Ranhoff[12][a]	Artaso [6][b]	Post/phone	Visit/phone	Post – Visit
Transfer	0.30	0.66	0.46	0.27	0.76	[c]	0.47	0.54	0.42
Walking	0.50	0.67	0.78	0.21	0.85	[c]	0.71	0.59	0.58
Stairs	0.58	n.a.	0.68	0.49	0.69	0.55	0.39	0.53	0.47
Toilet use	0.63	0.67	0.57	0.14	0.65	0.46	0.76	0.51	0.72
Dressing	0.32	0.57	0.51	n.a.	0.77	0.39	0.73	0.63	0.50
Feeding	0.43	0.85	0.51	n.a.	0.54	0.31	0.65	0.56	0.76
Bladder	0.53	0.75	0.33	0.22	0.59	0.32	0.55	0.47	0.81
Bowel	0.27	0.81	0.19	n.a.	0.41	0.17	0.60	0.49	0.62
Grooming	0.50	0.61	0.31	0.11	0.50	0.31	0.69	0.74	0.57
Bathing	0.68	n.a.	0.87	–0.10	0.71	0.29	0.55	0.72	0.61

Face-to-face assessments of the BI unless otherwise stated. Interpretation of kappa: <0.2 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, 0.81–1.0 very good agreement [21].

[a] Weighted kappa.

[b] Dichotomised kappa.

[c] In this study all patients were independently mobile.

MMSE, Mini-Mental State Examination [20].

articles. It is therefore unlikely that additional studies have been overlooked. The scope of the 12 studies identified included the various clinical settings in which the BI is commonly used. The summary finding of 'fair' to 'moderate' agreement of individual BI items, and high intraclass correlation coefficients of total BI scores obtained by different interview methods (e.g. postal questionnaire, telephone or face-to-face), or by testing with different observers, is reassuring. However, there are several caveats that apply to this conclusion.

Firstly, the BI may be less reliable in patients with cognitive impairment and when scores obtained by interview are compared to scores obtained by testing, but the available studies addressing these issues are too small and diverse for robust conclusions.

Secondly, reliability of the BI is influenced by the degree of disability in the population examined and is higher in subjects who are independent [3, 15]. Scoring of individual items may be least reliable when patients' scores lie in the middle categories [3] or when greater disability is present [10]. Only two studies reported the prevalence of disability in their patients [13, 15]. Therefore the effect on the results cannot be estimated.

Lastly, the demonstration of reliability is dependent on the choice of statistical methods. Although there are several statistical methods available to investigate reliability, there is no consensus on the preferred approach [17, 19]. Most studies used several different statistical methods owing to this lack of consensus.

However, some studies used inappropriate methods based on parametric correlation statistics [9, 14], or comparisons of group means [6, 9, 14, see Table 1]. High percentage agreement between raters [3, 10, 11, 14] does not preclude systematic over- or under-rating by one assessor. Simplification of data such as the use of threshold scores by Korner-Bitensky [10] can mask differences in reliability.

Kappa and ICC are acceptable for measuring agreement. They reach higher numerical values when there is a wide range of different scores in the patient group studied [17]. Some studies did not provide the range of scores in their patient group [6, 8], making interpretation of their results difficult. The value of ICC or kappa, which constitutes adequate reliability, is ambiguous and depends on the clinical context in which the measure is used [17].

Studies using the Bland and Altman method [12, 13, 16] showed little systematic bias but a clinically worrying imprecision with a 95% confidence interval of ±4 points or more in a 20 point score instrument. This is partly due to the small sample size of the studies involved. It is of concern that the BI has become so widely adopted for use with older people in the face of this imprecision. Ideally, a sufficiently large repeatability study based on older people with mixed chronic medical conditions (multiple pathology) is required to investigate this further. Additionally, important questions relating to supporting guidance and/or training (with the attendant resource implications) have yet to be adequately addressed.

Key points

- The reliability of the BI has been investigated in the major clinical settings relevant to older people.
- The BI was found to be reliable when administered by face-to-face interview and by telephone (ICC 0.89) and on testing by different observers (ICC 0.95–0.97) but has a considerable imprecision (95% CI of ±4 points).
- The individual studies address different aspects of reliability and use various different statistics, limiting comparability.
- Test–retest reliability has not been investigated on older people with multiple diagnoses.
- A large repeatability study on patients with multiple diagnoses is required to investigate the inter-observer disagreement demonstrated with the Bland and Altman method and to clarify the importance of assessor training.

References

1. Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Maryland State Med J 1965; 14: 61–5.
2. Report of joint workshops of the Research Unit of the Royal College of Physicians and the British Geriatrics Society. Standardised assessment scales for elderly people. London: Royal College of Physicians 1992.
3. Collin C, Wade D. The Barthel Index: a reliability study. Int Disabil Stud 1988; 10: 61–3.
4. Murdock C. A critical evaluation of the Barthel Index, Part 2. Br J Occup 1992; 4: 153–6.
5. Granger CV, Albrecht GL, Hamilton BB. Outcome of comprehensive medical rehabilitation: measurement by PULSES profile and Barthel Index. Arch Phys Med Rehabil 1979; 60: 145–54.
6. Artaso Irigoyen B, Goni Sarries A, Gomez Martinez AR, Garcia Nicholas MA. Direct and indirect assessment of the patient with dementia. Geriatrika 2002; 18: 45–9 (Spanish).
7. Edwards M, Feightner J, Goldsmith CH. Inter-rater reliability of assessments by individuals with and without a background in health care. Occup Ther J Res 1995; 15: 103–10.
8. Fricke J, Unsworth CA. Inter-rater reliability of the original and modified Barthel Index, and a comparison with the Functional Independence Measure. Aust Occup Ther J 1997; 44: 22–9.
9. Hartig MT, Engle VF, Graney MJ. Accuracy of nurse aides' functional health assessment of nursing home residents. J Gerontol Ser A Biol Sci Med Sci 1997; 52: M142–8.
10. Korner-Bitensky N, Wood-Dauphinee S, Siemiatycki J, Shapiro S, Becker R. Health-related information post discharge: telephone versus face-to-face interviewing. Arch Phys Med Rehabil 1994; 75: 1287–96.
11. Ranhoff AH, Laake K. The Barthel ADL Index: scoring by the physician from patient interview is not reliable. Age Ageing 1993; 22: 171–4.
12. Ranhoff AH. Reliability of nursing assistants' observations of functioning and clinical symptoms and signs. Aging (Milano) 1997; 9: 378–80.
13. Richards SH, Peters TJ, Coast J, Gunnell DJ, Darlow MA, Pounsford J. Inter-rater reliability of the Barthel ADL index: how does a researcher compare to a nurse? Clin Rehabil 2000; 14: 72–8.
14. Roy CW, Togneri J, Hay E, Pentland B. An inter-rater reliability study of the Barthel Index. Int J Rehabil Res 1988; 11: 67–70.
15. Skruppy M. Activities of daily living evaluations: is there a difference in what the patient reports and what is observed? Phys Occup Ther Geriatr 1993; 11: 13–25.
16. Yeo D, Faleiro R, Lincoln NB. Barthel ADL index: a comparison of administration methods. Clin Rehabil 1995; 9: 34–9.
17. Streiner DL, Norman GR. Health Measurement Scales. A practical guide to their development and use. Third edn. Oxford Medical Publications, 2003, ch 8: Reliability; 104–27.
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–10.
19. Wade DT. Measurement in Neuro-rehabilitation. Oxford University Press, 1992.
20. Folstein MF, Folstein SE. 'Mini-Mental State'—a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 2: 189–98.
21. Altman DG. Practical Statistics for Medical Research. London, 1991.

Received 17 May 2004; accepted in revised form 17 January 2005