The Berg Balance Scale has high intra- and inter-rater reliability but absolute reliability varies across the scale: a systematic review

Stephen Downs[1], Jodie Marquez[2] and Pauline Chiarelli[2]

[1]Transitional Aged Care, Bellingen Hospital, [2]Discipline of Physiotherapy, University of Newcastle
Australia

Questions: What is the intra-rater and inter-rater relative reliability of the Berg Balance Scale? What is the absolute reliability of the Berg Balance Scale? Does the absolute reliability of the Berg Balance Scale vary across the scale? **Design**: Systematic review with meta-analysis of reliability studies. **Participants**: Any clinical population that has undergone assessment with the Berg Balance Scale. **Outcome measures**: Relative intra-rater reliability, relative inter-rater reliability, and absolute reliability. **Results**: Eleven studies involving 668 participants were included in the review. The relative intra-rater reliability of the Berg Balance Scale was high, with a pooled estimate of 0.98 (95% CI 0.97 to 0.99). Relative inter-rater reliability was also high, with a pooled estimate of 0.97 (95% CI 0.96 to 0.98). A ceiling effect of the Berg Balance Scale was evident for some participants. In the analysis of absolute reliability, all of the relevant studies had an average score of 20 or above on the 0 to 56 point Berg Balance Scale. The absolute reliability across this part of the scale, as measured by the minimal detectable change with 95% confidence, varied between 2.8 points and 6.6 points. The Berg Balance Scale has a higher absolute reliability when close to 56 points due to the ceiling effect. We identified no data that estimated the absolute reliability of the Berg Balance Scale among participants with a mean score below 20 out of 56. **Conclusion**: The Berg Balance Scale has acceptable reliability, although it might not detect modest, clinically important changes in balance in individual subjects. The review was only able to comment on the absolute reliability of the Berg Balance Scale among people with moderately poor to normal balance. **[Downs S, Marquez J, Chiarelli P (2013) The Berg Balance Scale has high intra- and inter-rater reliability but absolute reliability varies across the scale: a systematic review. *Journal of Physiotherapy* 59: 93–99]**

Key words: Postural Balance, Meta-Analysis, Reproducibility of Results

Introduction

The Berg Balance Scale was developed in 1989 via health professional and patient interviews that explored the various methods used to assess balance (Berg et al 1989). Initially, 38 balance tests were selected as potential components of the score and then refined through further interviews and trials to 14 items. Each of these items is scored from 0 to 4, which are summed to make a total score between 0 and 56, with a higher score indicating better balance. Although the Berg Balance Scale was originally developed to measure balance in the elderly, it has since been used to measure balance in a wide variety of patients.

All clinical measurement tools need to be reliable. Absolute reliability is clinically relevant and appears to be the most useful way of describing the reliability of the Berg Balance Scale (Bland and Altman 1986). The absolute reliability of the Berg Balance Scale provides a confidence interval, within which one can be confident that a change in balance is real change. The most common way of expressing this is the minimal detectable change with 95% confidence (MDC95). With regard to balance, intra-rater reliability refers to the reproducibility of a balance score when tested and retested by the same assessor. Inter-rater reliability refers to the reproducibility of a balance score when measured by different assessors. Relative reliability

provides information about the variation in a score due to measurement error relative to variation within a population. This measure of reliability appears commonly in the literature, usually expressed as intra-class correlation (ICC) where a score of 1 represents perfect agreement and a score of 0 represents no relationship. Relative reliability provides perspective of the reliability of the Berg Balance Scale compared to other measurements, but is less useful clinically and is dependent on variability within the study sample. Studies of heterogeneous populations may find a very high relative reliability, even when the test is unable to detect clinically important changes reliably (Bland and Altman 1986). Three commonly used methods of

> **What is already known on this topic:** The Berg Balance Scale rates balance from 0 (very poor) to 56 (normal) and is widely used in many clinical populations. The reliability of the scale in people with stroke has previously been reviewed but its reliability across all clinical populations has not been summarised.
>
> **What this study adds**: Relative intra- and inter-rater reliability of the Berg Balance Scale are high. Absolute reliability was assessable between 20 and 56 on the scale. Absolute reliability varied within this range.

calculating ICC are used, generally referred to as Type 1, Type 2, and Type 3 (Shrout and Fleiss 1979). If a Type 1 calculation is incorrectly used, the reported ICC is likely to be an underestimate. Use of a Type 3 calculation is likely to result in a higher ICC, however Type 3 calculations cannot be generalised validly to assessors not involved in the study (Shrout and Fleiss 1979).

The objective of this review was to summarise the available evidence for the reliability of the Berg Balance Scale across all age groups and conditions where the Berg Balance Scale was used as a balance measurement tool.

Intra-rater reliability is measured by having an assessor measure balance and then repeat the measurement of the same person after a specified time lapse. Inter-rater reliability can be measured either by repeated measures by different assessors or by one assessor performing the test and other assessors rating the test. In the case of the Berg Balance Scale, the second rating can be done either in person or by reviewing a videorecording. Repeated measurements have the disadvantage that a person's underlying balance might change between two measurements and therefore may underestimate the actual reliability of the Berg Balance Scale.

Simultaneous testing of the Berg Balance Scale to measure inter-rater reliability has different disadvantages. The Berg Balance Scale instructions may be interpreted and delivered in slightly different ways by different assessors. Non-verbal components such as demonstrating how to perform balance tests may vary between assessors. Safety considerations may lead some assessors not to attempt components of the Berg Balance Scale that other assessors might consider safe to attempt. An assessor might stand very close to a subject while performing balance testing, and so demonstrate that supervision is required. Simultaneous Berg Balance Scale testing, either in person or by video, can assess the reliability of how different assessors interpret a subject performing the Berg Balance Scale, but will not detect differences in how assessors instruct subjects to perform Berg Balance Scale testing and may therefore overestimate the actual reliability of the Berg Balance Scale.

It is reasonable to speculate that the reliability of the Berg Balance Scale may vary for each of the test items and for different populations. For example, in healthy community-dwelling people, reliability might be affected by disagreement about how Item 14 'standing on one leg' is measured, while easier items such as Item 3 'sitting balance' might be expected to have almost complete agreement of 4/4 among assessments. Conversely, when applied to people with stroke who are unable to stand, the reliability of 'sitting balance' may be more affected, while more difficult tasks such as 'standing on one leg' are likely to be universally assessed as 0/4. An additional factor that might cause variation in the reliability of the Berg Balance Scale is the underlying health conditions of subjects whose balance is tested. Individual studies are unlikely to be able to investigate the Berg Balance Scale over the full range of the scale and over the broad spectrum of causes of disordered balance. This review describes the range of subjects in whom the reliability of the Berg Balance Scale has been studied, reporting both their balance as well as any underlying health condition.

A previous literature review of the Berg Balance Scale (Blum and Korner-Bitensky 2008) considered the relative reliability of the Berg Balance Scale in patients with stroke and found it to have strong reliability. The current review covers important aspects of the reliability of the Berg Balance Scale not considered by the earlier review, including absolute reliability, and the reliability of the Berg Balance Scale in patients with conditions other than stroke.

Floor or ceiling effects occur when a significant proportion of a tested population achieve the lowest or highest possible score on a test, respectively (Everitt 2010). In groups where the mean Berg Balance Scale score is close to 0 or 56, the scale is unlikely to be useful in discriminating between individuals and will exhibit floor or ceiling effects. In such cases the scale is unlikely to be able to detect a change in balance, even if there is a real change. While floor and ceiling effects can potentially impair the clinical and research usefulness of the Berg Balance Scale, they are also likely to inflate its absolute reliability. A person with extremely poor balance is likely to be uniformly rated at 0/4 on most elements of the Berg Balance Scale. Conversely, a person with extremely good balance is likely to be uniformly rated 4/4 on most items of the Berg Balance Scale. Floor and ceiling effects involve groups with lower variability, which in turn lead to lower estimates of relative reliability compared to groups with more variable scores. Therefore, absolute and relative reliability should be interpreted with reference to floor and ceiling effects.

The specific study questions for this systematic review were:
1. What is the relative intra-rater and inter-rater reliability of the Berg Balance Scale?
2. What is the absolute reliability of the Berg Balance Scale, defined as the minimal detectable difference able to be determined with 95% confidence?
3. Does the absolute reliability of the Berg Balance Scale vary across the scale?

Method

Identification and selection of studies

A literature search was undertaken to locate eligible published studies. Electronic searches of Medline, CINAHL, Embase, and the Cochrane Library from 1980 to August 2010 were conducted using 'Berg Balance Scale' as a search term. No search terms were used for intervention type or health condition and no methodological filter was used for study design. See Appendix 1 on the eAddenda for the detailed search strategy. All potentially relevant papers were identified from abstracts and assessed for inclusion. The reference lists of included studies were searched for additional relevant papers. Data were extracted from the included studies by two authors (SD and PC) with any disagreements adjudicated by a third author (JM).

The inclusion criteria for studies are presented in Box 1. Studies investigating the relative reliability of the Berg Balance Scale had to supply a confidence interval around the estimate of the reliability of the scale or data allowing a confidence interval to be calculated. A minimum sample size of 10 was also applied, as recommended by Walter et al (1998). Studies examining translated versions of the scale were included if the study was reported in English. Studies examining a modified or partial version of the scale were excluded. Studies that excluded people who were

Box 1. Inclusion criteria.

Design
- Reliability studies examining the Berg Balance Scale
- Published in English
- Sample size ≥ 10 participants

Participants
- Any clinical population

Outcomes
- Relative intra- and inter-rater reliability
- Absolute reliability

unable to attempt some items of the scale were excluded. Studies that used incorrect or unclear methods to calculate the intra-class correlation coefficient (ICC) and articles not containing original data, such as letters and reviews, were also excluded. Cognitive impairment initially was not a basis for excluding papers. However, only one paper studied people who predominantly had substantial cognitive impairment, so this paper was considered separately.

Assessment of the characteristics of the studies

The following data were extracted from each included study: the number of participants and their age, diagnosis, disease severity, and distribution of scores of the Berg Balance Scale. Any exclusion criteria applied in the original studies were also recorded.

Data analysis

Meta-analyses of the relative intra-rater and inter-rater reliability were performed. Confidence intervals were assessed at 95%. Sensitivity analysis was conducted on studies examining translations of the Berg Balance Scale by individually omitting studies, repeating the analysis and determining if results were significantly different without any study. If not specifically stated, it was assumed that studies conducted in predominantly non-English speaking locations used translations.

To calculate the relationship between absolute reliability and samples of Berg Balance Scale data, samples were weighted for sample size and the mean Berg Balance Scale was plotted against the MDC95. A quadratic line of best fit was used because the floor and ceiling effects can be expected to cause increased absolute reliability as the mean Berg Balance Scale approaches 0 or 56. Meta-analysis of absolute reliability was not conducted due to the confounding effect of the sample mean Berg Balance Scale score on MDC95.

Results

Flow of studies through the review

Of the 511 papers identified (510 from electronic searches and 1 from reference lists), 27 were identified as being related to reliability based on information in the title and abstract. We excluded 15 studies, primarily for having inadequate detail about the methods or insufficient data to include in the meta-analysis. Eleven studies were included in analysis of the reliability of the Berg Balance Scale. The flow of studies through the review is presented in Figure 1.
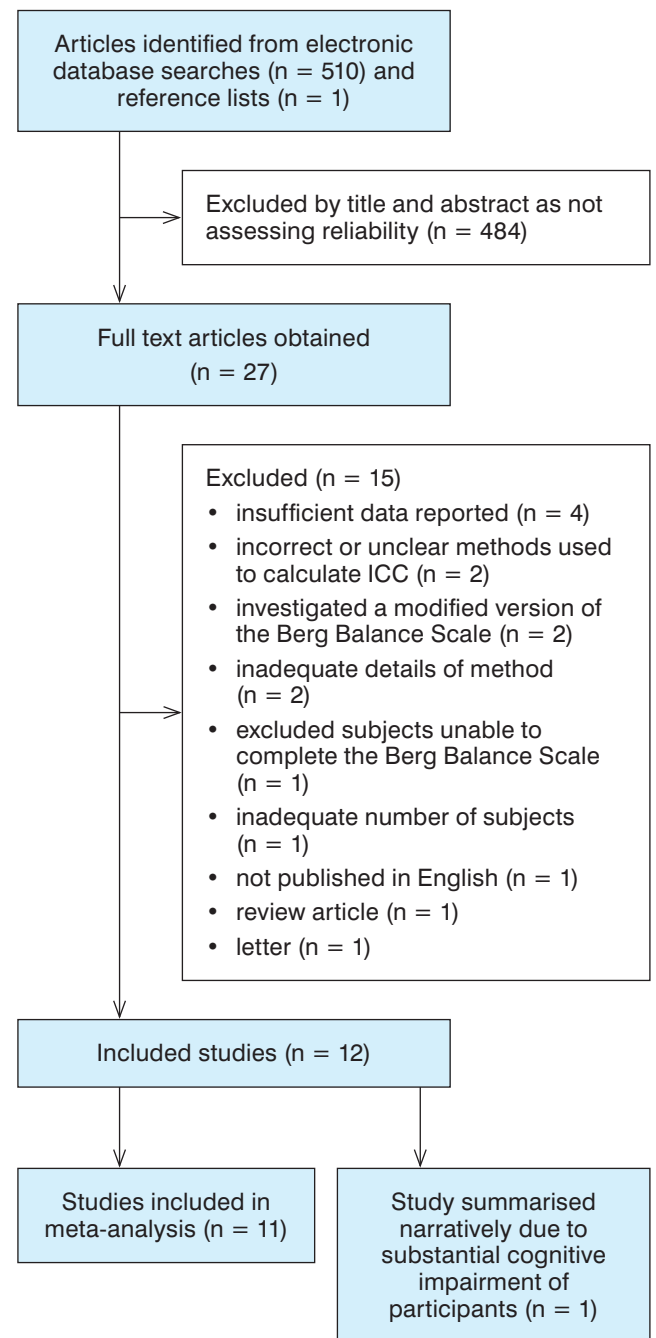

Figure 1. Flow of studies through the review.

Characteristics of the studies

Table 1 presents the characteristics of the included studies, including a description of the participants. These studies included elderly patients (Donoghue et al 2009), elderly residents of an aged care facility (Berg et al 1995), and patients with stroke (Liaw et al 2008, Mao et al 2002, Stevenson 2001), multiple sclerosis (Cattaneo et al 2007, Paltamaa et al 2005), spinal cord injury (Wirz et al 2010), and Parkinson's disease (Lim et al 2005, Steffen and Seney 2008).

Table 1. Summary of included studies (n = 12).

Study	Reliability examined	n	Setting	Diagnosis	Age *(yr)* mean (SD)	Severity	Exclusion criteria	Berg Balance Scale score[a] (0−56)
Berg 1995	Relative	63	Acute stroke ward	Acute stroke	73 (9)	At least some motor impairment	Medically unstable	Full range
			Aged care facility	Not stated	84 (5)	Independently mobile		
Cattaneo 2007	Both	25	Multiple Sclerosis clinic	Multiple Sclerosis	42 (13)	Able to walk 6 metres	Cognitive impairment that might hinder testing	47.3 (7.7)
Halsaa 2007	Relative	83	Geriatric rehab. inpatient unit, day hospital	Varied	82 (6)	Able to walk	Significant cognitive impairment, fracture	44.4
Mao 2002	Relative	112	Hospital inpatients	Recent stroke	69 (11)	Able to follow commands	Unable to give informed consent, able to consent by proxy not excluded	34.8 (18.6)
Wirz 2010	Relative	40	Outpatient rehabilitation centre	Spinal cord injury	49 (12)	Able to walk 15 metres	Balance affecting co-morbidity	41.1 (15.2)
Liaw 2008	Both	52	Outpatient rehab. centre	Chronic stroke	60 (13)	Able to follow verbal instructions	Cognitive impairment, other major diseases	Full range
Donoghue 2009	Absolute	118	Outpatient rehab. centre	Various	81 (7)	n/s	Cognitive impairment, Parkinson's disease, unable to consent, stroke, recent hip replacement	38.6 (9.6)
Lim 2005	Absolute	26	Home visits	Parkinson's disease	63 (8)	Able to walk without gait aid	Cognitive impairment, co-morbidities affecting balance	53.8 (2)
Steffen & Seney 2008	Absolute	37	University	Parkinson's disease	n/s	Able to walk independently	Cognitive impairment, activity limiting heart disease	50 (7)
Paltamaa 2005	Absolute	19	Physiotherapy outpatient department	Multiple sclerosis	Intra-rater: 43 (9) Inter-rater: 49 (9)	Able to walk 20 m	Unable to give written informed consent	54.3 (2.1)
Conradsson 2007	Absolute	45	Aged care facility	Various	83 (7)	Dependant for personal care, mean MMSE 17.5	Unable to stand from chair	30.1 (15.6)
Stevenson 2001	Absolute	48	Rehab. inpatients	Sub acute stroke patients	74 (7)	n/s	Unable to consent	IQR 36.5 to 47

[a]mean (SD) unless otherwise stated, rehab. = rehabilitation, MMSE = Mini Mental State Examination (Folstein et al 1975), n/s = not stated, IQR = interquartile range
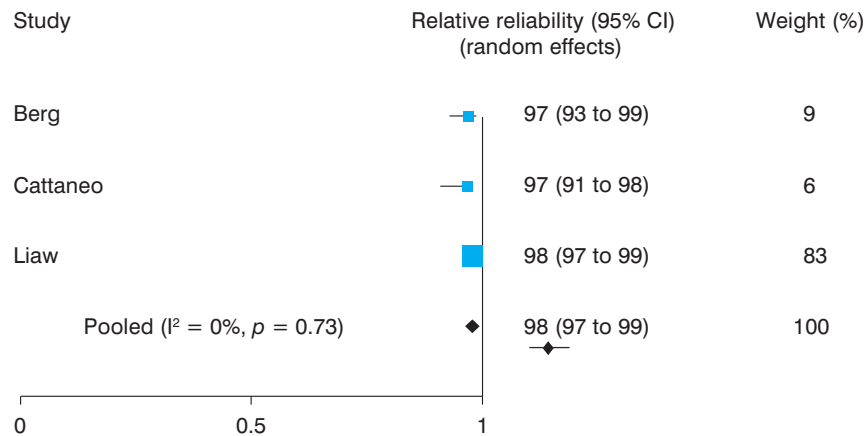
Study	Relative reliability (95% CI) (random effects)		Weight (%)
Berg		97 (93 to 99)	9
Cattaneo		97 (91 to 98)	6
Liaw		98 (97 to 99)	83
Pooled (I² = 0%, *p* = 0.73)		98 (97 to 99)	100

Figure 2. Intra-rater relative reliability of the Berg Balance Scale.

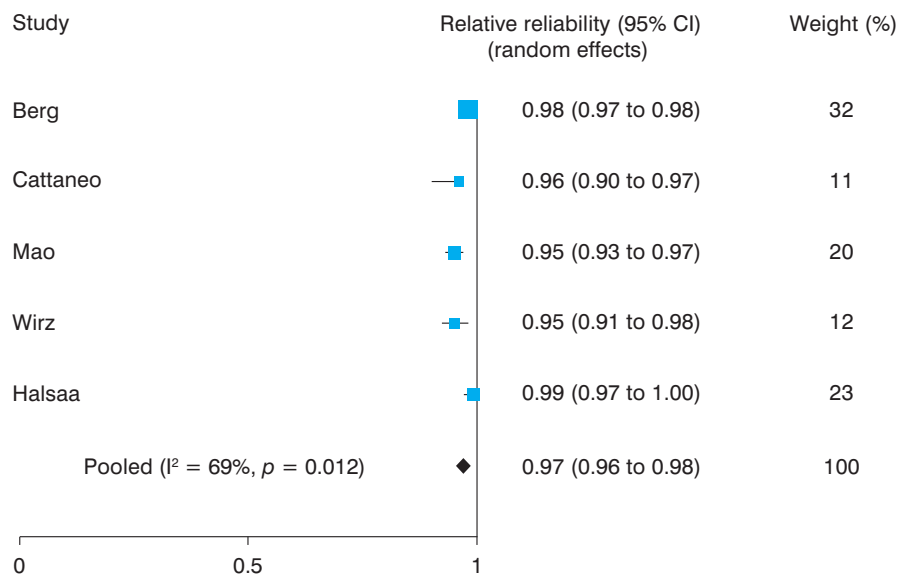| Study	Relative reliability (95% CI) (random effects)		Weight (%)
Berg		0.98 (0.97 to 0.98)	32
Cattaneo		0.96 (0.90 to 0.97)	11
Mao		0.95 (0.93 to 0.97)	20
Wirz		0.95 (0.91 to 0.98)	12
Halsaa		0.99 (0.97 to 1.00)	23
Pooled (I² = 69%, *p* = 0.012)		0.97 (0.96 to 0.98)	100

Figure 3. Inter-rater relative reliability of the Berg Balance Scale.

Relative reliability

The intra-rater relative reliability of the Berg Balance Scale was estimated by meta-analysing data from three studies with a total of 101 subjects. The pooled estimate of the intra-rater relative reliability of the Berg Balance Scale was 0.98 (95% CI 0.97 to 0.99), as presented in Figure 2. A further analysis was conducted to examine the inter-rater relative reliability of the Berg Balance Scale by meta-analysing data from five studies with a total of 345 subjects. The pooled estimate of the inter-rater reliability was 0.97 (95% CI 0.96 to 0.98), as presented in Figure 3. These studies included participants from a variety of clinical populations with balance abilities across the full spectrum of the Berg Balance Scale, although only one study had a sizeable number of subjects with very low Berg Balance Scale scores (Berg et al 1995).

Sensitivity analyses did not find evidence that translations of the Berg Balance Scale into languages other than English have different reliability to the English version. In all cases repeating the analysis omitting translations of the Berg Balance Scale changed the relative reliability by less than 1%. All papers used Shrout and Fleiss Type 2 calculation to calculate ICC except Berg et al (1995), which used Type 1.

Absolute reliability

Studies investigating the absolute intra-rater reliability of the Berg Balance Scale show that the MDC95 varies in relation to the mean Berg Balance Scale scores of the sample, as presented in Figure 4. The review did not identify data about the absolute reliability of the Berg Balance Scale within its lower range of 0 to 20. Only one study examined the absolute inter-rater reliability of the Berg Balance Scale (Cattaneo et al 2007). This found very similar results for absolute intra- and inter-rater reliability.
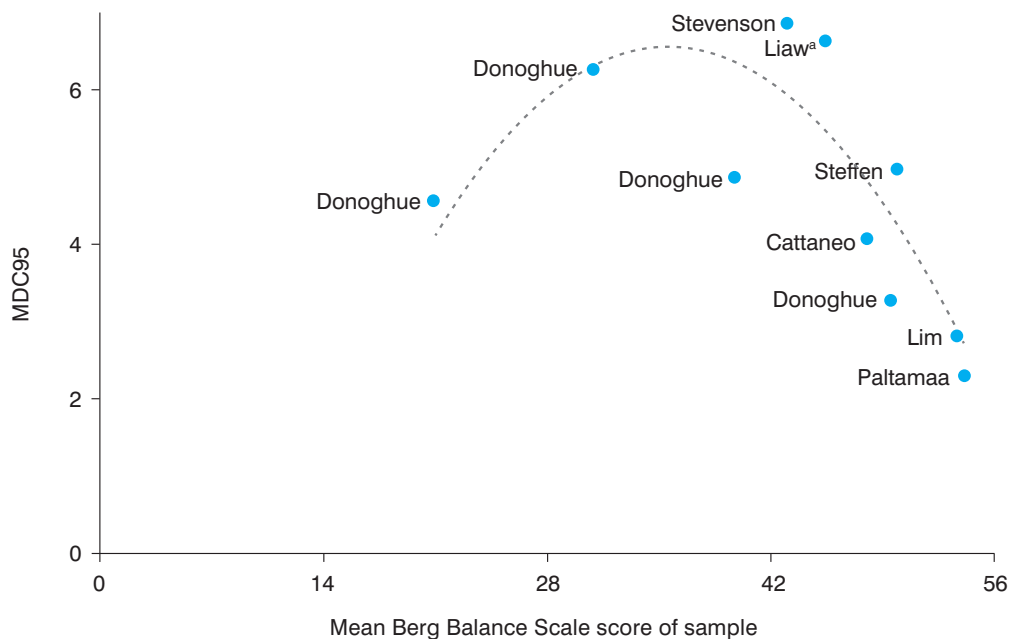
Figure 4. Minimal detectable change with 95% confidence (MDC95) compared to mean of sample population. Dotted line = fitted values. [a]Liaw only reports median value.

Sensitivity analysis was conducted individually on all papers studying the absolute reliability of the Berg Balance Scale using translations. A Swedish translation studying the reliability of the Berg Balance Scale in residential aged care facilities with substantially cognitively impaired residents found a significantly lower absolute reliability with a MDC95 of 7.7 (mean Berg Balance Scale 30.1) (Conradsson et al 2007). These study findings were not included in our analysis of the absolute reliability of Berg Balance Scale. In all other cases the line of best fit with the individual study excluded was almost identical to the analysis presented.

Discussion

Our review identified substantial and consistent evidence of high intra-rater and inter-rater relative reliability of the Berg Balance Scale. Absolute reliability data were also favourable, although some people might experience moderate change in balance that would not be reliably detected by the scale. Furthermore, the absolute reliability data were only available for people with Berg Balance Scores above 20.

The reliability of the Berg Balance Scale has been investigated among a wide variety of subjects, although both studies investigating the reliability of the Berg Balance Scale in patients with Parkinson's disease used subjects with high Berg Balance Scale scores which incurred a ceiling effect. The results of these studies might therefore be considered invalid in terms of describing the reliability of the Berg Balance Scale for patients with Parkinson's disease whose balance scores are in the middle or lower range of the Berg Balance Scale.

This review found little evidence describing the reliability of the English language Berg Balance Scale in people with substantial cognitive impairment, although a Swedish language Berg Balance Scale translation (Conradsson et al 2007) suggests the Berg Balance Scale may be less reliable in people with substantial cognitive impairment.

While the high relative reliability suggests the Berg Balance Scale is clinically useful, there is little specific guidance as to how confident one can be that a real change in balance has occurred between tests across time for individual patients. This review suggests that if an individual has a Berg Balance Scale score of between 20 and 56 and experiences a change of between 3 and 7 (see Figure 4), one can be 95% confident that there has been a real change in balance. Individuals may experience clinically relevant changes in balance that cannot be reliably detected. Downs et al (2012) found hospital inpatients with a Berg Balance Scale of 20 have approximately a 30% probability of being discharged to a nursing home, while those with a Berg Balance Scale of 25 have approximately 20% probability of being discharged to a nursing home, suggesting that a difference in balance which is only barely detectable with 95% confidence in any individual may in fact be highly clinically relevant.

Changes in the average Berg Balance Scale score of patient or research groups have a smaller minimal detectable change than individual subjects. Thus, while moderately clinically important balance changes might not always be detectable with 95% confidence in individuals, they can be expected to be reliably detectable within groups. Researchers or clinicians who find clinically important changes in the average Berg Balance Scale score of a group of individuals might therefore be confident that the change was not caused by random variation.

This literature review did not find data describing the absolute reliability in groups with very low Berg Balance Scale scores, although data presented by Cattaneo et al (2007) suggest that the absolute reliability of the Berg Balance Scale might be higher in the 0 to 20 range than the 20 to 56 range. Bimodal distribution of the Berg Balance Scale has been reported previously (Berg et al 1995, Downs et al 2012), suggesting subjects might be categorised into two distinct groups: those able to stand independently and those unable to stand independently. Where people were able to stand independently, they were also able to attempt and usually achieve a score on several items, generally achieving a Berg Balance Scale score greater than 20. Those unable to stand independently are unable to attempt these items and usually score less than 15. The dichotomous nature of these two groups suggests that the absolute reliability of the lower Berg Balance Scale between 0 and 20 cannot be validly inferred from data related to the higher 20 to 56 range.

This review was underpinned by very broad inclusion criteria which may have impacted the findings. Although studies published in non-English journals were excluded, most of the studies in this review were performed in countries predominantly speaking a language other than English and may have used translations of the Berg Balance Scale.

Our meta-analysis has shown that the Berg Balance Scale has high intra- and inter-rater relative reliability. Several studies of absolute reliability suggest that the Berg Balance Scale is able to detect many clinically significant changes in balance with 95% confidence, although some individuals might experience moderate change in balance that cannot be reliably detected by the Berg Balance Scale. This review found little evidence describing the absolute reliability of the Berg Balance Scale for people with a Berg Balance Scale score between 0 and 20. ∎

Correspondence: Stephen Downs, Transitional Aged Care, Bellingen Hospital, Church Street, Bellingen, 2454, NSW, Australia. Email: Stephen.downs@ncahs.health.nsw.gov.au

References

Berg K, Wood-Dauphine S, Williams JI (1995) The balance scale: reliability assessment with elderly residents and patients with an acute stroke. *Scandinavian Journal of Rehabilitation Medicine* 27: 27–36.

Berg K, Wood-Dauphine S, Williams JI, Gayton D (1989) Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada* 41: 304–311.

Bland J, Altman D (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327: 307–310.

Blum L, Korner-Bitensky N (2008) Usefulness of the Berg Balance Scale in stroke rehabilitation: a systematic review. *Physical Therapy* 88: 559–566.

Cattaneo D, Jonsdottir J, Repetti S (2007) Reliability of four scales on balance disorders in persons with multiple sclerosis. *Disability and Rehabilitation* 29: 1920–1925.

Conradsson M, Lundin-Olsson L, Lindelof N, Littbrand H, Malmqvist L, Gustafson Y, et al (2007) Berg Balance Scale: intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. *Physical Therapy* 87: 1155–1163.

Donoghue D, Murphy A, Jennings A, McAuliffe A, O'Neil S, Charthaigh EN, et al (2009) How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine* 41: 343–346.

Downs S, Marquez J, Chiarelli P (2012) Balance outcomes from two small rural hospitals. *Australian Journal of Rural Health* 20: 275–280.

Everitt BS (2010) *The Cambridge Dictionary of Statistics*. Cambridge: Cambridge University Press.

Folstein MF, Folstein SE, McHugh PR (1975) 'Mini Mental State' A practical method of grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12: 189–198.

Liaw LJ, Hsieh CL, Lo SK, Chen HM, Lee S, Lin JH (2008) The relative and absolute reliability of two balance performance measures in chronic stroke patients. *Disability and Rehabilitation* 30: 656–661.

Lim LI, van Wegen EE, de Goede CJ, Jones D, Rochester L, Hetherington V, et al (2005) Measuring gait and gait-related activities in Parkinson's patients own home environment: a reliability, responsiveness and feasibility study. *Parkinsonism & Related Disorders* 11: 19–24.

Mao HF, Hsueh IP, Tang PF, Sheu CF, Hsieh CL (2002) Analysis and comparison of the psychometric properties of three balance measures for stroke patients. *Stroke* 33: 1022–1027.

Paltamaa J, West H, Sarasoja T, Wilkstrom J, Malkia E (2005) Reliability of physical functioning measures in ambulatory subjects with MS. *Physiotherapy Research International* 10: 93–109.

Shrout P, Fleiss J (1979) Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86: 420–428.

Steffen T, Seney M (2008) Test-retest reliability and minimal detectable change on balance and, ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism. *Physical Therapy* 88: 733–746.

Stevenson TJ (2001) Detecting change in patients with stroke using the Berg Balance Scale. *Australian Journal of Physiotherapy* 47: 29–38.

Walter S, Eliasziw M, Donner A (1998) Sample size and optimal designs for reliability studies. *Statistics in Medicine* 17: 101–110.

Wirz M, Muller R, Bastiaenen C (2010) Falls in persons with spinal cord injury: validity and reliability of the Berg Balance Scale. *Neurorehabilitation and Neural Repair* 24: 70–77.