Development of a Walking-Safety Scale for Older Adults, Part II: Interrater and Test–Retest Agreement of the GEM Scale

François Dubé, Jacqueline Rousseau, Christine Kaegi, Renée Boudreault, Sylvie Nadeau

ABSTRACT

Purpose: The GEM scale is an objective assessment tool, specifically developed for older adults, to evaluate walking safety using standardized tasks. The purpose of this study was to estimate the interrater and test–retest agreement of the GEM scale.

Method: Participants ($n = 41$; ≥ 65 years) were recruited from geriatric units and assessed independently and simultaneously by three raters on two occasions using the GEM scale. Kappa coefficients and percentage agreement were calculated for each item of the scale.

Results: A majority of walking items ($n = 22$) showed fair to substantial interrater agreement ($\kappa \geq 0.25$) and substantial to almost perfect test–retest agreement ($\kappa \geq 0.60$). Mean percentage agreement was high for both interrater and test–retest agreement ($79\% \pm 15\%$ and $83\% \pm 16\%$ respectively). Moreover, detailed analyses demonstrated that the relatively low agreement of some items resulted from changes in the performance of some participants and the low variability of scores. Although some walking items showed less agreement, the final decision regarding the participants' ability to walk safely resulted in moderate to substantial interrater and test–retest agreement.

Conclusion: The GEM scale is a new assessment tool that can now be used with estimated interrater and test–retest properties to allow therapists to objectively evaluate walking safety among the elderly.

Key Words: assessment, elderly, gait, physiotherapy, psychometric properties

Dubé F, Rousseau J, Kaegi C, Boudreault R, Nadeau S. Development of a walking safety scale for older adults, part II: interrater and test–retest agreement of the GEM scale. *Physiother Can.* 2008;60:274-282.

RÉSUMÉ

Objet : La GEM est un instrument d'évaluation objectif, développé pour les aînés dans le but d'évaluer la sécurité à la marche à l'aide de tâches standardisées. Cette étude avait pour but d'évaluer la concordance interjuges et test-retest de la GEM.

Méthodologie : Les participants ($n = 41$; ≥ 65 ans) ont été recrutés parmi les unités gériatriques et ont été évalués à deux reprises, de façon indépendante et simultanée, par trois évaluateurs à l'aide de la GEM. Les coefficients kappa et le pourcentage de concordance ont été calculés pour chaque item de l'instrument.

Résultats : La majorité des items de marche ($n = 22$) ont démontré une concordance interjuges de « acceptable » à « très bonne » ($\kappa \geq 0,25$) et une concordance de « très bonne » à « presque parfaite » pour le test-retest ($\kappa \geq 0,60$). La moyenne des pourcentages d'accord était élevée pour l'accord interjuges et test-retest ($79\% \pm 15\%$ et $83\% \pm 16\%$, respectivement). En outre, des analyses détaillées ont démontré que la concordance relativement faible de certains items était due à des changements dans la performance de certains participants et à la variabilité peu élevée des cotes.

François Dubé, PT: MSc candidate, Université de Montréal, Institut universitaire de gériatrie de Montréal, Montreal, Quebec.

Jacqueline Rousseau, PhD: Associate Professor, École de réadaptation, Université de Montréal, and researcher at the Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montreal, Quebec.

Christine Kaegi, PT: Institut universitaire de gériatrie de Montréal, Montreal, Quebec.

Renée Boudreault, PT: Institut universitaire de gériatrie de Montréal, Montreal, Quebec.

Sylvie Nadeau, PhD: Associate Professor, École de réadaptation, Université de Montréal, and researcher at the Centre de recherche interdisciplinaire en réadaptation du Montréal métropolitain, Institut de réadaptation de Montréal, Montreal, Quebec.

This project was supported in part by a grant from the Comité aviseur de la recherche clinique of the Institut universitaire de gériatrie de Montréal and by the Ordre professionnel de la physiothérapie du Québec.

The authors thank the individuals who volunteered for this study. Special thanks to all the members of the Physiotherapy Department for the recruitment of subjects and their constant support and help throughout the study. The authors also acknowledge Mary-Grace Paniconi for reviewing the manuscript.

The results of this study were partially presented in poster sessions: at the XVIIIth Conference of the International Society for Postural and Gait Research (ISPGR), May 2005, Marseille, France; the scientific meeting of the Réseau québécois de recherche sur le vieillissement (RQRV), April 2005, Montreal, Canada; and the Annual Symposium of the Ordre professionnel de la physiothérapie du Québec (OPPQ), September 2005, Montreal, Canada.

Address correspondence to *François Dubé, PT*, Institut universitaire de gériatrie de Montréal, 4565 Queen Mary Road, Montreal, QC H3W 1W5 Canada; Tel: (514) 340-3518; Fax: (514) 340-2826; E-mail: francoisdube@hotmail.com.

DOI:10.3138/physio.60.3.274

Bien que certains items de marche aient montré une concordance plus faible, la décision finale concernant la sécurité à la marche a démontré un accord interjuges et test-retest variant de « modéré » à « très bonne ».

Conclusion : La GEM est un nouvel instrument d'évaluation dont les qualités métrologiques de fidélité interjuges et test-retest estimé permettent aux thérapeutes d'évaluer objectivement la sécurité à la marche chez les aînés.

Mots clés : évaluation, aînés, marche, physiothérapie, qualités métrologiques

BACKGROUND

In Canada, it has been estimated that 30% of people aged 65 and older experience at least one fall each year.¹ The cost associated with falls is an important concern in geriatric rehabilitation and will certainly increase as our population ages. Falls in the elderly are often associated with fear of falling and loss of independence.^{2,3} From 1997 to 2002, the rate of fall-related deaths for people aged 65 and older increased and was related to increasing age.⁴ Gait changes also occur with ageing, and some changes in gait patterns are related to falls.⁵

In a rehabilitation context, gait evaluation is an important part of daily clinical practice.⁶ Based on their expertise, physiotherapists regularly assess walking ability in the elderly. Assessment of a person's safety in walking is multidimensional and is aimed at determining whether a person can walk independently, with or without an ambulation aid, or whether supervision or assistance is needed. According to Ruchinskas et al., in an article on clinical decision making in falls prediction, "Clinicians are frequently asked to make determinations of critical importance regarding patient care."⁷ (p.263) Determining the safety level of an elderly patient during ambulation is a challenging task for physiotherapists and has important implications for patients and their caregivers. Therefore, it is important for physiotherapists to have access to reliable and objective tools to assess walking safety in the elderly.

A literature review revealed that there are many assessment tools available to evaluate gait.⁸ Until now, the assessment of walking safety by physiotherapists was primarily subjective. The Grille d'évaluation de la sécurité à la marche (GEM) scale was developed as an objective assessment tool using standardized items to address the specific issue of walking safety for the elderly. It was important for the authors of the GEM to develop a scale accessible to professionals in physiotherapy with various levels of clinical experience and two different levels of education (physical therapist and physical rehabilitation therapist). Based on two overall concepts, walking and safety, the GEM scale includes 33 walking items distributed within three sub-scales: A, B, and C. (For a complete description of the GEM scale, see Kaegi et al.⁸) Sub-scales A and B are used to evaluate safety for indoor walking, while sub-scale C is used as a pretest for safe outdoor walking. The GEM scale also includes questions addressed to the patient regarding his or her

own perception of walking safety in different environments.

Each walking item is scored on a four-level ordinal scale, the "safety scale." The rater assesses whether the patient is walking safely (2 = safe), requires supervision or cueing (1S = supervision/cueing), requires assistance (1A = assistance), or performs the item dangerously even with physical assistance (0 = danger). At the end of sub-scales A and B (indoor walking), a "rater score" is assigned using a three-level ordinal scale based on an analysis of the walking item scores and the patient's perception of his own walking safety. For sub-scales A and B, the evaluator concludes that (a) the patient is not safe in walking, (b) the patient is somewhat or sometimes safe in walking, or (c) the patient is safe in walking. In sub-scale C (pretest for outdoor walking), the rater score is a two-level ordinal scale. The rater assigns a score of either (a) the patient does not have the prerequisites for safe outdoor walking or (c) the patient has the prerequisites for safe outdoor walking. To be consistent with the scoring criteria for sub-scales A and B, there is no (b) rating for sub-scale C.

The GEM scale underwent a three-phase content-validation process, which demonstrated the relevance of this assessment tool and its representativeness as a walking-safety assessment for the elderly.⁸ The aim of the present study was to continue the validation process by estimating the interrater and test-retest agreement of GEM scale scores.

METHODS

Characteristics of the participants and raters, procedures, and data analyses are described below. The Ethics Committee of the Institut universitaire de gériatrie de Montréal (IUGM) approved this study.

Raters

Two physiotherapists (university degree) and one physical rehabilitation therapist (3-year college-level diploma) were recruited from the Montreal region. To participate in the study, raters were required to be Francophone and unfamiliar with the GEM scale. Their practice experience varied from 1.5 to 21 years. The raters worked with different types of patients (orthopaedic, neurological, geriatric, and paediatric/obstetric) and in

Table 1 Subject Characteristics for Interrater and Test–Retest Agreement Studies

	<i>Interrater (n = 41)</i>		<i>Test–Retest (n = 31)</i>	
	<i>Mean (SD)</i>	<i>n (%)</i>	<i>Mean (SD)</i>	<i>n (%)</i>
Age	78.4 (7.9)		78.9 (8.1)	
Gender				
Female		20 (49)		13 (42)
Male		21 (51)		18 (58)
Functional tests				
Berg Balance Scale (max. /56)	42 (11)		38 (14)	
Walking speed (m/s)	0.6 (0.2)		0.5 (0.3)	
Ambulation aid				
None		17 (42)		12 (39)
Cane		11 (27)		9 (29)
Wheeled walker		8 (19)		7 (23)
Quad cane, narrow base		2 (5)		1 (3)
Quad cane, wide base		2 (5)		2 (6)
Forearm crutches		1 (2)		0 (0)
Primary diagnosis				
Functional loss and deconditioning		12 (29)		11 (35)
Orthopedic problems (surgery, fracture, pain)		11 (27)		7 (23)
Neurological disorder (stroke, Parkinson or Alzheimer disease)		10 (24)		9 (29)
Falls, balance and gait disorders		6 (15)		2 (6)
Cardio-respiratory impairments		2 (5)		2 (6)
Health care unit				
Rehabilitation		21 (51)		17 (55)
Short-term care		14 (34)		10 (32)
Day hospital		4 (10)		2 (6)
Long-term care		2 (5)		2 (6)

various settings (rehabilitation unit, day hospital, out-patient clinic, long-term care, and private practice).

Participants

Participant characteristics for interrater and test–retest assessments are presented in Table 1. Participants ($n=58$; ≥ 65 years) were recruited from the Institut universitaire de griatrie de Montral (IUGM) between November 2003 and July 2004. For various reasons—refusal ($n=12$), discharge from the ward ($n=3$), cancellation of assessment session because of participant’s medical condition ($n=7$), or rater’s illness ($n=5$)—17 participants did not complete both sessions (T1, T2) and 10 did not complete either assessment session. Therefore, data from participants who attended T1 or T2 were included in the interrater agreement analysis ($n=41$) and data from those who completed T1 and T2 were used in the test–retest agreement analysis ($n=31$).

Participants had to (1) understand and speak French and (2) be identified by their treating physiotherapist as needing an assessment of their walking safety. Participants were excluded if they were medically unstable or if weight bearing on one or both legs was contraindicated. Participants were referred to the principal author by their treating physiotherapist, and written informed consent was obtained from all participants before the study, in accordance with the procedures of the Ethics Committee of the IUGM.

Procedure

Prior to the study, the raters met with the principal author to obtain information about the procedures for the assessment sessions. They were provided with the GEM scale (administration manual and scoring sheets) one week before the study began and asked to master and fully understand the GEM scale prior to and throughout the study. The raters received no formal training, additional information, or clarification about the scale. Furthermore, the raters were asked not to use the GEM scale in their own practice and to avoid discussing it with colleagues.

Each participant was scheduled for two testing sessions (T1 and T2), three days apart. The mean duration of the assessment sessions was 46 ± 11 minutes (minimum, maximum: 27, 75 minutes). As specified in the GEM administration manual, participants could rest at any time during testing. The equipment was in accordance with the administration manual of the GEM scale. A safety belt was placed around the waist of each participant by the principal author. To ensure consistency among participants, raters were asked to hold the belt loosely without limiting the participant’s movements. At the beginning of each session, raters were informed of the gender, age, and diagnosis of the patient and the walking aid to use during testing. The choice of walking aid was based on information given by the participant’s treating physiotherapist. The three raters evaluated each participant simultaneously.

One rater conducted the assessment while the other two observed. The lead rater was alternated for each participant, so that each rater conducted the testing of 13 or 14 participants. The lead rater gave the participants instructions while the other two (observers) placed themselves as close as possible to the participants while avoiding verbal or physical interaction with them. Using the GEM scale, the raters completed the score sheets independently according to the administration manual. At the end of each session, time was allowed for the raters to complete the score sheets if necessary. The raters were not allowed to make any changes to the score sheets after handing them to the principal author or to discuss the participants' performances, the scoring, the instructions, or the interpretation of the GEM scale.

Data Analysis

Analyses were performed on the safety scale (walking items) and the rater scale (rater score) using SPSS 12.0 statistical analysis software (SPSS Inc., Chicago, IL). Kappa coefficients⁹ and agreement percentages were calculated for each item of the GEM scale and for the rater score of each sub-scale. For interrater analysis ($n=41$), the data for the three raters collected at T1 were analyzed using the macro for the multi-rater kappa coefficient, available on the SPSS website (<http://www.spss.com/>). For test-retest agreement ($n=31$), the data for the three raters collected at T1 and T2 were analyzed with the macro for non-square kappa coefficient tables available on the SPSS website. We tested the hypothesis that kappa values were greater than zero. The test of significance was one-tailed with a significance level of $p<0.05$. The Landis and Koch scale¹⁰ was used to qualify the agreement, with the kappa coefficients categorized on a six-level agreement scale: (1) poor ($\kappa<0.00$), (2) slight ($\kappa=0.00-0.20$), (3) fair ($\kappa=0.21-0.40$), (4) moderate ($\kappa=0.41-0.60$), (5) substantial ($\kappa=0.61-0.80$), (6) almost perfect ($\kappa=0.81-1.00$).¹⁰

RESULTS

The interrater and test-retest results are presented in Tables 2 and 3 respectively. Kappa coefficients and agreement percentages of the 33 walking items and the rater score for the three sub-scales (A, B, and C) are presented for interrater and test-retest reliabilities. The SE is also presented for the rater score for the three sub-scales.

Interrater Agreement

Walking Items

Twelve walking items demonstrated significant ($p<0.05$) kappa coefficients ($\kappa_m \pm SD = 0.48 \pm 0.14$) varying between fair and substantial agreement on the Landis and Koch scale.¹⁰ Mean agreement percentages ($\pm SD$) of these walking items varied from 37% to 93% ($66 \pm 16\%$).

Twenty-one walking items had kappa coefficients representing poor to moderate agreement ($\kappa \leq 0.41$). However, the majority of those items (20/21) had agreement percentages between 80% and 93% (mean = 87%); the final item had a lower percentage agreement (58%). The mean percentage agreement ($\pm SD$) for all walking items was $79 \pm 15\%$.

Rater Scores

The rater score for sub-scales A, B, and C had significant ($p<0.001$) kappa coefficients (minimum, maximum: 0.54, 0.64) varying between moderate and substantial agreement on the Landis and Koch scale.¹⁰ The mean percentage agreement ($\pm SD$) for the rater score was $70 \pm 6\%$.

Test-Retest Agreement

Walking Items

Data from the three raters were analyzed. For rater #1, all the walking items except one had significant ($p<0.05$) kappa coefficients ($\kappa_m \pm SD = 0.64 \pm 0.20$) with a mean percentage agreement ($\pm SD$) of $87 \pm 11\%$. For rater #2, 14 of the 33 items had significant kappa coefficients ($\kappa_m \pm SD = 0.52 \pm 0.17$) with a mean percentage agreement ($\pm SD$) of $84 \pm 10\%$. The other 19 items showed low kappa coefficients ($\kappa_m \pm SD = -0.01 \pm 0.01$) but high percentage agreement (mean $\pm SD = 95 \pm 3\%$). Likewise, for rater #3, 21 items had significant kappa coefficients ($\kappa_m \pm SD = 0.57 \pm 0.19$) with a mean percentage agreement ($\pm SD$) of $83 \pm 11\%$; the other 12 items had low kappa coefficients ($\kappa_m \pm SD = 0.01 \pm 0.07$) but high percentage agreement (mean $\pm SD = 95 \pm 5\%$).

Rater Scores

The rater score for sub-scales A, B, and C showed significant ($p<0.05$) kappa coefficients for the three raters (minimum, maximum 0.39, 0.80) and qualified as fair to substantial agreement.¹⁰ The mean percentage agreement ($\pm SD$) for the rater score was $77 \pm 9\%$.

DISCUSSION

Ratings for most of the walking items showed high percentage agreement and fair to almost perfect kappa coefficients. The rater score for the three sub-scales reflected fair to substantial interrater and test-retest agreement. Some kappa coefficients were highly influenced by the prevalence of scores (see below).

Distribution of Scores

The kappa coefficient is used as an index of concordance for categorical data. It differs from percentage agreement in that it adjusts for agreement that occurs by chance.^{11,12} The kappa coefficient is influenced

Table 2 Inter-rater Agreement Results

	κ	Landis and Koch Scale ¹⁰	% Agreement
Sub-scale A: Basic Walking Items			
A1. Stand up from a chair (or wheelchair) and walk 10 m	0.74**	Substantial	90
A2. Walk then turn 180°	0.18§	Slight	88
A3. Walk and turn head to the right	0.41	Moderate	88
A4. Walk and turn head to the left	0.17§	Slight	90
A5. Walk and stop abruptly	-0.03§	Poor	90
A6. Walk backwards	0.40*	Fair	78
A7. Walk sideways to the right	-0.04§	Poor	88
A8. Walk sideways to the left	0.34	Fair	88
A9. Walk and make an 'S' around 2 chairs and walk	0.29	Fair	80
A10. Walk and sit down on a chair (or wheelchair)	-0.04§	Poor	88
Rater score A (SE)	0.60** (0.08)	Moderate	68
Sub-scale B: Advanced Walking Items			
B1. Walk and sit on a chair without armrests	0.30§	Fair	90
B2. Get up from a chair without armrests and walk	0.65*	Substantial	93
B3. Walk over a doorsill	0.26§	Fair	83
B4. Walk and pick up a shoe from the floor	0.19§	Slight	80
B5. Walk and open then close a door	-0.04	Poor	85
B6. On carpet, walk 5 m	0.22§	Fair	93
B7. On carpet, walk then turn 180°	0.34	Fair	88
B8. On carpet, walk backwards	0.26§	Fair	80
B9. On carpet, walk sideways to the right	0.22§	Fair	93
B10. On carpet, walk sideways to the left	0.12§	Slight	88
B11. On carpet, walk and make an 'S' around 2 chairs and walk	0.12§	Slight	88
B12. Walk and climb stairs	0.54**	Moderate	71
B13. Walk and descend stairs	0.36*	Fair	66
Rater score B (SE)	0.64*** (0.07)	Substantial	66
Sub-scale C: Pretest for Outdoor Walking Items			
C1. Walk and step up onto a platform	0.54***	Moderate	64
C2. Step down from a platform and walk	0.60***	Moderate	67
C3. On mat, walk 2 m	0.39*	Fair	58
C4. On mat, walk then turn 180°	0.26*	Fair	45
C5. On mat, walk and stop abruptly	0.25	Fair	58
C6. On mat, walk backwards	0.30**	Fair	37
C7. Walk up an incline	0.09§	Slight	85
C8. Walk down an incline	-0.04§	Poor	87
C9. Walk and climb stairs	0.49**	Moderate	61
C10. Walk and descend stairs	0.50**	Moderate	67
Rater score C (SE)	0.54** (0.16)	Moderate	77

* $p < 0.05$ ** $p < 0.001$ *** $p < 0.0001$ §Kappa coefficient explained by the first paradox described by Feinstein and Cicchetti¹¹

by prevalence.¹¹ In the present study, the results were highly influenced by the low variability of scores for two of the three raters. Raters #2 and #3 showed lower variation in scores and generally used the maximum scoring level (2, or Safe). Rater #1, on the other hand, showed a higher variation in scoring (i.e., often used a wider range on the safety scale). The low variability in scores produced a high prevalence of the maximum score level on the safety scale. This high prevalence produced symmetrically unbalanced marginal totals in the matrix, which lowered the kappa coefficients despite high values of percentage agreement.

This phenomenon is explained by the first of two paradoxes described by Feinstein and Cicchetti,¹¹ whereby high percentage agreement is associated with low kappa coefficients. In our study, this paradox explains some of the low kappa coefficients, which are due to low variability in the scores' distribution. If the

analysis was limited to the kappa coefficient, the results could be misinterpreted. Most low kappa coefficients (marked § in Tables 2 and 3) actually represent high inter-rater agreement as a result of the high prevalence of scores in the kappa matrix.

The lack of variability in scores may have different causes, each with possible consequences for the distribution of scores. First, in designing the GEM scale, the authors' aim was to develop an instrument that could be used without any specific training by the two types of professionals responsible for carrying out physiotherapy within Canada: physiotherapists, who have a university degree, and physical rehabilitation therapists (in Quebec), who have a 3-year college-level diploma. (Both are licensed professionals within the province of Quebec.) In addition, the authors of the GEM scale wanted an instrument that would be accessible to these professionals regardless of their clinical experience in geriatrics.

Table 3 Test-retest Agreement Results

	κ	Rater #1 Landis and Koch Scale ¹⁰	% Agreement	κ	Rater #2 Landis and Koch Scale ¹⁰	% Agreement	κ	Rater #3 Landis and Koch Scale ¹⁰	% Agreement
Sub-scale A: Basic Walking Items									
A1. Stand up from a chair (or wheelchair) and walk 10 m	0.51**	Moderate	84	0.50*	Moderate	88	0.52***	Moderate	85
A2. Walk then turn 180°	0.84***	Almost perfect	97	0.00§	Poor	97	0.00§	Poor	100
A3. Walk and turn head to the right	0.58**	Moderate	90	−0.03§	Poor	91	0.48***	Moderate	94
A4. Walk and turn head to the left	0.84***	Almost perfect	97	0.00§	Poor	97	1.00***	Almost perfect	100
A5. Walk and stop abruptly	0.78***	Substantial	97	−0.03§	Poor	94	0.00§	Poor	100
A6. Walk backwards	0.65***	Substantial	87	0.29*	Fair	88	0.23	Fair	85
A7. Walk sideways to the right	0.84***	Almost perfect	97	−0.01§	Poor	94	0.00§	Poor	97
A8. Walk sideways to the left	0.62**	Substantial	91	−0.01§	Poor	94	−0.03§	Poor	91
A9. Walk and make an 'S' around 2 chairs and walk	0.72***	Substantial	94	0.00§	Poor	91	0.25*	Fair	85
A10. Walk and sit down on a chair (or wheelchair)	0.72***	Substantial	94	0.00§	Poor	91	0.00§	Poor	100
Rater score A (SE)	0.49*** (0.12)	Moderate	68	0.69*** (0.11)	Substantial	79	0.80*** (0.09)	Substantial	88
Sub-scale B: Advanced walking Items									
B1. Walk and sit on a chair without armrests	†	†	94	0.00§	Poor	97	0.35*	Fair	91
B2. Get up from a chair without armrests and walk	0.45*	Moderate	87	0.48**	Moderate	94	0.30*	Fair	88
B3. Walk over a doorkill	0.84***	Almost perfect	97	0.00§	Poor	100	0.53***	Moderate	85
B4. Walk and pick up a shoe from the floor	0.67**	Substantial	90	0.47*	Moderate	91	0.47*	Moderate	91
B5. Walk and open then close a door	0.50*	Moderate	97	−0.03§	Poor	94	0.00§	Poor	100
B6. On carpet, walk 5 m	0.78***	Substantial	97	0.00§	Poor	100	0.00§	Poor	97
B7. On carpet, walk then turn 180°	0.84***	Almost perfect	97	0.48**	Moderate	94	0.00§	Poor	94
B8. On carpet, walk backwards	0.67**	Substantial	90	−0.03§	Poor	94	0.60***	Moderate	88
B9. On carpet, walk sideways to the right	0.35*	Fair	90	0.00§	Poor	100	0.65***	Substantial	97
B10. On carpet, walk sideways to the left	0.47*	Moderate	94	0.00§	Poor	97	−0.05§	Poor	88
B11. On carpet, walk and make an 'S' around 2 chairs and walk	0.72***	Substantial	94	0.00§	Poor	97	1.00***	Almost perfect	100
B12. Walk and climb stairs	0.85***	Almost perfect	81	0.63***	Substantial	82	0.63***	Substantial	77
B13. Walk and descend stairs	0.77***	Substantial	90	0.72***	Substantial	85	0.65***	Substantial	77
Rater score B (SE)	0.44** (0.12)	Moderate	61	0.55*** (0.12)	Moderate	71	0.72*** (0.10)	Substantial	82
Sub-scale C: Pretest for Outdoor Walking Items									
C1. Walk and step up onto a platform	0.40*	Fair	69	0.38*	Fair	69	0.69***	Substantial	72
C2. Step down from a platform and walk	0.37*	Fair	66	0.26*	Fair	63	0.48***	Moderate	63
C3. On mat, walk 2 m	0.68***	Substantial	86	0.64***	Substantial	94	0.50***	Moderate	72
C4. On mat, walk then turn 180°	0.65**	Substantial	83	0.45*	Moderate	88	0.56**	Moderate	75
C5. On mat, walk and stop abruptly	0.52*	Moderate	86	0.00§	Poor	94	0.50*	Moderate	75
C6. On mat, walk backwards	0.38*	Fair	62	0.44**	Moderate	75	0.48**	Moderate	66
C7. Walk up an incline	0.63**	Substantial	93	0.00§	Poor	100	−0.05§	Poor	88
C8. Walk down an incline	1.0***	Almost perfect	100	0.00§	Poor	97	−0.03§	Poor	94
C9. Walk and climb stairs	0.72***	Substantial	86	0.60***	Moderate	75	0.71***	Substantial	78
C10. Walk and descend stairs	0.66***	Substantial	83	0.89***	Almost perfect	88	0.71***	Substantial	78
Rater score C (SE)	0.47* (0.16)	Moderate	79	0.57*** (0.16)	Moderate	81	0.39* (0.20)	Fair	84

* $p < 0.05$ *** $p < 0.001$ *** $p < 0.0001$ §Kappa coefficient explained by the first paradox described by Feinstein and Cicchetti¹¹

†Kappa coefficient impossible to calculate due to lack of variation of scoring for T1 (see Discussion)

The raters recruited for this study had a variety of levels of clinical experience and education. It is possible that the three raters used the safety scale differently in this study, as a result of their differing clinical experience and/or educational level, and this may have led to lower variability in scores. Rater #1 was a physiotherapist with four years' work experience in geriatric rehabilitation; rater #2 was a physical rehabilitation therapist with 21 years' experience, predominantly in an orthopaedic outpatient programme as well as in geriatric long-term care; and rater #3, also a physiotherapist, had 1.5 years of clinical experience in paediatrics. We postulate that rater #1 showed higher variability in score distribution because of his greater experience in geriatrics. We hypothesize that, in this study, he was confronted with situations similar to those in his everyday clinical practice, whereas raters #2 and #3 may have had less exposure to the decision-making process with respect to walking safety in this target population. Both experience and education may influence the capacity to adequately interpret the concept of safety and to use the safety scoring scale.

Second, it is possible that the description of the safety scale was somewhat unclear to the raters. It was noted that the raters primarily scored the top level of the safety scale (2 = safe) and rarely the second level (1S = supervision/cueing), suggesting that they had difficulty in discriminating between the two. This may be because safety is not well defined in the literature and is a new concept for a walking-assessment tool. On a clinical level, therapists also recognize that, whereas the need for assistance (code 1A) is more clear cut, there is sometimes a fine line between a patient who needs supervision and one assessed as walking safely. Because the differences among the four levels of the safety scale have important clinical implications for physiotherapy professionals, clarifying the definition of 1S (supervision/cueing) could conceivably remedy this situation. Offering formal training might also improve clinicians' ability to distinguish among the four safety levels.

Lastly, the low variability in scores may be attributed in part to limits in the methodology used for data collection, participant recruitment, and sample size. Because of the importance of maintaining close proximity to the participants, the study was limited to three raters; some items (e.g., B5, walk and open then close a door) cannot be performed with more than three raters. Our objective of attaining a sample size of 30 for the test-retest agreement was achieved, but a larger sample might have led to greater variability in the rater score. Although a convenience sample was used, the participants' characteristics were heterogeneous in terms of diagnoses, walking aids, and type of care unit (day hospital, rehabilitation unit, chronic-care unit). Furthermore, the participants presented with a wide range of balance and walking abilities, as reflected by their Berg Balance Scale scores (mean $\pm$ SD = $42/56 \pm 11$; minimum, maximum: 9, 56)

and walking speed (mean $\pm$ SD = 0.6 ± 0.2 m/s; minimum, maximum: 0.1, 1.0). Based on their characteristics, the participants recruited likely represented a wide range of safety levels and were representative also of the population for which the GEM scale is intended. Therefore, the low variability in the distribution of scores is most likely explained by a rater bias due to different levels of experience and education and lack of clarity in the safety-level description rather than by characteristics of the sample.

Agreement on Walking Items

Interrater Agreement

On the Landis and Koch scale,¹⁰ rater agreement on the walking items varied from poor to substantial. Scores for 20 of 21 items had low kappa coefficients despite high percentage agreement. The majority of these items (16/21) were likely affected by the first paradox described by Feinstein and Cicchetti¹¹ (marked § in Table 2). Only 5 items were neither significantly reliable nor explained by the paradox.

Four of these 5 items had percentage agreement $\geq 80\%$. One item (C5) demonstrated a fair score with low percentage agreement (58%). For this item, partial or complete disagreement among raters was present for 16 participants, for 15 of whom interrater disagreement occurred between scores 2 and 1S. As mentioned previously, the top two levels of the safety scale are perhaps more difficult to distinguish, and the different levels of experience and educational backgrounds of the raters may have affected the agreement results. Based on the percentage agreement and the influence of the prevalence of scores on the kappa coefficient, most of the walking items (28/33) showed high interrater agreement.

Other factors may have influenced interrater agreement, but we are unable to confirm these hypotheses. First, each rater conducted the testing sessions for one-third of the participants and observed for the other two-thirds. The difference in proximity to the patient between the lead rater and the observer may have influenced interrater agreement. Second, some walking items (e.g., B4, B5, C9) are multiple-task items that include a series of actions to be taken in a certain order, while others are single-task items (e.g., A6, B9, C7). Agreement between raters for multiple-task items is probably more difficult to achieve because of the greater complexity of these items.

Test-Retest Agreement

Overall, rater #1 showed high variability in scores and demonstrated moderate test-retest agreement on the GEM scale. Rater #1 was significantly reliable on all but one of the items ($p < 0.05$); for that item (B1), rater #1 demonstrated high percentage agreement (94%). The kappa coefficient could not be calculated for this item, as there was no variability of scores; rater #1 scored the

highest level (2 = safe) for all participants at T1 for that item. Generally, raters #2 and #3 presented lower kappa coefficients because of the low variability of their scores. The vast majority of low kappa coefficients found for raters #2 and #3 are explained by Feinstein and Cicchetti's first paradox.¹¹

Test-retest agreement for three specific items (C1, C2, and C6) may have been influenced by changes in performance on the part of certain participants. For these items, the performance of some participants changed between T1 and T2; on one occasion they performed the walking item without physical help (score of 2 or 1S), and on the other occasion they required physical help to complete the item (score of 1A or 0). These changes in performance for some participants may explain the lower kappa coefficients for those three items.

Rater Score Agreement

The rater score represents the therapist's final decision for each sub-scale and is of the utmost importance in decisions about walking safety. The GEM scale is structured so that scores are attributed to the walking items. After each sub-scale (A, B, and C), these scores are then interpreted and result in a final decision, the rater score. Overall, the three raters showed moderate to substantial interrater agreement and fair to substantial test-retest agreement for the rater score. There was greater variability on the rater score across the three raters, and, consequently, higher agreement, than for the individual walking items.

An internal-consistency study of the GEM scale is underway and, ideally, should have been performed prior to the agreement studies to reduce the number of items in the scale. For budgetary and recruitment reasons, however, the agreement studies were performed prior to the internal-consistency study. In addition, physiotherapy professionals involved in the three-phase content-validation process demonstrated high clinical interest for each of the walking items retained in the GEM scale (see Kaegi et al.⁸). Therefore, the objective of the internal-consistency study is not to reduce the number of items but to further assess the level of agreement in scoring the scale by analyzing the correlations matrix and to complement the construct-validation study that is also underway.

CONCLUSION

In the current study, raters demonstrated greater test-retest agreement than interrater agreement for the walking items. Our results revealed that raters could reliably score the majority of the walking items (moderate to almost perfect kappa coefficients).¹⁰ Two principal biases contributed to lower agreement

for some items: (1) low variability in the scores given to participants and (2) real changes in levels of performance for some participants between T1 and T2. The first paradox described by Feinstein and Cicchetti¹¹ allowed us to interpret some of these kappa coefficients in more depth by analyzing the distribution of the scores and considering the percentage agreement. This analysis suggested that some of the low kappa coefficients might have resulted from limited variability in scores, which may have been due to a rater bias or to the description of the levels in the safety scale. On the other hand, there was great variability among scores for the rater's final decision, with moderate to substantial interrater and test-retest agreement. Thus, although raters showed low levels of agreement for some individual walking items, their final decisions on participants' ability to walk safely showed adequate interrater and test-retest agreement.

The GEM scale is a new tool to enable therapists to assess walking safety in older-adult patients. Interrater and test-retest agreement and percentage agreement were assessed across three raters who scored the GEM scale independently. Studies are presently being conducted to determine the internal consistency and construct validity of the GEM scale and further studies, including the predictive validity of the scale, are recommended also.

KEY MESSAGES

What Is Already Known on This Subject

The GEM scale was developed to assess walking safety in the elderly. It underwent a three-phase content-validation process. Further psychometric tests are now needed to continue the study of this instrument, including construct and predictive validity.

What This Study Adds

Results of this study will enable clinicians to use the GEM scale knowing the various issues affecting reliability. Our results indicated that the rater's final decision on a patient's walking safety showed acceptable interrater and test-retest agreement.

REFERENCES

1. O'Loughlin JL, Robitaille Y, Boivin JF, Suissa S. Incidence of and risk factors for falls and injurious falls among the community-dwelling elderly. *Am J Epidemiol.* 1993;137:342-54.
2. Maki BE. Gait changes in older adults: predictors of falls or indicators of fear? *J Am Geriatr Soc.* 1997;45:313-20.
3. Stel VS, Smit JH, Pluijm SM, Lips P. Consequences of falling in older men and women and risk factors for health service use and functional decline. *Age Ageing.* 2004;33:58-65.

4. Report on seniors' falls in Canada. Ottawa: Public Health Agency of Canada; 2005.
5. Kemoun G, Thoumie P, Boisson D, Guieu JD. Ankle dorsiflexion delay can predict falls in the elderly. *J Rehabil Med.* 2002;34:278–83.
6. Mulder T, Nienhuis B, Pauwels J. Clinical gait analysis in a rehabilitation context: some controversial issues. *Clin Rehabil.* 1998;12:99–106.
7. Ruchinskas RA, Macciocchi SN, Howe GL, Newton RA. Clinical decision making in the prediction of falls. *Rehab Psychol.* 2001;46:262–70.
8. Kaegi C, Boudreault R, Rousseau J, Bourbonnais D, Nadeau S, Dubé F. Development of a walking safety scale for older adults, part I: content validity of the GEM scale. *Physiother Can.* 2008;60:264–273.
9. Cohen JA. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
10. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–74.
11. Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol.* 1990;43:543–9.
12. Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. *J Clin Epidemiol.* 1990; 43:551–8.