ORIGINAL ARTICLE

Smallest Real Difference of 2 Instrumental Activities of Daily Living Measures in Patients With Chronic Stroke

Wen-Shian Lu, MS, Christine C. Chen, ScD, Sheau-Ling Huang, MS, Ching-Lin Hsieh, PhD

ABSTRACT. Lu W-S, Chen CC, Huang S-L, Hsieh C-L. Smallest real difference of 2 instrumental activities of daily living measures in patients with chronic stroke. Arch Phys Med Rehabil 2012;93:1097-1100.

Objective: To estimate the smallest real difference (SRD) values of 2 instrumental activities of daily living measures (the Nottingham Extended Activities of Daily Living [NEADL] and the Frenchay Activities Index [FAI]) in patients with chronic stroke.

Design: Test-retest reliability study.

Setting: Physical rehabilitation units of 5 hospitals.

Participants: Chronic stroke patients (N=52; 37 men, 15 women) who were discharged from the hospital for more than 6 months.

Interventions: Not applicable.

Main Outcome Measures: Both measures were administered twice about 2 weeks apart to participants. The SRD was calculated on the basis of standard error of measurement: SRD = $1.96 \times \sqrt{2} \times$ Standard error of measurement. SRD% (the value of SRD divided by total score of a measure) was used to compare measurement errors across both measures. Reproducibility between successive measurements of the measures was investigated with intraclass correlation coefficients (ICCs).

Results: The SRD (SRD%) values of the NEADL and the FAI were 12.0 (21.1%) and 6.7 (14.9%), respectively. Test-retest reproducibility of both measures was high (ICC: NEADL=.89, FAI=.89).

Conclusions: Because of substantial SRD values of the NEADL and the FAI, prospective users should be cautious in using both measures to detect real change for a single subject.

Key Words: Activities of daily living; Bias (Epidemiology); Cerebrovascular accident; Rehabilitation; Reproducibility of results.

PARTICIPATION IN ACTIVITIES of daily living is an important goal of rehabilitation for patients with chronic stroke. According to the *International Classification of Functioning, Disability and Health*,[1] participation refers to the engagement in life situations through assuming different social roles and involvement in communal activities. Limitations in the ability to perform instrumental activities of daily living (IADLs) (eg, housekeeping, recreation, hobbies, and social interaction) may affect the social participation in patients with chronic stroke who live in the community.[2,3] Measuring IADL performance is important for quantifying the level of participation in chronic stroke patients, and a reliable IADL measure is essential in planning intervention and conducting research.

Test-retest reliability refers to the extent of reproducibility of repeated measurements or the absence of random measurement error.[4] Random measurement error of a test can be quantified by standard error of measurement and smallest real difference (SRD).[5] Standard error of measurement is a commonly known concept; however, clinicians are less familiar with SRD. In fact, the two are mathematically related (as will be demonstrated later). Like standard error of measurement, SRD is expressed in the same unit as the original measurement; however, the two focus on different aspects of measurement error. The standard error of measurement provides an estimate of measurement error of each test/assessment; the SRD, on the other hand, is the smallest change that indicates real improvement or deterioration for an individual—that is, real change beyond measurement error.[6,7] When a repeated measurement difference is greater than the SRD, such a difference can be considered a real difference.[8] Thus, the SRD of a measure is critical for clinicians and researchers to determine whether a patient's change is true.

The extent of the reproducibility of repeated measurements is commonly examined by the intraclass correlation coefficient (ICC). However, SRD can also be used to examine the test-retest reliability.[9] An ideal clinical test should have a high ICC and a small magnitude of standard error of measurement (and by extension, small SRD). Such a test can then be used with confidence to monitor clinical changes in patients over time and across settings.

The Nottingham Extended Activities of Daily Living (NEADL) and the Frenchay Activities Index (FAI) are widely used to assess IADL performance in patients with stroke.

List of Abbreviations

FAI	Frenchay Activities Index
IADLs	instrumental activities of daily living
ICC	intraclass correlation coefficient
LOAs	limits of agreement
NEADL	Nottingham Extended Activities of Daily Living
SRD	smallest real difference
SRD%	smallest real difference percentage
SRD_{group}	smallest real difference at the group level
$SRD_{individual}$	smallest real difference at the individual level

From the School of Occupational Therapy, Chung Shan Medical University, and Occupational Therapy Room, Chung Shan Medical University Hospital, Taichung, Taiwan (Lu); Programs in Occupational Therapy, Department of Rehabilitation and Regenerative Medicine, Columbia University, New York, NY (Chen); and the School of Occupational Therapy, College of Medicine, National Taiwan University, and Division of Occupational Therapy, Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan (Huang, Hsieh).

Supported by the National Health Research Institute in Taiwan (grant no. NHRI-EX99-9512PI), the E-Da Hospital (grant no. 97-EDN11), and College of Medicine, National Taiwan University (grant no. 98HM00021).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Reprint requests to Sheau-Ling Huang, MS, School of Occupational Therapy, College of Medicine, National Taiwan University, 4F, No 17, Xuzhou Rd, Taipei 100, Taiwan, e-mail: *cathy@ntu.edu.tw*.

In-press corrected proof published online on Apr 5, 2012, at www.archives-pmr.org.

0003-9993/12/9306-01103$36.00/0
doi:10.1016/j.apmr.2012.01.015

However, the amount of random measurement error (ie, SRD and standard error of measurement values) of these 2 tests remains unknown, limiting their utility in clinical settings to assess the IADL change for a single subject. Comparison studies of these 2 measures, which are important for selecting an appropriate IADL measure, are rare. The purpose of this study was, therefore, to determine and compare the amount of random measurement error and the extent of the reproducibility of the 2 tests. Our results would be useful for potential users to select appropriate measures and to improve clinical data interpretation in patients with chronic stroke.

METHODS

Participants

Patients with stroke were being treated as outpatients and recruited from the physical rehabilitation units of 5 hospitals in Taiwan. Participants enrolled in the study met the following inclusion criteria: (1) had either a cerebral infarction or hemorrhage; (2) were older than 20 years; (3) could complete the interviews and describe coherently activities of their typical day; and (4) were at least 6 months after discharge from the hospital and had stable IADLs. Patients who had other diseases (eg, heart disease or multiple sclerosis) that were likely to affect IADL function were excluded. All subjects voluntarily participated in this study and signed the consent forms. The study was approved by the medical ethics committee of the participating universities and hospitals.

Procedure

All subjects were interviewed by 1 trained occupational therapist in 2 sessions, about 2 weeks apart. Between the 2 testing sessions, all subjects received regular therapies as they usually did. During the interview, the NEADL and FAI were administered. To control for possible bias from the testing sequence, half of the subjects were given the NEADL first, and the other half, the FAI first. Subjects were not told during the first session that the same questionnaires would be administered again (in a later session), to reduce possible memory effect. Before the second session, each subject confirmed that no significant change in his/her daily life (eg, injury, disease, or living place) had occurred within the past 2 weeks.

Measures

The NEADL, designed for stroke patients, contains 22 items in 4 areas: mobility, kitchen, domestic, and leisure.[10] The subjects were asked to recall the difficulty in performing IADLs in the previous week. In this study, we used a 19-item culturally adapted Taiwanese version, in which 3 items from the original 22-item version were deleted because of lack of discriminative values.[2] Each item was scored on a 4-level scale (0, unable to perform the activity; 1, able to perform the activity with help; 2, able to perform the activity with some difficulty; 3, able to perform the activity easily). The total score ranges from 0 to 57, with higher scores implying less difficulty the patients perceive in performing IADLs.[2,10,11]

The FAI was developed to measure social activities or lifestyle for patients with stroke.[12] It contains 15 items of IADLs in 3 areas: domestic, leisure/work, and outdoors. The subjects were asked to recall the frequency of performing 10 everyday IADLs in the past 3 months and 5 seasonal activities in the past 6 months.[3,13] Each item was scored on a 4-level scale, from 0 (never) to 3 (highest frequency). The total score ranges from 0 to 45. Higher FAI scores imply a higher frequency in performing IADLs.

Data Analysis

Data were analyzed with the SPSS 15.0 for Windows statistical program.[a] Paired *t* tests were first performed to ensure that no systematic bias existed between test-retest assessments. Pearson product-moment correlation coefficient was then used to examine the extent of association between the scores of the NEADL and those of the FAI, at both testing sessions. In addition, the following analyses were conducted. The significance level was set at .05 for all analyses.

SRD and SRD percentage of the 2 measures. Measurement error makes the assessed value of a measure differ from the true value. We first determined the SRD of each of the 2 measures (ie, the NEADL and FAI). The SRD indicates whether a person's change score is real (or true change) at the 95% confidence level[14] and can be calculated as follows: $SRD = 1.96 \times \sqrt{2} \times$ Standard error of measurement (1.96 because of the 95% confidence, $\sqrt{2}$ because of the difference of 2 variances). The standard error of measurement can be derived from the square root of the within-subject variance.[14-17]

In addition, we used the smallest real difference percentage (SRD%) to compare the amount of random measurement error between the 2 measures. The SRD% can be calculated by dividing the SRD with the maximum score of the measures and multiplying by 100%. We used SRD% because it is independent of measurement units.[4] Because there is no commonly accepted criterion available for judging an acceptable value of SRD%, we used Smidt et al's[18] standard as a reference.[18] An SRD% of 10% or less was considered an acceptable measurement error for a measure in this study.

Score reproducibility between 2 testing sessions. To estimate the level of reproducibility between test-retest measurements, we examined the ICC[4] and the limits of agreement (LOAs).[19,20] $ICC_{2,1}$ estimates the consistency of the total test score between 2 testing sessions and is based on the 2-way random effects of analysis of variance of the scores.[4] A test with an ICC of .80 or greater is considered as having good reproducibility (between test-retest measurements).[21]

LOA is a term coined by Bland and Altman.[20] These authors developed a method to visually examine test agreement by plotting the difference scores (d) against the mean scores of each pair of assessments (between first and second sessions).[19,20] Assuming that differences follow the standard normal distribution, 95% of the differences should lie between $d \pm 1.96 \times SD$, where d is the mean difference between the 2 test sessions and SD represents the SD of differences. Ideally, there should not show a systematic trend in a Bland and Altman plot, because the mean scores of the 2 assessments at test and retest are usually not associated with the difference scores between the 2 testing sessions.

RESULTS

Sample Description

A total of 67 subjects with chronic stroke participated in this study and completed the first assessment. Of the 67 subjects, 15 were lost to follow-up either because of loss of contact or the patient's refusal to be retested. Results of *t* tests showed that there were no significant differences in mean baseline scores of the NEADL and the FAI between those who completed both sessions of testing and those who did not ($P=.20$ for NEADL and $P=.39$ for FAI). Demographic and clinical characteristics of the remaining 52 subjects are shown in table 1. The mean age was 59.4 years, and 71% of the subjects were men. The mean baseline scores of the NEADL and the FAI were 17.3 and 9.3 points, respectively.

Table 1: Demographic and Clinical Characteristics of the Subjects

Characteristics	Subjects Completed 2 Assessments (n=52)	Subjects Did Not Complete 2 Assessments (n=15)
Age (y)	59.4±11.6	57.0±10.2
Sex (male/female)	37/15	10/5
Lesion side (left/right/ bilateral)	20/30/2	7/8/0
NEADL baseline scores	17.3±13.3	11.7±10.3
FAI baseline scores	9.3±7.4	7.5±7.9
Interval between 2 assessments (d)	13.7±2.5	

NOTE. Values are mean ± SD or n.

SRD and SRD% of the 2 Measures

The mean difference (1.0 for NEADL and 0 for FAI) between test-retest scores was small and nonsignificant ($P=.248$ for NEADL, $P=.938$ for FAI), indicating no systematic bias between test and retest in either measure. In addition, we found high associations (Pearson $r=.89$ at both sessions) between the scores of the NEADL and those of the FAI.

The SRDs of the NEADL and the FAI were 12.0 and 6.7, respectively. The SRD% was 21.1% for the NEADL and 14.9% for the FAI (table 2), suggesting that the levels of random measurement error of both measures were beyond the acceptable range.

Reproducibility Between Repeated Measurements of the 2 Tests

The ICC between test and retest of the NEADL and FAI were both .89. Such high levels of ICCs indicate that at the test level, the test-retest reproducibility of these 2 tests is excellent (see table 2).

The LOAs for the NEADL (−11.1, 13.1) and the FAI (−6.9, 6.9) are shown in figure 1. The Bland and Altman plot of the NEADL (see fig 1) showed a trend: the higher the mean test values, the larger the differences between the 2 test sessions. In contrast, the differences did not vary in any systematic way over the range of the FAI.

DISCUSSION

Although the NEADL and the FAI are commonly used in clinical research, to our knowledge, the SRDs for these tests have not been reported in previous studies. The results of this study provide some important references to help researchers and clinicians judge whether the change in performance is real change or an artifact of random measurement error. If a patient's score difference between 2 consecutive measurements is larger than the SRD, the change score can be viewed as statistically significant.

The SRD values of the NEADL (12.0) and the FAI (6.7) reported in this study appear quite large for a single subject with chronic stroke. Putting it in practical context, in order for a patient to improve 12.0 points in the NEADL, the patient needs to achieve 1 level of improvement (eg, progressing from "able to perform the activity with some difficulty" to "able to perform the activity easily") in at least 12 items or to achieve 2 levels of improvement in at least 6 items among the total of 19 items. Similar conditions were also noted in the FAI test. Thus, prospective users should bear in mind these substantial SRD values when using both measures to detect real change for a single subject.

Clinical trials usually report only group differences. However, in a given trial, significant group improvement does not guarantee that all patients in the group have achieved significant improvement. In other words, even though the mean changes within a study group are significant, the individual change of a substantial proportion of the study group might not achieve SRD at the individual level ($SRD_{individual}$). As previous authors have indicated,[4,8,22] ICC is a relative reliability index that only shows the reliability of ranking of positions of group members, while SRD is an absolute index that gives the "threshold" of real change. For future clinical trials, to help interpret research findings into clinical practice, researchers can report the proportion of patients who have achieved improvement beyond the SRD.

We compared the extent of the reproducibility of the 2 measures. The results showed that the ICC values of the NEADL and the FAI were high (ICC=.89 for both measures), indicating that both measures were equally reliable. The Bland and Altman graph (see fig 1) shows that the higher the NEADL score, the larger the variability between repeated assessments. Such a trend was not found in the FAI (see fig 1). We further compared the proportion of random measurement error relative to possible highest score (ie, SRD%) of these 2 measures. The SRD% of FAI (14.9%) was slightly lower than that of the NEADL (21.1%). These observations support that the FAI is more appropriate than the NEADL for clinical use.

Although both the NEADL and the FAI assess patients' IADL performance, they focus on different perspectives. The NEADL focuses on the extent of difficulty that a patient perceives in performing the IADLs. The FAI focuses on the frequency of a patient's performing the IADLs. However, we found very high associations (Pearson $r=.89$ at both sessions) between the scores of the NEADL and those of the FAI. Thus, the difference between both measures might be trivial for potential users.

We can also calculate the amount of random measurement error at the group level (ie, for research purpose). According to de Vet,[23] SRD at the group level (SRD_{group}) = $SRD_{individual} \div \sqrt{n}$ (n is the sample size); therefore, we can calculate the SRD_{group} value from the $SRD_{individual}$ value for a group level comparison. For example, the $SRD_{individual}$ for the NEADL in this study is 12.0 points, so the SRD_{group} will be 1.7 (for sample size of 52). It means that for a group of stroke patients, if the mean difference between 2 measurements is more than 1.7 points for the NEADL, we can claim that the change is real rather than the result of random measurement errors. The SRD_{group} is small given that the sample size is large. Thus, the SRD_{group} is seldom a concern when the sample size of a research study is substantial.

Table 2: The ICC, SRD, and SRD% of the NEADL and FAI

Test	1st Session	2nd Session	1st and 2nd Sessions	*d*	Standard Error of Measurement	CV (%)	ICC (95% CI)	SRD (SRD %)
NEADL	17.3±13.3	18.3±13.0	17.8±13.1	1.0±6.2	4.3	73.8	.89 (.82–.94)	12.0 (21.1)
FAI	9.3±7.4	9.3±7.2	9.3±7.3	0.0±3.5	2.4	78.4	.89 (.81–.93)	6.7 (14.9)

NOTE. Values are mean ± SD or as otherwise indicated.
Abbreviations: CI, confidence interval; CV, coefficient of variation; *d*, mean difference of scores between the 2nd and the 1st sessions.

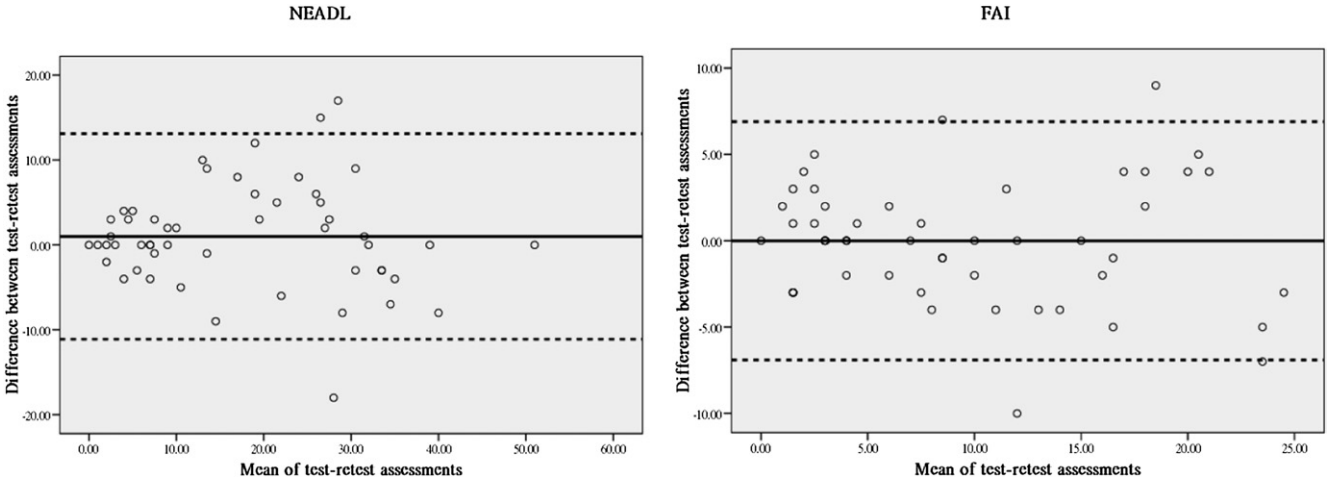

Fig 1. Bland and Altman graphs of the differences between 2 assessments plotted against the means of the 2 assessments of the NEADL and the FAI. The solid line represents the mean of the differences. The 2 dashed lines define LOA (mean of the difference ±1.96 SD).

Study Limitations

Our sample was a convenience sample. In addition, we recruited participants who could complete the interviews (ie, could coherently describe activities of their typical day), to make sure that their cognitive capabilities were good enough to answer our questionnaires. However, we did not perform a formal cognitive assessment. Our participants were mostly men and had mild disability. These characteristics of the sample might threaten the generalizability of our findings. Future studies that recruit patients with moderate or severe disability to further validate our findings may be needed.

CONCLUSIONS

Because of substantial SRD values of the NEADL and the FAI, prospective users should be cautious in using both measures to detect real change for a single subject.

References

1. World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001.
2. Hsueh IP, Huang SL, Chen MH, Jush SD, Hsieh CL. Evaluation of stroke patients with the extended activities of daily living scale in Taiwan. Disabil Rehabil 2000;22:495-500.
3. Schuling J, de Haan R, Limburg M, Groenier KH. The Frenchay Activities Index. Assessment of functional status in stroke patients. Stroke 1993;24:1173-7.
4. Flansbjer UB, Holmback AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. J Rehabil Med 2005;37:75-82.
5. Schuck P, Zwingmann C. The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. Int J Rehabil Res 2003;26:85-91.
6. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993;73:386-95; discussion 396-401.
7. Schreuders TA, Roebroeck ME, Goumans J, van Nieuwenhuijzen JF, Stijnen TH, Stam HJ. Measurement error in grip and pinch force measurements in patients with hand injuries. Phys Ther 2003;83:806-15.
8. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005; 19:231-40.
9. Chen HM, Chen CC, Hsueh IP, Huang SL, Hsieh CL. Test-retest reproducibility and smallest real difference of 5 hand function tests in patients with stroke. Neurorehabil Neural Repair 2009;23:435-40.
10. Nouri FM, Lincoln NB. An extended activities of daily living scale for stroke patients. Clin Rehabil 1987;1:301-5.
11. Lincoln NB, Gladman JR. The Extended Activities Of Daily Living Scale: a further validation. Disabil Rehabil 1992;14:41-3.
12. Holbrook M, Skilbeck CE. An activities index for use with stroke patients. Age Ageing 1983;12:166-70.
13. Piercy M, Carter J, Mant J, Wade DT. Inter-rater reliability of the Frenchay Activities Index in patients with stroke and their careers. Clin Rehabil 2000;14:433-40.
14. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10:571-8.
15. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000;30:1-15.
16. Stratford P. Reliability: consistency or differentiating among subjects? Phys Ther 1989;69:299-300.
17. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Pr; 1993.
18. Smidt N, van der Windt DA, Assendelft WJ, et al. Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis. Arch Phys Med Rehabil 2002;83:1145-50.
19. Portney LG, Watkins MP. Foundations of clinical research: application to practice. Upper Saddle River: Prentice-Hall; 2000.
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
21. Prince B, Makrides L, Richman J. Research methodology and applied statistics. Part 2: the literature search. Physiother Can 1980;32:201-6.
22. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. Phys Ther 2006;86:735-43.
23. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures: theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001;17:479-87.

Supplier

a. SPSS Inc, 233 S Wacker Dr, 11th Fl, Chicago, IL 60606.