Inter-Rater Differences in Post-Stroke Rehabilitation Assessments: Towards Optimising Stroke Recovery

Mohd Azri Abd Mutalib[1], Norsinnira Zainul Azlan[2], Nor Mohd Haziq Norsahperi[3], Hafizu Ibrahim Hassan[4]

[1]*Advanced Automation Section, Smart Manufacturing Centre, SIRIM Berhad, Lot 1A, Persiaran Zurah, Kawasan Perindustrian Rasa, 44200 Rasa, Selangor, Malaysia*

[2]*Department of Mechatronics Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia*

[3]*Department of Electrical & Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

[4]*Department of Mechatronics Engineering, Ahmadu Bello University, 810211 Zaria, Nigeria*

mohdazri@sirim.my, sinnira@iium.edu.my, nmhaziq@upm.edu.my, sirhafiz01@gmail.com

Article history

Received:
20 Feb 2025

Received in revised form:
20 Jul 2025

Accepted:
14 Aug 2025

Published online:
30 Aug 2025

*Corresponding author
mohdazri@sirim.my
sinnira@iium.edu.my

Abstract

Stroke assessment is a crucial process for evaluating post-stroke patients' functional abilities, offering essential insights into recovery progress, residual impairments, and the capacity to perform daily activities. These assessments are designed to identify specific deficits in motor function, coordination, and strength, facilitating the development of tailored rehabilitation plans. However, the reliance on manual methods and subjective judgments often leads to inconsistencies, as evaluations heavily depend on individual therapists' perceptions. This subjectivity introduces variability, where the same patient's performance may be interpreted differently by different therapists, resulting in inconsistent outcomes. Inter-rater variability further compounds the issue, with differences in scoring methods and interpretations among therapists. This study is conducted to determine the impact of these variations on stroke rehabilitation assessments. A total of 56 stroke subjects are assessed by three therapists with varying levels of experience from Sultan Ahmad Shah Medical Centre (SASMEC), Kuantan, Pahang. The assessment focuses on the patients' ability to perform 10 Activities of Daily Living (ADLs) derived from the Motor Activity Log (MAL), an established standard clinical assessment protocol. Statistical analysis using ANOVA indicates that the p-values for the Amount of Use (AOU) and Quality of Movement (QOM) scales are 0.189 and 0.515, respectively, demonstrating no statistically significant differences among the therapists' ratings despite observed variations. These findings suggest that the differences among therapists are not substantial to undermine the reliability of the assessment process. To enhance consistency and reduce individual biases, averaging the scores from multiple therapists is recommended as an effective approach for obtaining a more robust and standardised representation of a patient's functional abilities.

Keywords: *Activity of Daily Living (ADL), ANOVA analysis, Inter-rater differences, Motor Activity Log (MAL), Post-stroke assessment.*

* *Corresponding author mohdazri@sirim.my;sinnira@iium.edu.my*

1. Introduction

Stroke is one of the leading causes of long-term disability worldwide, affecting millions of individuals each year. Rehabilitation plays a vital role in restoring lost functions and enhancing the quality of life for post-stroke patients. Accurate and consistent assessment of functional abilities is essential for designing effective rehabilitation plans, tracking recovery progress, and evaluating therapeutic interventions.

Various established clinical assessment tools are available and widely used to evaluate the abilities of post-stroke patients, providing valuable insights into their recovery progress. Below is a brief overview of five standard clinical assessments commonly used in hospitals and rehabilitation centers. Further explanations are provided in Table 1.

a. Action Research Arm Test (ARAT):
The ARAT is used to assess upper limb motor function in stroke patients, specifically evaluating their ability to perform tasks such as grasping, gripping, and lifting objects. It is widely employed to track rehabilitation progress related to arm and hand functionality [1].

b. Barthel Index (BI):
The Barthel Index measures a patient's ability to perform ten basic activities of daily living (ADLs), including feeding, bathing, and walking. This tool provides an overall score reflecting the patient’s level of functional independence [2][3].

c. Fugl-Meyer Assessment (FMA):
The Fugl-Meyer Assessment is a comprehensive tool for evaluating physical function in post-stroke patients. It assesses motor function, balance, sensation, and joint range of motion, providing a detailed score that quantifies physical impairments across multiple domains [1][2][4][5].

d. Motor Activity Log (MAL):
The Motor Activity Log assesses a patient’s ability to use their affected limb in daily activities. It focuses on the Amount of Use (AOU) and Quality of Movement (QOM) during tasks, offering valuable insights into functional recovery and progress in stroke rehabilitation [1][2][6][7].

e. Wolf Motor Function Test (WMFT):
The Wolf Motor Function Test evaluates upper limb motor function by measuring the time it takes to complete specific tasks involving the arm and hand. It simultaneously assesses both the speed and quality of movement, making it an essential tool for tracking motor recovery and the effectiveness of rehabilitation interventions [2][6][7].

Table 1. Standard Clinical Assessments for Post-Stroke Patient Assessment

Clinical Assessment	Focus	Task	Scale	Duration for normal subject	Relevant Information
ARAT	Upper limb motor function	Grasping, gripping, pinching, and reaching tasks; includes functional tasks like picking up objects, stacking, and lifting.	0-cannot performs 1-partially performs 2-long time to completely performs 3-normally performs	5-10 minutes	Designed for upper limb recovery in stroke patients, focuses on the hand and arm functions
BI	ADL	Feeding, bathing, dressing, grooming, toileting, transfers, walking, climbing stairs, and bowel and bladder control	1-cannot performs 2-attempts task but unsafe 3-moderate help required 4-minimal help required 5- fully independent	20 minutes to an hour	Provides an overall score of functional independence; often used for determining the level of assistance needed
FMA	Physical function, including motor function, balance, and sensation	Motor function (upper and lower limbs), balance, sensation (light touch, proprioception), joint range of motion.	0- cannot perform. partially perform 2- fully perform	30-45 minutes	Comprehensive, including multiple domains (motor, balance, sensation); widely used in stroke rehabilitation studies
MAL	Functional use of the affected limb	The amount of use (AOU) and quality of movement (QOM) of the affected arm during ADLs.	0-never 1-very rarely 2-rarely 3-fair 4-almost normal	20 minutes	Focused on how well the patient uses their affected arm in everyday tasks, providing insight into the

			5-normal		functional recovery of the arm
WMFT	Upper limb motor function, speed, and quality of movement	15 functional tasks like lifting objects, moving items, and transferring them. Includes time to complete tasks and quality of movement ratings.	0-cannot perform -partially perform 2-perform with assistance 3-perform but slowly 4-perform with a slightly slow 5-normally perform	15-30 minutes	Measures both speed (timed) and quality of movement; useful for assessing rehabilitation progress in the upper limbs

Stroke assessments are often influenced by subjective judgments, as they primarily depend on therapists' observations and interpretations, which can result in variability in scoring and assessment outcomes. Although standardised assessment protocols exist, significant variability persists. Differences in therapists' experience, training, and perception can lead to inconsistent ratings, undermining the reliability and accuracy of the assessments. This inter-rater variability is a crucial issue that must be addressed to enhance the consistency and objectivity of stroke assessments.

This study uniquely contributes to the field by investigating the impact of inter-rater differences in stroke rehabilitation assessments. It specifically examines whether variations in therapists' rating significantly influence assessment outcomes and explores methods to address these discrepancies to improve the standardisation and reliability of stroke rehabilitation assessment protocols. This contribution is crucial for ensuring consistency in assessment, particularly in clinical and research settings.

The following sections present the methodology used to assess 56 stroke patients by three therapists with varying experience, the statistical analysis performed to evaluate inter-rater variability, and the implications of the findings for improving stroke assessment practices. The conclusion highlights recommendations for standardising stroke assessments to reduce bias and enhance reliability.

2. Methodology

The Motor Activity Log (MAL) is employed in this study due to its widespread use at Sultan Ahmad Shah Medical Centre (SASMEC) and its strong relevance to ADLs. A total of ten ADLs are selected, with eight tasks directly extracted from the MAL and two additional tasks proposed by SASMEC therapists based on their popularity and relevance to daily activities. Additionally, the MAL standard includes the flexibility to incorporate “other optional activities” as needed. All ten

activities are evaluated using the MAL scoring scale. Table 2 presents an overview of the MAL scale, which includes the Amount of Use (AOU) and Quality of Movement (QOM) measures, while Table 3 lists the ten ADLs utilised in this study along with their terminology used and respective sources.

Table 2. The MAL Scale.

Score	AOU	QOM
0	The weaker arm was not used at all for that activity.	The weaker arm was not used at all for that activity
1	Occasionally used the weaker arm, but only very rarely.	The weaker arm was moved during the activity but was not very helpful.
2	Sometimes used the weaker arm but did the activity most of the time with the stronger arm.	The weaker arm was of some use during the activity but needed some help from the stronger arm but moved very slowly or with difficulty.
3	Used the weaker arm about half as much as before the stroke.	The weaker arm was used for that activity, but the movements were slow or were made only with some effort.
4	Used the weaker arm almost as much as before the stroke	The movements made by the weaker arm for the activity were almost normal but not quite as fast or accurate as normal
5	Used the weaker arm as often as before the stroke.	The ability to use the weaker arm for that activity was as good as before the stroke.

Table 3. 10 ADLs Utilised with Their Respective Sources.

No	ADLs	Terminology	Source
1	Engage and release plug top	Plug	Therapist suggestion
2	Turning on a light switch	Switch	MAL
3	Turning a fan regulator	Fan	Therapist suggestion
4	Turning a water faucet	Faucet	MAL
5	Turning a doorknob	Doorknob	MAL
6	Opening a drawer	Drawer	MAL
7	Opening a door	Door	MAL
8	Combing a hair	Comb	MAL
9	Using a spoon for eating	Spoon	MAL
10	Brushing a tooth	Toothbrush	MAL

The IIUM Research Ethics Committee (IREC) has approved data collection from under the approval number IREC 2023-078. Data collection involves 56 post-stroke patients across all scales in the MAL as study subjects. The data is limited to post-stroke patients who have regained a minimum level of motor power and can lift their affected hand. Patients who cannot lift their affected hand are excluded from the scope of this study. Each subject is required to perform all ten ADLs in every session, with the test conducted for three times for consistency. The ten ADLs are standardised, and the sequence of tasks was kept consistent for all participants to ensure uniformity in testing across all subjects.

Three therapists from SASMEC with professional's experience and extensive clinical expertise, participated in the assessments. Each therapist conducted the assessment independently, aligned with the three trials performed by each subject. To eliminate bias, the scores provided by one therapist were not disclosed to the other two therapists. The assessments are carried out in a controlled setting, with all therapists using the same standardised tools to evaluate the patients. The results are systematically documented to analyse the level of agreement and identify any discrepancies in the therapists' ratings.

3. Result

ANOVA analysis is performed on the collected data from post-stroke patients to evaluate differences in scores assigned by the therapists for each subject. Table 4 presents a summary of the ANOVA results for the ratings given by the three therapists. To further analyse the differences in scores, the Estimated Marginal Means (EMM) of the ratings provided by the three therapists are calculated and presented in Table 5. T1, T2, and T3 denote Therapist 1, Therapist 2, and Therapist 3, respectively. The corresponding graphs for AOU and QOM are illustrated in Figures 1 and 2.

Table 4. ANOVA Analysis for The Three Therapist Rating.

	Dependent variable	
	AOU	QOM
Sum of square (SS)	10.80	3.68
Degrees of freedom (df)	2	2
Mean square	5.40	1.84
F Statistic	1.67	0.66
Sig./ p-value	0.19	0.52

Table 5. The EMM of Three Therapist Rating.

ADLs	T1 AOU	T1 QOM	T2 AOU	T2 QOM	T3 AOU	T3 QOM
Plug	2.45	2.46	2.52	2.46	2.50	2.61
Switch	3.34	3.32	3.46	3.36	3.43	3.36
Fan	2.23	2.46	2.25	2.32	2.18	2.32
Faucet	3.46	3.20	3.45	3.27	3.59	3.36

Doorknob	2.93	3.13	2.98	3.00	3.02	2.88
Drawer	3.23	3.20	3.30	2.96	3.38	3.20
Door	3.14	3.14	3.27	3.27	3.34	3.32
Spoon	2.96	2.91	3.07	3.14	3.11	3.07
Comb	2.95	3.07	3.13	3.16	3.16	3.30
Toothbrush	3.00	3.25	3.13	3.32	3.07	3.21

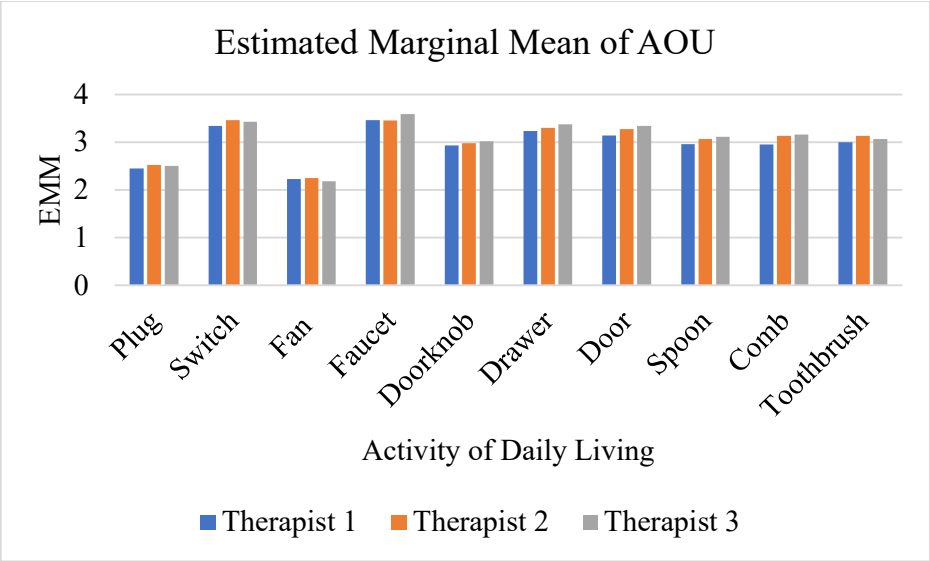

Figure 1: EMM for AOU

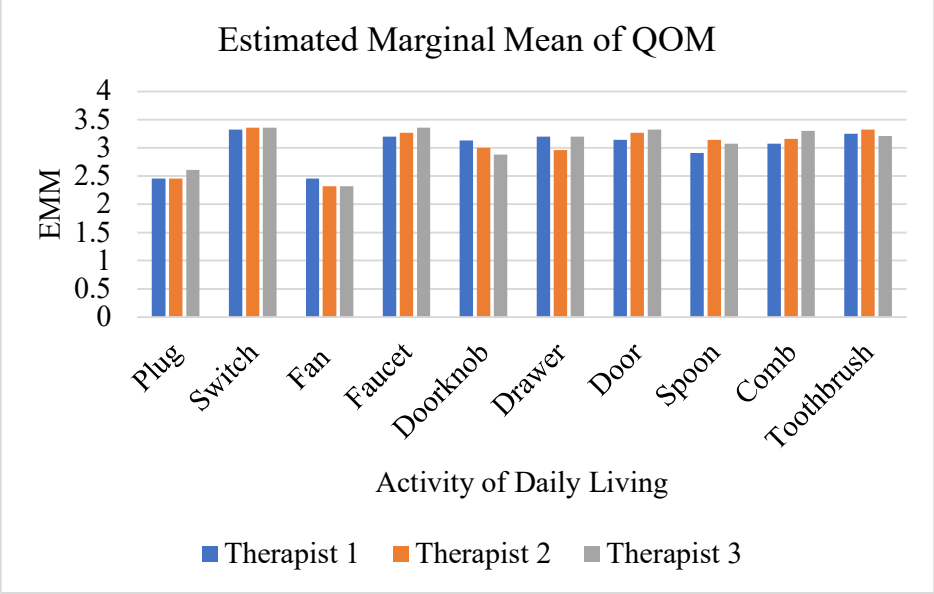

Figure 2: EMM for QOM

Since the differences in ratings among the three therapists are found to be statistically non-significant, the average rating across therapists is used as a reference for comparison. To evaluate the consistency of each therapist's rating relative to the average, the percentage of Root Mean Square Error (RMSE) is calculated separately for AOU and QOM, as they represent distinct outcome measures in the MAL. Table 6 presents the resulting percentage RMSE values for each therapist in relation to the average score per ADL.

Table 6: RMSE between the average and therapist's individual ratings.

ADL		RMSE (%)		
		T1	T2	T3
Plug	AOU	11.95	10.00	8.02
	QOM	11.02	7.07	8.02
Switch	AOU	10.00	5.98	5.98
	QOM	8.86	7.07	4.63
Fan	AOU	9.26	8.86	**2.67**
	QOM	14.14	11.34	10.00
Faucet	AOU	10.69	8.018	5.98
	QOM	9.64	8.02	9.26
Doorknob	AOU	10.70	8.02	7.07
	QOM	12.82	5.35	5.98
Drawer	AOU	10.36	8.86	7.07
	QOM	9.26	11.65	3.78
Door	AOU	12.82	7.56	6.55
	QOM	17.32	11.20	10.00
Spoon	AOU	12.82	10.35	10.35
	QOM	14.64	5.98	8.86
Comb	AOU	12.25	9.64	7.07
	QOM	10.35	9.26	10.69
Toothbrush	AOU	10.35	10.00	**15.35**
	QOM	9.26	8.45	10.69

4. Discussion

Table 4 presents the ANOVA results, which indicate that there are no statistically significant differences in the ratings from the three therapists for both AOU and QOM. This is supported by the p-values of 0.19 for AOU and 0.52 for QOM, both of which exceed the commonly accepted threshold of 0.05. A p-value greater than 0.05 suggests that any observed differences in the means are likely due to random variation rather than a genuine effect. The sum of squares (SS) values for AOU (10.80) and QOM (3.68) represent the total variation explained by the model, while the corresponding mean square values (5.40 for AOU and 1.84 for QOM) are insufficient to yield significant results in this context. These findings suggest that the variance observed across the therapists' ratings is minimal and does not reflect a substantial difference in their assessments.

Table 5, Figures 1 and 2 illustrate the ANOVA data analysis in terms of estimated marginal means (EMM). The results show a high degree of inter-rater reliability in both the AOU and QOM ratings across the three therapists (T1, T2, and T3). Specifically, the AOU ratings show minimal inter-rater variability, as supported by the closeness of the ratings across therapists and the non-significant p-values (greater than 0.05). This consistency indicates a strong agreement among the therapists on their assessments of the tasks, reinforcing the reliability of the AOU evaluation in this context.

Similarly, the QOM ratings demonstrate strong alignment across all three therapists, further supporting the notion of good inter-rater reliability in assessing the quality of movement. The minor discrepancies observed across tasks are likely attributable to thin differences in individual perception or the inherent complexity of specific tasks, rather than significant inconsistencies among the therapists.

The overall consistency observed across therapists for both AOU and QOM ratings suggests that these assessments are relatively standardised. Any variations in task-specific scores are likely due to factors such as the inherent difficulty of the tasks being assessed or the performance of the patients themselves, rather than inconsistencies in the raters' assessments.

Table 5 presents the RMSE values between the average and individual therapist ratings. The lowest error is observed with Therapist 3 (Fan, AOU) at 2.67%, while the highest is 15.35% for Therapist 3 (Toothbrush, AOU). RMSE values ranging from 2.67% to 15.35% can be considered within an acceptable range. According to [8], the general guide for acceptable inter-rater accuracy is between 71% to 99%, with physical human assessments ideally falling within the 80% to 90% range, which is considered good and acceptable. Accuracy greater than 90% is regarded as excellent [9]. Based on the results in Table 5, the lowest observed accuracy is 85.65%, which is already within the acceptable range for human physical assessments.

However, to achieve even better results, this research should be continued. The subjective qualitative approach should be upgraded to a more standardised quantitative method. The use of more consistent sensors for recording post-stroke patient parameters could further enhance the consistency of rehabilitation assessment. Additionally, the adaptation of advanced technologies such as Artificial Intelligence (AI), Machine Learning Algorithm (MLA), and Deep Learning (DL) to analyse the data could lead to more accurate and reliable results. Therefore, this research should be continued to maximise its potential impact on the post-stroke rehabilitation field.

5. Conclusion

This study demonstrates strong inter-rater reliability in the MAL assessments of AOU and QOM ratings by three therapists, as indicated by the ANOVA results. The absence of statistically significant differences across the therapists' ratings, supported by non-significant p-values, reinforces the reliability and consistency of the assessment process. Additionally, the relatively low RMSE values between the average ratings and individual therapist ratings suggest that the assessments are

accurate and consistent, with the observed errors falling within an acceptable range for physical human assessment.

Despite these promising findings, further improvements can be made to achieve even greater accuracy and reliability. Transitioning from a subjective qualitative approach to a more standardised quantitative method will improve the consistency of assessments. Additionally, incorporating advanced technologies such as Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) can optimise data analysis, leading to more precise and reliable outcomes. Continuing this research is essential to further enhancing rehabilitation assessments and maximising its impact on the rehabilitation field.

Acknowledgments

The authors would like to extend their appreciation to Sultan Ahmad Shah Medical Centre (SASMEC) for facilitating the data collection involving post-stroke patients at their rehabilitation centre.

References

[1] M. F. Tsai, R. H. Wang, and J. Zariffa, "Validity of Novel Outcome Measures for Hand Function Performance After Stroke Using Egocentric Video," *Neurorehabil. Neural Repair*, vol. 37, no. 2–3, pp. 142–150, 2023, doi: 10.1177/15459683231159663.

[2] Y. Z. Shou, X. H. Wang, and G. F. Yang, "Verum versus Sham brain-computer interface on upper limb function recovery after stroke: A systematic review and meta-analysis of randomized controlled trials," *Med. (United States)*, vol. 102, no. 26, p. E34148, 2023, doi: 10.1097/MD.0000000000034148.

[3] S. Ferfeli, A. Galanos, I. A. Dontas, A. Triantafyllou, I. K. Triantafyllopoulos, and E. Chronopoulos, "Reliability and validity of the Greek adaptation of the Modified Barthel Index in neurorehabilitation patients," *Eur. J. Phys. Rehabil. Med.*, vol. 60, no. 1, pp. 44–54, 2024, doi: 10.23736/S1973-9087.23.08056-5.

[4] M. Goliwąs, J. Małecka, K. Adamczewska, M. Flis-Masłowska, J. Lewandowski, and P. Kocur, "Polish Cultural Adaptation and Reliability of the Fugl-Meyer Assessment of Motor Performance and Sensory Assessment Scale in Stroke Patients," *J. Clin. Med.*, vol. 13, no. 13, 2024, doi: 10.3390/jcm13133710.

[5] B. P. Huynh *et al.*, "Sensitivity to Change and Responsiveness of the Upper Extremity Fugl-Meyer Assessment in Individuals With Moderate to Severe Acute Stroke," *Neurorehabil. Neural Repair*, vol. 37, no. 8, pp. 545–553, 2023, doi: 10.1177/15459683231186985.

[6] A. Omer *et al.*, "Effects of Kinesio Taping and Modified Constraint-Induced Movement Therapy on Upper Extremity Function, Quality of Life, and Spasticity in Individuals Recovering from Stroke," *J. Heal. Rehabil. Res.*, vol. 4, no. 1, pp. 167–172, 2024, doi: 10.61919/jhrr.v4i1.347.

[7] P. Psychouli, I. Mamais, and C. Anastasiou, "An Exploration of the Effectiveness of Different Intensity Protocols of Modified Constraint-Induced Therapy in Stroke: A Systematic Review," *Rehabil. Res. Pract.*, vol. 2023, no. 3, 2023, doi: 10.1155/2023/6636987.

[8] K. A. Szucs and E. V. D. Brown, "Rater reliability and construct validity of a mobile application for posture analysis," *J. Phys. Ther. Sci.*, vol. 30, no. 1, pp. 31–36, 2018, doi: 10.1589/jpts.30.31.

[9] G. Carta *et al.*, "Discovering the Vagus: Validation and Inter-Rater Reliability of the Vagus Nerve Neurodynamic Test Among Healthy Subjects," *SSRN, the LANCET*, 2020, [Online]. Available: https://ssrn.com/abstract=3529450