J Rehabil Med 2016; 48: 265–272

ORIGINAL REPORT

RATER EXPERIENCE INFLUENCES RELIABILITY AND VALIDITY OF THE BRIEF INTERNATIONAL CLASSIFICATION OF FUNCTIONING, DISABILITY, AND HEALTH CORE SET FOR STROKE

Shanjia Chen, MSc[1,2#], Jing Tao, PhD[1,2,3#], Qian Tao, PhD[4,5], Yunhua Fang, MSc[1,2], Xiaoxuan Zhou, MSc[1,2], Hongxia Chen, PT, MSc[6], Zhuoming Chen, OT, PhD[7], Jia Huang, OT, PhD[1,2], Lidian Chen, PhD[1,2,3] and Chetwyn C. H. Chan, PhD[5]

From the [1]College of Rehabilitation Medicine, Fujian University of Traditional Chinese Medicine, [2]Fujian Key Laboratory of Exercise Rehabilitation, Fujian provincial rehabilitation industrial institution, [3]Affiliated Rehabilitation Hospital, Fujian University of Traditional Chinese Medicine, Fuzhou, [4]Psychology Department, School of Medicine, Jinan University, Guangzhou, [5]Applied Cognitive Neuroscience Laboratory, Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, [6]The Second Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine and [7]The First Affiliated Hospital of Jinan University, Guangzhou, China. [#]These authors contributed equally to this work.

Objective: **To investigate how clinical experience and access to patient information regarding functional capability influence inter-rater reliability and validity of the Brief International Classification of Functioning, Disability, and Health Core Set for Stroke (ICF) assessment.**

Methods: **Study 1 involved expert (clinical experience ≥5 years) and novice (clinical experience <2 years) rater-pairs, each evaluating the same post-stroke patients using the ICF assessment (*n* = 149). Study 2 involved novice raters separately evaluating a different cohort of post-stroke patients with the ICF assessment (*n* = 78). The novice raters had prior knowledge of patient functioning through conducting 6 clinical tests.**

Results: **For Study 1, the expert rater-pairs (kappa = 0.50–0.85 for categories; intra-class correlation (ICC) = 0.76–0.96 for components) had higher reliability coefficients than novice rater-pairs (kappa = 0.18–0.69 for categories; ICC = 0.63–0.88 for components). For Study 2, the novice raters with prior knowledge of patient's functioning yielded significantly higher ICF component scores than those without prior knowledge. The former raters' component scores were comparable to those of the expert rater-pairs.**

Conclusion: **Clinical experience in post-stroke rehabilitation enhances inter-rater reliability of ICF assessment. Knowledge of patient's functional capability, such as conducting common clinical tests in post-stroke rehabilitation, is useful for improving assessment validity.**

Key words: ICF; stroke; reproducibility of results; clinical experience; disabled persons.

J Rehabil Med 2016; 48: 265–272

Correspondence address: Lidian Chen, College of Rehabilitation Medicine, Fujian University of Traditional Chinese Medicine,1 Huatuo Road, Minhou Shangjie, Fuzhou, Fujian 350122, China. E-mail: cld@fjtcm.edu.cn and Qian Tao, Psychology Department, School of Medicine, Jinan University, 510632, Guangzhou, China. E-mail: tracy.tao@connect.polyu.hk

Accepted Dec 18, 2015; Epub ahead of print Feb 17, 2016

INTRODUCTION

The International Classification of Functioning, Disability, and Health (ICF) Core Set for Stroke captures prototypical problems relevant to post-stroke patients irrespective of their stage of rehabilitation (1, 2). The Brief Core Set for Stroke includes 18 second-level categories grouped into 4 components: 6 categories for Body Function (BF), 2 for Body Structure (BS), 7 for Activity and Participation (AP), and 3 for Environmental Factors (EF) (3). The generic qualifier scale under each category quantifies severity of impairment, which has 5 response levels ranging from "0" (no impairment) to "4" (completely impairment). The instructions for test administration suggest that health professionals gather information from clinical records in order to facilitate assigning scores on the ICF (4). The instructions, however, do not stipulate how this could impact the reliability of the assessment or any requirements for rater experience and background. This study aimed to reveal how different clinical experience levels of raters and access to patient functional capacity information influences the assignment of scores on the Brief ICF Core Set for Stroke (hereafter called the ICF assessment).

Consistency of scores across raters is an inter-rater reliability issue. Clinical experience of raters is a key factor contributing to accuracy and reliability of results generated from clinical assessments (5, 6). For instance, the reliability of fracture classifications and accuracy in making clinical decisions can be improved with clinical experience (7). Common measures of clinical experience are years of clinical practice, level of professional training, credentials, and seniority in the field (7, 8). Clinicians with more experience have been found to be able to relate the case at hand to similar cases previously encountered (9). Researchers explained that more experience would enable clinicians to develop higher reasoning ability for identifying social/behavioural cues beyond patient self-reports (10). Another study postulated that an advantage of having more clinical experience is the accumulation of a larger dataset of patient-related problems and solutions (11).

 doi: 10.2340/16501977-2063

 ISSN 1650-1977

Previous reliability studies on the ICF have reported low to moderate consistency among raters (12–15). One study on the difference of test-retest reliability of the ICF assessment between novice and experienced raters indicated that more experienced raters had significantly higher consistency than less experienced raters on BF and AP categories among older adults (12). Another study focused on influences of confidence and core competence of raters on inter-rater reliability with the Extended ICF Core Set for Stroke (13) and the results indicated that neither variable had significant effects on reliability. The researchers postulated that factors that could have influenced consistency would be rater clinical experience, skills, and knowledge about stroke. These higher level attributes would enable clinicians to observe and identify patient problems relevant to ICF factors and categories. Nevertheless, those postulations were not substantiated by their findings. Besides clinical experience, the type of information that raters can access would influence their assignment of scores for performance during clinical assessment. The different ways of collecting patient information, such as direct observation and clinical records, might influence scoring of ICF categories (16, 17), thus affecting their validity.

The aim of the current study was to investigate how clinical experience and access to patient information regarding functional capability influences inter-rater reliability and validity on the ICF assessment. Two research questions were investigated: (*i*) how raters with different clinical experience (more or less) influence the reliability of ICF assessment; (*ii*) how knowledge of patient functional capability influences the validity of ICF assessment. Study 1 investigated the impact of clinical experience on inter-rater reliability of the ICF assessment in a group of post-stroke patients. Study 2 investigated how common rehabilitation clinical tests administered by less experienced therapists influenced ICF assessment results and hence, its validity. It was hypothesized that raters who had more clinical experience would yield higher inter-rater reliability indices than those with less experience (Study 1). Equipping rehabilitation therapists with knowledge of patient functional capability is posited to positively affect the validity of ICF assessment scores among raters with less clinical experience (Study 2).

METHODS

Participants

The 10 raters participating in this study were rehabilitation therapists working in 2 rehabilitation in-patient settings located in Southern China. Among them, 6 had less clinical experience (mean 1.45 years; range 0.9–1.5 years) and 4 had more clinical experience (mean 8.65 years; range 7.9–9 years). Raters with less experience (< 2 years) in stroke rehabilitation were defined as novices, and raters with more experience (≥ 5 years) were defined as experts. All raters received undergraduate training in rehabilitation and were certified as rehabilitation therapists. None of the raters had experience in administering the ICF assessment prior to this study. Study 1 involved 2 expert and 2 novice rater-pairs. Study 2 involved another 2 novice raters, who adopted a reverse assessment sequence to that of novice raters in Study 1. There was no significant difference between the 2 studies in the years of clinical experience for novice raters.

Participants were post-stroke in-patients recruited from the 2 rehabilitation settings. Inclusion criteria were: (*i*) diagnosed as first stroke and confirmed by brain scan; (*ii*) aged between 40 and 80 years; (*iii*) 4–12 weeks post-stroke and receiving active rehabilitation; (*iv*) hemiplegia resulting in contralesional paralysis; and (*v*) severe neurological function according to the National Institutes of Health Stroke Scale (NIHSS) (score ≥ 6) (18). Patients were excluded if they had aggravated disease or failed to complete the clinical assessment protocol. A total of 268 patients were screened; 32 did not meet the inclusion criteria, and 9 did not complete all assessments. The final sample size was 227 (37.0% female) with a mean age of 61.7 years (Table I). The participants were randomly assigned to expert rater-pairs (Study 1, $n = 72$), novice rater-pairs without prior knowledge of patient functioning (Study 1, $n = 77$), and single novice rater with prior knowledge of patient functioning (Study 2, $n = 78$). The study obtained ethics approval from the Institutional Review Board of each clinical setting. Written informed consent was obtained from all participants or their proxies.

Instruments

ICF assessment. The ICF assessment consists of 18 categories, each of which is accompanied by a detailed description of its content. The generic qualifier serves as the rating scale for the BF, BS, and AP components (13), which has 5 response options ranging from "0" to "4" (no/mild/moderate/severe/complete impairment) indicating the extent of impairment (Table II). The EF qualifier has 9 response options ranging from "–4" to "4". Options "–4" to "–1" denote different levels of environmental barrier; "0" denotes no influence; options "1" to "4" denote different levels of environmental facilitator enhancing the function of the individual; option "8" denotes not specified and option "9" denotes not applicable. The raters conduct the assessment by means of: (*i*) observing individual's performance in executing a task or a movement in life situations; (*ii*) interviewing the individual and his/her proxies; (*iii*) asking questions such as "Where are you now" or "What is the time?" for evaluating orientation; and (*iv*) reviewing clinical history and clinical examination results. Options "8" and "9" were treated as nominal categories in Cohen's kappa reliability statistics. A category score was derived directly from the response option of the qualifier (not including "8" and "9"). Category scores were summed to form component scores for BF, BS, AP, and EF. A higher score for the BF, BS, AP and a lower score for the EF indicated more serious dysfunction and barrier conditions in patients.

Table I. *Demographic and clinical characteristics of the participants allocated to each of the groups of expert rater-pairs, novice rater-pairs and single novice rater*

	Expert rater-pairs ($n = 72$)	Novice rater-pairs ($n = 77$)	Single novice rater ($n = 78$)	F($\times^2$), sig
Age, years, mean (SD)	62.6 (11.7)	62.9 (10.0)	59.5 (12.1)	2.09, 0.13
Sex, male, %	70.8	61.0	61.5	1.96, 0.38
Education, years, mean (SD)	7.7 (4.8)	8.0 (4.2)	8.8 (5.0)	1.19, 0.31
Onset of stroke, days, mean (SD)	45.2 (19.3)	39.8 (15.7)	46.1 (17.7)	2.87, 0.06
Ischaemic stroke, %	75.0	71.4	69.2	0.62, 0.73
Affected side, left, %	48.6	48.1	38.5	0.42, 0.81
NIIHSS, mean (SD)	10.4 (4.2)	10.3 (3.4)	10.5 (3.8)	0.18, 0.91

NIHSS: NIH stroke scale; SD: standard deviation.

Table II. *Generic qualifier scale for stroke assessment and a general guide for scoring*

Score	Description	Adaptation description
0	Normal	Enable to execute a task or an action independently.
1	Mildly impaired	Enable to execute most part of a task or an action with minimum assistance from others such as supervision.
2	Moderate impaired	Enable to execute some part of a task or an action with moderate assistance from others.
3	Severe	Enable to execute a small part of a task or an action with maximal assistance from others. For instance, 2 people are required to execute a task.
4	Complete impairment	Need total assistance to execute a task or an action.

Short Form Health Survey (SF-36). The SF-36 (19) is widely used for assessing health-related quality of life. It is a self-administered questionnaire that contains 36 categories grouped under 8 dimensions of health: physical functioning, role limitations due to physical problems, bodily pain, general health perception, vitality, and social function, role limitations due to emotional problems, and mental health, and 1 single-category scale on health change over the past year (19). A concurrent study was conducted to translate the SF-36 into Chinese (Appendix SI[1]).

Other clinical tests. Patient function was assessed with 5 other common clinical tests, all of which have been validated for usage in Chinese populations in previous studies: the Montreal Cognitive Assessment (MoCA) measures the overall cognitive profile (20), the Chinese version of Modified Barthel Index (MBI-C) assesses self-care independence (21), the Chinese version of Frenchay Activity Index (FAI-C) measures activities of daily living (22), the brief version of the Fugl-Meyer Assessment (FMA) assesses upper and lower limb functions (23), and the modified Rankin Scale (mRS) measures global disability and prognosis (24).

Procedure

In Study 1, expert and novice rater-pairs administered the ICF assessment to each patient assigned to their group (Fig. 1). One rater from each pair contacted the patient and completed the first ICF assessment on the second day after subject recruitment. Then, the second rater in the same rater-pair administered the ICF assessment on the same patient within 24 h to minimize possible changes in patient functional status with time. After completing both ICF assessments, 1 rater in each pair randomly conducted 6 clinical tests over the next 3 days. The sequence for administering the tests was fixed: MoCA, SF-36, MBI-C, FMA, FAI-C, and mRS. The sequence in which raters within each rater-pair completed the ICF assessment, as well as the rater who administered the 6 clinical tests, were randomly determined by drawing lots. The assessment design and testing schedule were the same for both expert and novice rater-pairs at each institution.

Study 2 involved an additional 2 single novice raters who conducted their patient assessments in reverse of Study 1. Each single novice rater started to contact the patient on the second day after patient recruitment and completed the 6 clinical tests within the next 3 days. The ICF assessment was then administered immediately after completing these 6 clinical tests. The procedures for completing all 6 clinical tests were the same as those followed by raters in Study 1.All raters from both studies received training on administering the ICF assessment and 6 clinical tests. Manuals that described test administration procedures were distributed to all raters. Training began with explaining the purpose and content of each test. Demonstration of all testing procedures involved real patients. The training session lasted for 3 days. ICF assessment training included familiarization with the ICF framework and classifications and illustrating methods for interviewing patients and proxies. Raters were familiarized with the types and format of clinical information contained in patient case files, such as results of clinical and laboratory examinations conducted by their medical and rehabilitation teams. All assessments were conducted in a quiet room within the rehabilitation department. Only a few patients were too weak to walk (or use a wheelchair) to the room, and for these subjects

[1]http://www.medicaljournals.se/jrm/content/?doi=10.2340/16501977-2063

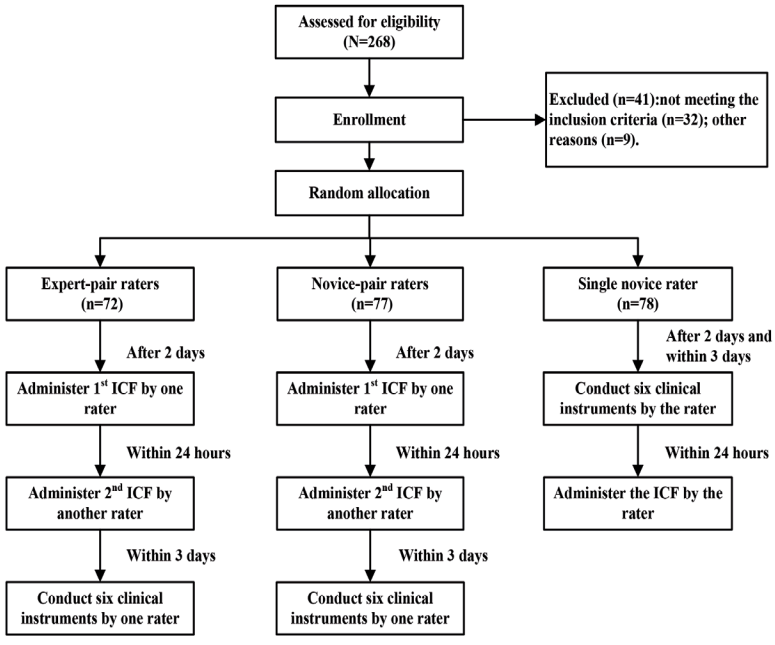

Fig. 1. Study design and summary of implementation of the study.

data collection was carried out at the bedside. As a control, the results of 6 clinical assessments were not included in the list of information available to the raters.

Data analysis

For Study 1, inter-rater reliability of the ICF assessment used percentage of agreement and Cohen's kappa with 95% confidence intervals (95% CI) at the category score level, and ICC at the component score level. The strengths of the kappa for expert rater-pairs were compared with those for novice rater-pairs (25). Pearson's correlation coefficients were used to test relationships between component scores on the ICF assessment with those of the 6 common clinical tests.

For Study 2, the data collected was combined with that from Study 1. There were 3 rater group conditions: expert raters without prior knowledge of patient functional capability, novice raters without prior knowledge of patient functional capability, and novice raters with prior knowledge of patient functional capability. Since each participant was assessed twice, results of the first ICF assessment were selected for comparison. ICF assessment ratings for each patient from each rater group condition were compared using multivariate analysis of variance, followed by *post-hoc* comparisons. The level of statistical significance was set at $p \leq 0.05$. All analyses were performed with STATA (version 8.17) and SPSS (version 18) statistical software.

RESULTS

Study 1

For expert rater-pairs, the percentage agreement for the ICF categories was 82.0% (95% CI 79.2–84.8%) and Cohen's kappa coefficient was 0.73 (95% CI 0.68–0.77). For novice rater-pairs, the percentage of agreement was 65.0% (95% CI 61.3–68.8%) and Cohen's kappa coefficient was 0.45 (95% CI 0.38–0.51) (Table III) For expert rater-pairs, 17 out of 18 categories (94.4%) showed Cohen's kappa coefficients above 0.60, compared with only 2 categories (11.1%) for novice rater-pairs. Most novice rater-pair categories (72.2%) showed moderate reliability. At the component score level, the ICC coefficients for expert rater-pairs ranged from 0.76 (95% CI 0.63–0.84) for EF, 0.89 (95% CI 0.83–0.93) for AP, and 0.96 (95% CI 0.93–0.97) for both BF and BS. In contrast, the ICC coefficients for novice rater-pairs ranged from 0.63 (95% CI 0.48–0.75) for EF, 0.85 (95% CI 0.78–0.90) for BS, 0.88 (95% CI 0.82–0.92) for BF, and 0.88 (95% CI 0.81–0.92) for AP.

For expert rater-pairs, the strongest Pearson's correlation coefficients were found between MoCA scores and BF (r = –0.81) and AP (r = –0.69) (Table IV). No significant correlations were found between scores on any of the 6 clinical tests and EF. For novice rater-pairs, stronger correlations were found between the clinical test scores and AP. No significant correlation coefficients were revealed for the EF.

Study 2

One-way multivariate analysis of variance revealed significant differences among the 3 rater groups ($F_{8,426=}5.40$, $p < 0.001$). *Post-hoc* comparisons also revealed significant differences between rater groups except for EF scores (Fig. 2). There were no significant differences between expert rater-pairs and single novice raters with prior knowledge for BF ($p = 0.32$) and BS ($p = 0.48$). However, novice rater-pairs without prior knowledge were found to yield significantly lower mean (standard deviation (SD)) BF ($p = 0.02$; 5.92 (SD 3.73) vs 7.72 (SD 5.10)) and BS ($p < 0.001$; 2.82 (SD 1.49) vs 4.11 (SD 1.69)) component scores than single novice raters with prior knowledge. A similar pattern of differences between novice rater-pairs without prior knowledge and expert rater-pairs were revealed for mean BF ($p = 0.001$; 5.92 (SD 3.73) vs 8.51 (SD 5.28), respectively) and BS ($p < 0.001$; 2.82 (SD 1.49) vs 3.91 (SD 1.66), respectively). On the contrary, for AP component

Table III. *Percentage agreement and Cohen's kappa statistics for stroke assessment*

		Expert rater-pairs ($n = 72$)		Novice rater-pairs ($n = 77$)	
Items	ICF code title	Percentage agreement	Cohen kappa (95% CI)	Percentage agreement	Cohen kappa (95% CI)
b110	Consciousness functions	88.73	0.76 (0.68–0.81)	80.00	0.20 (–0.09–0.32)
b114	Orientation functions	88.89	0.85 (0.79–0.89)	75.00	0.61 (0.55–0.64)
b140	Attention functions	81.43	0.76 (0.74–0.81)	63.64	0.44 (0.36–0.56)
b144	Memory functions	84.51	0.80 (0.74–0.84)	58.44	0.40 (0.35–0.54)
b167	Mental functions of language	80.56	0.74 (0.71–0.77)	59.21	0.40 (0.30–0.48)
b730	Muscle power functions	84.51	0.79 (0.73–0.88)	61.84	0.47 (0.34–0.57)
s110a	Structure of brain: extent	90.28	0.83 (0.69–0.86)	68.83	0.50 (0.45–0.58)
s730a	Structure of upper extremity: extent	87.14	0.84 (0.79–0.88)	71.43	0.59 (0.55–0.69)
d310p	Performance of communicating with - receiving - spoken messages	81.94	0.76 (0.68–0.81)	63.16	0.46 (0.29–0.53)
d330p	Performance of speaking	85.92	0.81 (0.71–0.83)	62.34	0.45 (0.42–0.55)
d450p	Performance of walking	76.39	0.69 (0.55–0.77)	48.05	0.33 (0.29–0.39)
d510p	Performance of washing oneself	76.06	0.68 (0.60–0.72)	64.94	0.54 (0.51–0.72)
d530p	Performance of toileting	74.65	0.66 (0.56–0.73)	57.14	0.46 (0.35– 0.51)
d540p	Performance of dressing	77.46	0.69 (0.55–0.74)	61.04	0.50 (0.40–0.52)
d550p	Performance of eating	72.86	0.61 (0.48–0.68)	57.14	0.43 (0.40–0.55)
e310	Immediate family	77.78	0.50 (0.33–0.58)	67.53	0.41 (0.29–0.48)
e355	Health professionals	93.06	0.67 (0.50–0.86)	68.83	0.18 (–0.24–0.60)
e580	Health services, systems and policies	73.91	0.62 (0.48–0.71)	81.82	0.69 (0.65–0.71)

ICF: International Classification of Functioning, Disability, and Health; 95% CI: 95% confidence interval.

Table IV. *Pearson correlation coefficients between the component scores on the International Classification of Functioning, Disability, and Health Core Set for Stroke (ICF) assessment and scores on the 6 clinical instruments*

Groups and Components	FMA	FAI-C	MBI-C	MoCA	mRS	SF-36
Expert rater-pairs						
Body Function	–0.276*	–0.329**	–0.449***	–0.811***	0.280*	–0.112
Body Structure	–0.322**	–0.044	–0.406***	–0.280*	0.099	–0.247*
Activity and Participation	–0.505***	–0.417***	–0.665***	–0.685***	0.593***	–0.210
Environmental Factors	0.124	0.222	0.113	0.157	–0.144	–0.103
Novice rater-pairs (without prior knowledge)						
Body Function	–0.515***	–0.390**	–0.717***	–0.685***	0.519***	–0.612***
Body Structure	–0.503***	–0.116	–0.450***	–0.225*	0.371**	–0.395***
Activity and Participation	–0.733***	–0.414***	–0.852***	–0.479***	0.710***	–0.750***
Environmental Factors	–0.122	0.040	–0.028	0.151	–0.023	–0.068
Single novice rater (with prior knowledge)						
Body Function	–0.279*	–0.162	–0.455**	–0.768**	0.345**	–0.487**
Body Structure	–0.073	0.39	–0.02	0.024	0.048	–0.059
Activity and Participation	–0.622**	–0.378**	–0.752**	–0.499**	0.633**	–0.671**
Environmental Factors	0.112	0.137	0.007	0.125	–0.102	–0.144

*$p<0.05$, **$p<0.01$, ***$p<0.001$.
FMA: Fugl-Meyer Assessment; FAI; Frenchay Activity Index; MBI: Modified Barthel Index; MoCA: Montreal Cognitive Assessment; mRS: modified Rankin scale; SF-36: Short Form Health Survey.

scores, no significant difference was revealed between novice rater-pairs without prior knowledge and single novice raters with prior knowledge ($p=0.88$). Both single novice raters with prior knowledge ($p=0.006$; mean, 11.54 (SD 6.44)) and novice rater-pairs without prior knowledge ($p=0.004$; mean, 11.38 (SD 7.27)) were found to yield significantly lower scores on AP than expert rater-pairs (mean, 14.66 (SD 6.51)). Notably, the relationships between the 6 common clinical tests and ICF component scores for single novice raters with prior knowledge showed a pattern more similar to that for expert rather than novice rater-pairs without prior knowledge (Table II).

DISCUSSION

The most significant finding of Study 1 is that the expert raters yielded substantially higher inter-rater reliability on the ICF assessment than novice raters. This suggests that clinical experience confers greater consistency in ICF scoring for post-stroke patients. By gaining knowledge about patient functional capacity in Study 2, single novice raters yielded ICF assessment scores similar to expert raters. This compared more favourably with novice rater-pairs without prior knowledge about patient functional capability. Nevertheless, the

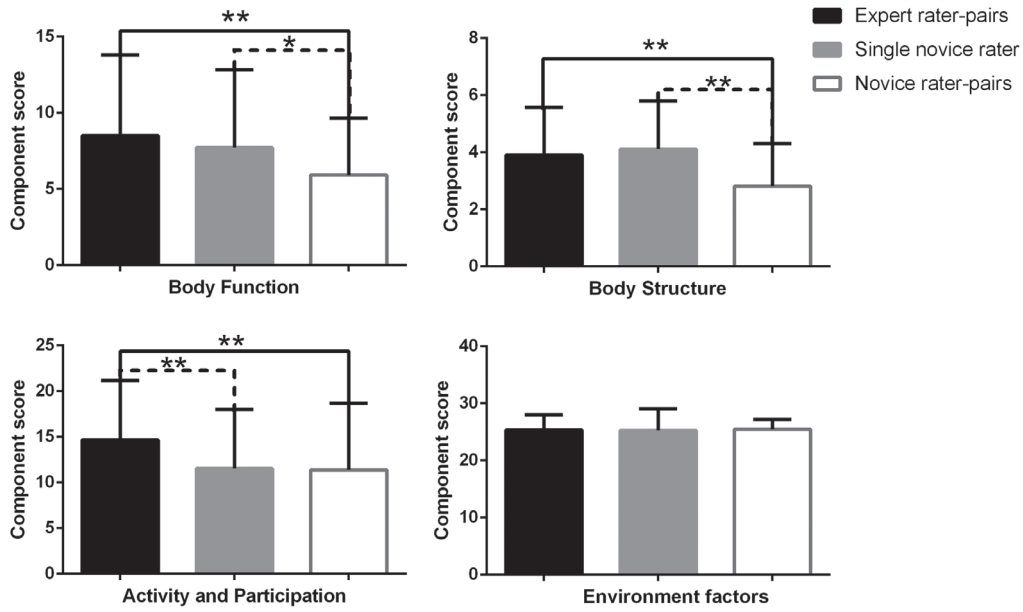

Fig. 2. Comparison of the International Classification of Functioning, Disability, and Health Core Set for Stroke (ICF) component scores (mean and standard deviation (SD)) among the expert rater-pairs, novice rater-pairs (without prior knowledge), and single novice rater (with prior knowledge). *p-value <0.05.

advantages of both clinical experience and prior knowledge of patient functioning seem to most benefit BF and BS scores and least benefit EF scores.

Our results showing raters with more clinical experience having more consistent ICF assessment scores than those with less experience concur with other types of reliabilities (5, 7, 17). Okochi et al. (12) adopted 8 years of experience as a cut-off and revealed higher test-retest reliability with BF and AP scores for more experienced evaluators. The benefit of clinical experience is proposed to account for higher clinical reasoning competence among experienced clinicians (26). One current theory of clinical competence suggests that experiences enable clinicians to accumulate and classify similar clinical problems based on a larger dataset (11), which would strengthen their classification and recognition capability (27). Experienced raters, such as experts in Study 1, could effectively conduct case-based reasoning by identifying the similarities between patients currently under assessment and those from the past (28). This, in turn, would promote consistency in assigning qualifier scores by the raters with more clinical experience for each ICF category. Other studies on expert clinicians have suggested that the expertise developed promotes narrative reasoning and a patient-centred approach (29–31). These skills facilitate effective communication between experienced clinicians and patients, as well as reading of the patient's history in a functional and psychological context (32). In contrast to experienced clinicians, those with less clinical experience were found to be more inclined to use hypothetical-deductive reasoning, which requires conscious effort to extract information while constructing hypotheses based on the patient's problems (6). As a result, less experienced clinicians have less clinician-patient interaction, which limits their understanding of the patient's problems (31). This perhaps explains why novice raters in Study 1 yielded lower inter-rater reliability on ICF assessment than expert raters.

In Study 1, novice raters yielded moderate inter-rater reliability on ICF scores. These results are in agreement with other studies involving novice raters (14) and reflect the dissatisfactory consistency of scores on the ICF assessment by raters who completed basic training on ICF framework and assessment methods. Our results on the expert raters are comparable to those reported by Gan et al. (33). Their study involved 2 experienced raters having more than 5 years of experience in rehabilitation administering the ICF-Children and Youth-based questionnaire to children with autism and produced good to excellent reliability (component ICC = 0.72–0.97). In contrast, coefficients yielded from scores of more experienced clinicians were higher than the moderate inter-rater reliability of the Comprehensive ICF assessment ($\kappa = 0.41$, 95% CI 0.39–0.43) reported by Starrost et al. (13). Their study recruited 2 physical therapists with more than 5 years of experience in neurorehabilitation assessing 30 post-stroke patients. The higher reliability coefficients yielded could be due to participating clinicians rating a greater number of post-stroke patients, as well as use of the Brief version of the ICF. Both factors would enhance the consistency of rating assignments.

Study 1 revealed relatively higher inter-rater reliability coefficients for BF, BS, and AP for both expert and novice raters. EF yielded the lowest inter-rater reliability coefficients, and clinical experience did not appear to have a significant impact. These findings are consistent with those reported for assessment of patients with low back pain (15), stroke survivors (13), and patients with rheumatoid arthritis (14). The relatively lower inter-rater reliability of EF ratings could be attributed to several things. First, the EF categories are not routinely found in medical records (16), such as category e580, which focuses on the role of "health services, systems, and policies" on preventing and treating health-related problems, providing medical rehabilitation, and promoting a healthy lifestyle. Raters would need to make a judgement without referencing existing information on these categories. Without neuropsychological records, rater judgements were based on subjective observations and information gained from patient self-reporting. This is supported by the fact that no significant correlations were revealed between EF and the 6 clinical tests. Secondly, a previous study commented that the EF categories were relatively broad compared with other components and that the qualifier scale (–4 to +4) was more complex (13). The content of the categories pertains to the physical, social, and attitudinal environment (34), such as category e310 "immediate family," which focuses on patient relationships with family members influencing rehabilitation outcome. When a category can be expressed as both a positive and negative factor, the rater may randomly make a judgement (14). Furthermore, the rater sometimes receives inconsistent or vague information from patients and their proxies (13). Finally, the technical terms were found to be unfamiliar to raters (14), such as category e580 "health services, systems, and policies," e310 "immediate family," and e355 "health professionals." The relatively low inter-rater reliability yielded for the EF indicates rooms for improvement in the design of the qualifier and test process. First, the number of response options can be decreased (14, 15, 35), such as from "–4" to "+4" to "–2" to "+2". This would lower the demand of quantifying the judgement to score process on the rater. Secondly, definitions can be given to each response option to further clarify its uniqueness using written descriptions (16) or pictorial presentation (12). Thirdly, the test process can be further standardized in terms of the sequence for collecting or reviewing specific types of information, and the associated decision-making and consensus processes (36). Last, but not least, self-rating by the patient may form one type of information to be considered.

The results obtained from Study 2 are intriguing, in that equipping raters with knowledge regarding the functional capability of post-stroke patients enabled novice raters to produce scores similar to that of expert raters. Such enhancement effects were found to be prominent for BF and BS, but not for AP and EF. Although producing comparable component scores does not necessarily mean an improvement in inter-rater reliability, increases in these scores suggest that knowledge about patient functioning could promote the validity of novice raters using the ICF assessment. The clinicians are encouraged

to be creative in choosing appropriate clinical measures for conducting the ICF assessment (37). However, no studies have yet addressed the issue of incorporating common clinical tests into the ICF rating protocol. Administering common clinical tests, such as those adopted in this study, before conducting the ICF assessment is recommended for raters with less clinical experience. Future studies are needed to investigate the extent to which this can enhance inter-rater reliability among novice raters. Researchers should further explore effective strategies to improve the validity of AP and EF scoring. Lastly, the strong correlations revealed between the 6 clinical tests and *BF and AP,* which is consistent with the findings of a previous study (38). In particular, the MoCA, MBI-C, and SF-36 were found to yield the highest correlation coefficients, suggesting that constructs of these 3 evaluations may well overlap with BF and AP on the ICF. To enhance validity of the ICF assessment, clinicians with less clinical experience may consider getting access to patient MoCA, MBI-C, and SF-36 results before administering the ICF assessment for post-stroke patients. Furthermore, future studies should target development of proxy measures for assisting less experienced clinicians to enhance the validity of EF ratings.

Further improvement in the reliability of the ICF assessment can be achieved by incorporating illustrations as supplementary material in the test. The ICF Illustration Library (http://www.icfillustration.com/top_e.html) can be a useful tool for use by the raters to seek clarification of the definition of specific components and codes. Clear operational definitions enable novice raters to delineate the scope and content of the evaluation, which would improve the reliability of the assessment. Other methods of improving the reliability include formulating standardized decision and consensus process on assigning ratings (36), and reducing the number of qualifiers of the ICF assessment (14).

Study limitations

The present study has several limitations. First, the expert and novice raters recruited are limited to the clinical experiences that they gained within the post-stroke rehabilitation settings in which they practiced. Generalization of the results to other raters practiced in settings and types of patients not similar to those for this study. Secondly, no inter-rater reliability was established for the novice raters recruited in Study 2. Therefore, the results obtained in Study 2 could not be interpreted as enhancing inter-rater reliability. Lastly, the content of the 6 common clinical tests employed in this study did not specifically cover EF. Instead, they primarily measured functional capability of post-stroke patients. This may explain why no significant correlations were established between the common clinical evaluations used and EF in the ICF assessment. Future studies are recommended to address these issues.

Conclusion

The present results indicate that novice raters were significantly less consistent than their expert counterparts on ICF assessment. These novice-to-expert discrepancies were primarily associated with BF, BS, and AP components. A lower competence associated with clinical reasoning and limited opportunity for exposure to diverse clinical cases probably account for the decreased consistency among the novice raters. The validity of scores made by novice raters was improved after administering functional capability measures to post-stroke patients prior to using the ICF assessment. In particular, the enhancement effect was found in BF and BS scores. Strategies on further refining the ICF assessment, such as simplifying the categories of EF and incorporating specific clinical measures in the rating protocol, are recommended.

ACKNOWLEDGEMENTS

This study was funded by the 12th Five-year Plan supporting project of Ministry of Science and Technology of the People's Republic of China (grant number 2013BAI10B01). It was supported by National Rehabilitation Research Center of Traditional Chinese Medicine and Fujian Rehabilitation Tech Co-innovation Center.

The authors declare no conflicts of interest.

REFERENCES

1. Paanalahti M, Lundgren-Nilsson A, Arndt A, Sunnerhagen K. Applying the Comprehensive International Classification of Functioning, Disability and Health Core Sets for stroke framework to stroke survivors living in the community. J Rehabil Med 2013; 45: 331–340.
2. Glassel A, Kirchberger I, Kollerits B, Amann E, Cieza A. Content validity of the Extended ICF Core Set for stroke: an international Delphi survey of physical therapists. Phys Ther 2011; 91: 1211–1222.
3. Geyh S, Cieza A, Schouten J, Dickson H, Frommelt P, Omar Z, et al. ICF Core Sets for stroke. J Rehabil Med 2004; 36: 135–141.
4. Rauch A, Cieza A, Stucki G. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. Eur J Phys Rehabil Med 2008; 44: 329–342.
5. Elgueta-Cancino E, Schabrun S, Danneels L, Hodges P. A clinical test of lumbopelvic control: development and reliability of a clinical test of dissociation of lumbopelvic and thoracolumbar motion. Man Ther 2014; 19: 418–424.
6. Ilgen JS, Bowen JL, Yarris LM, Fu R, Lowe RA, Eva K. Adjusting our lens: can developmental differences in diagnostic reasoning be harnessed to improve health professional and trainee assessment? Acad Emerg Med 2011; 18: S79–S86.
7. Randsborg PH, Sivertsen EA. Classification of distal radius fractures in children: good inter- and intraobserver reliability, which improves with clinical experience. BMC Musculoskelet Disord 2012; 13: 6.
8. Lwu S, Paolucci EO, Hurlbert RJ, Thomas KC. A scoring system for elective triage of referrals: Spine Severity Score. Spine J 2010; 10: 697–703.
9. Eva KW, Norman GR. Heuristics and biases – a biased perspective on clinical reasoning. Med Educ 2005; 39: 870–872.
10. Elstad EA, Lutfey KE, Marceau LD, Campbell SM, von Dem KO, McKinlay JB. What do physicians gain (and lose) with experience? Qualitative results from a cross-national study of diabetes. Soc Sci Med 2010; 70: 1728–1736.
11. Eva KW, Norman GR, Neville AJ, Wood TJ, Brooks LR. Expert-novice differences in memory: a reformulation. Teach Learn Med

2002; 14: 257–263.
12. Okochi J, Utsunomiya S, Takahashi T. Health measurement using the ICF: test-retest reliability study of ICF codes and qualifiers in geriatric care. Health Qual Life Outcomes 2005; 3: 46.
13. Starrost K, Geyh S, Trautwein A, Grunow J, Ceballos-Baumann A, Prosiegel M, et al. Interrater Reliability of the Extended ICF Core Set for Stroke applied by physical therapists. Phys Ther 2008; 88: 841–851.
14. Uhlig T, Lillemo S, Moe RH, Stamm T, Cieza A, Boonen A, et al. Reliability of the ICF Core Set for rheumatoid arthritis. Ann Rheum Dis 2007; 66: 1078–1084.
15. Hilfiker R, Obrist S, Christen G, Lorenz T, Cieza A. The use of the comprehensive International Classification of Functioning, Disability and Health Core Set for low back pain in clinical practice: a reliability study. Physiother Res Int 2009; 14: 147–166.
16. Kohler F, Connolly C, Sakaria A, Stendara K, Buhagiar M, Mojaddidi M. Can the ICF be used as a rehabilitation outcome measure? A study looking at the inter- and intra-rater reliability of ICF categories derived from an ADL assessment tool. J Rehabil Med 2013; 45: 881–887.
17. Soberg HL, Sandvik L, Ostensjo S. Reliability and applicability of the ICF in coding problems, resources and goals of persons with multiple injuries. Disabil Rehabil 2008; 30: 98–106.
18. Schlegel D, Kolb SJ, Luciano JM, Tovar JM, Cucchiara BL, Liebeskind DS, et al. utility of the nih stroke scale as a predictor of hospital disposition [J]. Stroke 2003; 34: 134–137.
19. Ware JJE. SF-36 health survey update. Spine 2000; 25: 3130–3139.
20. Lu J, Li D, Li F, Zhou A, Wang F, Zuo X, et al. Montreal cognitive assessment in detecting cognitive impairment in Chinese elderly individuals: a population-based study. J Geriatr Psychiatry Neurol 2011; 24: 184–190.
21. Leung SO, Chan CC, Shah S. Development of a Chinese version of the Modified Barthel Index- validity and reliability. Clin Rehabil 2007; 21: 912–922.
22. Imam B, Miller WC. Reliability and Validity of Scores of a Chinese Version of the Frenchay Activities Index. Arch Phys Med Rehab 2012; 93: 520–526.
23. Fu TS, Wu C, Lin K, Hsieh C, Liu J, Wang T, et al. Psychometric comparison of the shortened Fugl-Meyer Assessment and the streamlined Wolf Motor Function Test in stroke rehabilitation. Clin Rehabil 2012; 26: 1043–1047.
24. Yuan JL, Bruno A, Li T, Li SJ, Zhang XD, Li HY, et al. Replication and extension of the simplified modified Rankin scale in 150 Chinese stroke patients. Eur Neurol 2012; 67: 206–210.
25. Roberts C. Modelling patterns of agreement for nominal scales. Stat Med 2008; 27: 810–830.
26. Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Med Educ 2007; 41: 1140–1145.
27. Noll E, Key A, Jensen G. Clinical reasoning of an experienced physiotherapist: insight into clinician decision-making regarding low back pain. Physiother Res Int 2001; 6: 40–51.
28. Mamede S, Schmidt HG, Rikers RM, Penaforte JC, Coelho-Filho JM. Influence of perceived difficulty of cases on physicians' diagnostic reasoning. Acad Med 2008; 83: 1210–1216.
29. May S, Greasley A, Reeve S, Withers S. Expert therapists use specific clinical reasoning processes in the assessment and management of patients with shoulder pain: a qualitative study. Aust J Physiother 2008; 54: 261–266.
30. Doody C, McAteer M. Clinical reasoning of expert and novice physiotherapists in an outpatient orthopaedic setting. Physiotherapy 2002; 88: 258–268.
31. May S, Withers S, Reeve S, Greasley A. Limited clinical reasoning skills used by novice physiotherapists when involved in the assessment and management of patients with shoulder problems: a qualitative study. J Man Manip Ther 2010; 18: 84–88.
32. Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. Acad Med 1991; 66: S70–S72.
33. Gan SM, Tung LC, Yeh CY, Wang CH. ICF-CY based assessment tool for children with autism. Disabil Rehabil 2013; 35: 678–685.
34. Bickenbach J, Cieza A, Rauch A, Stucki G. ICF Core Sets: manual for clinical practice. Toronto (Canada): Hogrefe Publishing; 2012.
35. Dernek B, Esmaeilzadeh S, Oral A. The utility of the International Classification of Functioning, Disability and Health checklist for evaluating disability in a community-dwelling geriatric population sample. Int J Rehabil Res 2015; 38: 144–155.
36. Grill E, Gloor-Juzi T, Huber EO, Stucki G. Assessment of functioning in the acute hospital: operationalisation and reliability testing of ICF categories relevant for physical therapists interventions. J Rehabil Med 2011; 43: 162–173.
37. Schulz S. Application and use of the international classification of functioning, disability and health (ICF) in rehabilitation practice and research: an updated literature review. Leonardo internship report. Available from: http://www.marselisborgcentret.dk/fileadmin/filer/Publikationer/PDF_er/Leonardo_internship_report.pdf.
38. Schepers VP, Ketelaar M, van de Port IG, Visser-Meily JM, Lindeman E. Comparing contents of functional outcome measures in stroke rehabilitation using the International Classification of Functioning, Disability and Health. Disabil Rehabil 2007; 29: 221–230.