

Interrater Reliability of the Extended ICF Core Set for Stroke Applied by Physical Therapists

Klaus Starrost, Szilvia Geyh, Anke Trautwein, Jutta Grunow, Andres Ceballos-Baumann, Mario Prosiegel, Gerold Stucki, Alarcos Cieza

Background and Purpose. The World Health Organization's *International Classification of Functioning, Disability and Health* (ICF) is gaining recognition in physical therapy. The Extended ICF Core Set for Stroke is a practical tool that represents a selection of categories from the whole classification and can be used along with the ICF qualifier scale to describe patients' functioning and disability following stroke. The application of the ICF qualifier scale poses the question of interrater reliability. The primary objective of this investigation was to study the agreement between physical therapists' ratings of subjects' functioning and disability with the Extended ICF Core Set for Stroke and with the ICF qualifier scale. Further objectives were to explore the relationships between agreement and rater confidence and between agreement and physical therapists' areas of core competence.

Subjects and Methods. A monocentric, cross-sectional reliability study was conducted. A consecutive sample of 30 subjects after stroke participated. Two physical therapists rated the subjects' functioning in 166 ICF categories.

Results. The interrater agreement of the 2 physical therapists was moderate across all judgments (observed agreement=51%, kappa=.41). Interrater reliability was not related to rater confidence or to the physical therapists' areas of core competence.

Discussion and Conclusion. The present study suggests potential improvements to enhance the implementation of the ICF and the Extended ICF Core Set for Stroke in practice. The results hint at the importance of the operationalization of the ICF categories and the standardization of the rating process, which might be useful in controlling for rater effects and increasing reliability.

K Starrost, PT, MSc, is Head of Physical Therapy, Clinics Schmieder, Allensbach, Germany.

S Geyh, PhD, is Research Scientist, Swiss Paraplegic Research, Nottwil, Switzerland.

A Trautwein, PT, is Physical Therapist, Clinics Schmieder, Gerlingen, Germany.

J Grunow, PT, is Physical Therapist, Neurological Hospital Tristanstrasse, Munich, Germany.

A Ceballos-Baumann, MD, is Professor and Head, Neurological Hospital Tristanstrasse.

M Prosiegel, MD, is Head, Centre for Swallowing Disorders, Speciality Clinic for Physical Medicine and Medical Rehabilitation, Bad Heilbrunn, Germany.

G Stucki, MD, MS, is Professor and Head, Department of Physical Medicine and Rehabilitation, Ludwig-Maximilian University, Munich, Germany; Head, ICF Research Branch of the WHO CC FIC (DIMDI), Institute for Health and Rehabilitation Sciences, Ludwig-Maximilian University; and Head, Swiss Paraplegic Research, Nottwil, Switzerland. Dr Stucki's institutional mailing address is: Department of Physical Medicine and Rehabilitation, University Hospital Munich, Marchioninstrasse 15, 81377 Munich, Germany. Address all correspondence to Dr Stucki at: gerold.stucki@med.uni-muenchen.de.

A Cieza, PhD, is Research Scientist and Group Leader, ICF Research Branch of the WHO CC FIC (DIMDI), Institute for Health and Rehabilitation Sciences, Ludwig-Maximilian University, and Research Scientist and Group Leader, Swiss Paraplegic Research, Nottwil, Switzerland.

[Starrost K, Geyh S, Trautwein A, et al. Interrater reliability of the Extended ICF Core Set for Stroke applied by physical therapists. *Phys Ther.* 2008;88:841-851.]

© 2008 American Physical Therapy Association



Post a Rapid Response or
find The Bottom Line:
www.ptjournal.org

The World Health Organization's (WHO's) *International Classification of Functioning, Disability and Health* (ICF)¹ is gaining recognition in physical therapy and rehabilitation.²⁻⁵ The ICF provides a conceptual basis and a universal common language for understanding and describing patients' health status, reaching beyond mortality, diseases, and medical diagnoses. The use of the ICF promotes a comprehensive, multidisciplinary, and patient-centered perspective in health care. The ICF has been applied in physical therapy and rehabilitation, especially in the field of neurorehabilitation, to facilitate multidisciplinary team communication, to structure the rehabilitation process, for goal setting and assessment, for documentation, and for reporting.⁵⁻¹¹

However, with the practical application of the ICF, important challenges arise. The main challenge is the length of the classification, with more than 1,400 categories. To address this challenge, internationally agreed-on ICF Core Sets for various health conditions have been developed in a scientific evidence-based process.¹² The common standardized procedure for developing ICF Core Sets integrates evidence gathered from preliminary studies with a formal decision-making and expert consensus process. The methodological approaches in the preliminary studies include, for each health condition: (1) systematic literature reviews of outcome measurements used in clinical trials, (2) Delphi exercises capturing experts' views, and (3) collection of empirical data from people undergoing inpatient or outpatient rehabilitation. The results of the preliminary studies are the foundation for the subsequent decision-making and consensus process with a nominal group technique. The resulting ICF Core Sets are practical tools that represent selections of cat-

egories from the whole classification. They comprehensively describe the prototypical spectrum of problems in the functioning of patients with specific health conditions. They are based on the universal language of the ICF but enhance its applicability through their manageable size.

In the context of neurorehabilitation, stroke plays a prominent role. In this field, 3 ICF Core Sets can be applied, namely, the ICF Core Set for Stroke¹³ and the ICF Core Sets for patients with neurological conditions in acute care hospitals¹⁴ and early postacute care rehabilitation facilities.¹⁵ These ICF Core Sets have been combined to create the Extended ICF Core Set for Stroke. It contains all ICF categories that have been selected for any of the 3 ICF Core Sets mentioned above. The Extended ICF Core Set for Stroke contains 166 categories of the ICF, 59 categories of the component "body functions," 11 categories of the component "body structures," and 59 categories of the component "activity and participation." The influence of the component "environmental factors" is described by 37 categories.

A further challenge for the implementation of the ICF is the operationalization of ICF categories. The ICF comprises "qualifiers" to quantify the level of functioning or the severity of the problem in the various ICF categories. The WHO suggests that all categories of the classification be quantified with the same generic scale (Tab. 1).

According to the WHO, broad ranges of percentages are provided for situations in which calibrated assessment instruments or other standards are available to quantify the impairment, activity limitation, participation restriction, or environmental barrier or facilitator.¹ However, cali-

brated assessment instruments based on the ICF category system are scarcely available at present. Using existing instruments along with the ICF would also require the concepts measured by the instruments and their resulting scores to be translated into corresponding ICF categories and qualifiers. Accomplishing such a translation procedure in a scientific way would demand extensive research efforts, which have not been undertaken yet.

Therefore, the application of the ICF and the ICF Core Sets is a challenge to the user and poses the question of reliability and rater agreement when qualifiers are assigned to describe patients' functioning and disability. So far, only a few studies have dealt with the reliability of ICF qualifiers, and the interrater reliability of the qualifiers used with the ICF Core Set for Stroke has not been studied yet. Okochi et al¹⁶ used the ICF Checklist to examine test-retest reliability in geriatric patients and found moderate overall reliability during retesting after 1 week. Reliability varied among categories of the ICF (weighted kappa values = .46 for body functions and .55 for activity and participation). Van Triet et al¹⁷ studied the intertester reliability of a schedule based on the *International Classification of Impairments, Disabilities, and Handicaps* (ICIDH) in patients with musculoskeletal problems. The ICIDH¹⁸ is the predecessor of the ICF. Kappa values ranged from -.06 to 1.00 and were higher in "disability" categories than in "impairment" categories.

The study by van Triet et al,¹⁷ however, is clearly outdated because it is based on the ICIDH. The authors departed greatly from the categories of the classification and from its qualifier scale in creating their assessment schedule. Both studies^{16,17} were conducted with poorly specified mixed samples; thus, the results were not

generalizable to the functioning ratings for patients with stroke. In addition, the investigators in both studies made arbitrary selections of various areas of functioning, not covering the full scope of the ICF, as reflected in the carefully chosen categories of the Extended ICF Core Set for Stroke. In neither of the studies did the investigators consider the full qualifier scale in their analyses. The investigators in both studies applied designs in which different raters completed their recording of patients' functioning at different time points, mixing variation of time points with variation of raters. Thus, the type and the amount of information underlying the ratings might not have been comparable.

Although Okochi et al¹⁶ examined the influence of the experience of raters on retest reliability, rater confidence and core competence might be more proximate variables connected to reliability. Within the context of reliability, rater confidence is an important variable frequently examined in clinical research. For example, in studies dealing with the reliability of imaging techniques^{19,20} and behavioral observations,^{21,22} confidence ratings are often used as independent outcomes to demonstrate diagnostic accuracy. The results of these studies hint at a possible relationship between agreement and confidence. Confidence might serve as an explanatory factor for rater agreement. Thus, with regard to the reliability and the application of the Extended ICF Core Set for Stroke, the association of rater agreement and rater confidence is of interest.

Furthermore, in reliability studies, the experience and training of raters seemed to be highly relevant and were frequently reported.²³⁻²⁹ In these studies, "experience" and "training" referred not only to the handling of the specific rating instrument used, but also to the clinical

Table 1.

Generic *International Classification of Functioning, Disability and Health* Qualifier Scale

Qualifier	Description	Range (%)
Functioning ^a		
0	No problem (none, absent, negligible, . . .)	0-4
1	Mild problem (slight, low, . . .)	5-24
2	Moderate problem (medium, fair, . . .)	25-49
3	Severe problem (high, extreme, . . .)	50-95
4	Complete problem (total, . . .)	96-100
8	Not specified	
9	Not applicable	
Environmental factors ^b		
0	No barrier or facilitator	0-4
1	Mild barrier	5-24
2	Moderate barrier	25-49
3	Severe barrier	50-95
4	Complete barrier	96-100
+1	Mild facilitator	5-24
+2	Moderate facilitator	25-49
+3	Substantial facilitator	50-95
+4	Complete facilitator	96-100
8	Barrier or facilitator: not specified	
9	Not applicable	

^a Range represents impairment, restriction, or limitation.

^b Range represents barrier or facilitator.

experience of the raters within the field and the concepts to be rated and the patient group with the given disease condition. These studies drew an equivocal picture of the relationship between rater competence and interrater reliability and suggested that the different results might have depended on the specific rating instrument examined. Thus, the role of raters' areas of competence should be considered for any new rating tool.

Therefore, the overall objective of this investigation was to study the interrater reliability of physical therapists' ratings of the functioning of study participants with the Extended ICF Core Set for Stroke. The specific aims were: (1) to study the agreement of the 2 physical therapists

in rating participants' functioning with the Extended ICF Core Set for Stroke, (2) to explore the relationship between rater agreement and rater confidence, and (3) to explore rater agreement in relation to physical therapists' areas of core competence.

Method

Study Design, Participants, and Procedures

The study was conducted as a monocentric, cross-sectional reliability study at the Neurological Hospital Tristanstrasse, Munich, Germany, as part of a larger project aimed at the testing and validation of ICF Core Sets. All eligible patients admitted to the hospital on an inpatient or an outpatient basis between June and October 2005 were included in the

Table 2.
Sociodemographic and Stroke-Related Data for Participants (n=30)

Characteristic	No. of Participants ^a	% of Participants ^a
Age (y) ^b	66.6 (11.7)	37.1–87.6
Sex		
Male	19	63
Female	11	37
Etiology of stroke		
I61: Intracerebral hemorrhage	4	13
I62: Other nontraumatic intracranial hemorrhage	1	3
I63: Cerebral infarction	23	77
I64: Stroke, not specified	2	7
Chronicity ^b	6.7 (5.4) mo	1 mo–10 y
<3 mo	17	57
3–12 mo	12	40
>12 mo	1	3
Health care setting		
Inpatient	24	80
Day clinic	5	17
Outpatient	1	3
Rankin Scale disability grade		
1 (not significant)	1	3
2 (slight)	9	30
3 (moderate)	4	10
4 (moderately severe)	13	47
5 (severe)	3	10
Affected side		
Left	19	37
Right	11	63
Occupational status		
On sick leave	5	17
Retired before stroke	25	83

^a Unless otherwise indicated.

^b \bar{X} (SD), range.

study (n=41). Patients were included if their main diagnosis was stroke, if they were at least 18 years of age, if they understood the purposes of and reasons for the study, and if they had signed an informed consent form. Patients with secondary wound healing after surgery were excluded from the study because the aftermath of an invasive surgical procedure might have complicated the description of stroke-

related functioning in these patients. Patients who were not able to participate in an interview were excluded as well, because several categories of the Extended ICF Core Set for Stroke can be judged only according to patient self-report. Altogether, 9 patients were excluded. The main reasons for exclusion were communication problems and impaired vigilance. An additional 2 patients refused to participate because

of exhaustion. Table 2 shows the sociodemographic and stroke-related data for the participants. Two thirds (67%) of the participants were moderately or severely disabled (Rankin Scale grades 3, 4, and 5).³⁰

The participants' functioning was rated by 2 physical therapists with the Extended ICF Core Set for Stroke. The physical therapists had at least 5 years of experience in neuro-rehabilitation, had completed several postgraduate courses in the rehabilitation of motor performance, and were well trained in the use of the ICF. The ICF training consisted of workshops, a pilot phase with discussion of cases before initiation of the study, and the supervision of an expert from the ICF Research Branch, Munich, Germany. The ratings were based on information from the participants (interview and observation), from interviews with proxies (eg, a spouse, partner, or close relative, who also could have the role of a caregiver), from various health care professionals, and from medical records containing the results of standardized examinations. This approach was chosen because the accuracy and comprehensiveness of raters' clinical judgments are expected to increase when various sources of information and various perspectives are taken into account. Furthermore, people with neurological conditions often are not optimally situated to judge the severity of their deficits because of the common phenomenon of unawareness.³¹ Both physical therapists attended the interviews. The role of the interviewer was assigned at random. The 2 physical therapists completed the Extended ICF Core Set for Stroke independently of each other. For each category contained in the Extended ICF Core Set for Stroke, the raters judged their confidence in the ratings on a scale from 0% to 100%. The scale was subdivided into 10% steps.

To avoid bias caused by different levels of information, the participants were known to neither of the interviewers or to both. In addition to the disability and confidence ratings, sociodemographic and stroke-related data were recorded for each participant.

Analyses

To study the agreement of the 2 physical therapists in rating participants' functioning with the Extended ICF Core Set for Stroke, the overall percentage of observed agreement and an overall kappa coefficient³² with the associated bias-corrected 95% bootstrap confidence interval³³ were calculated. These calculations included all judgments across all ICF categories and participants. In addition, for each ICF category, the percentage of observed agreement, the kappa coefficient, and the associated bias-corrected 95% bootstrap confidence interval were calculated.

Because the ICF qualifiers "8: not specified" and "9: not applicable" cannot be integrated into the ordinal scale of the ICF qualifiers 0 to 4 and -4 to 4, respectively, a Cohen kappa statistic for nominal scale response categories was used. The kappa statistic is a measure of agreement that exists beyond the amount of agreement expected by chance alone. Kappa values generally range from 0 to 1, where 1 indicates perfect agreement and 0 indicates no additional agreement beyond that expected by chance alone. A negative kappa value indicates less agreement than that expected by chance alone. The bias-corrected 95% bootstrap confidence interval³³ allows determination of the precision of kappa values without assumptions being made about the homogeneity of the marginal distributions in the data. The bootstrap confidence interval is especially useful with small sample sizes.

It is well known that kappa values depend on the prevalence of the attribute measured and can show biased results with skewed marginal distributions.³⁴⁻³⁶ Consequently, kappa values for various ICF categories cannot be compared with each other properly because their baseline prevalences are unknown and their marginal distributions may be more or less skewed. Thus, in the present study, only information about whether kappa values exceeded chance or not was used for comparisons across ICF categories.³⁷ The percentage of observed rater agreement was preferred as the indicator for the level of agreement. Emphasizing the actual observed agreement can be further justified because the kappa statistic is a chance-corrected measure of agreement, but the definite role of chance in the rating process is not clear.^{34,35}

The levels of confidence of the 2 raters were compared across all judgments. For this purpose, first the confidence ratings of each rater were checked for a normal distribution with the Kolmogorov-Smirnov test. The *t* test for independent samples could be used to check the null hypothesis of no difference in the levels of confidence of the 2 physical therapists when the requirement for a normal distribution of the data was met. Otherwise, the nonparametric Mann-Whitney *U* test was applied.

For exploring the relationship between the level of rater agreement and rater confidence, Pearson correlation coefficients were calculated across all judgments for all ICF categories and participants. The level of agreement was quantified as the percentage of observed agreement in each category. The overall correlation coefficient and the percentage of variance in the level of rater agreement explained by confidence are reported. When the levels of confidence of the 2 raters were different,

the correlation between the level of agreement and confidence and the percentage of variance explained by confidence were reported for the 2 raters separately as well.

For exploring rater agreement in relation to physical therapists' areas of core competence, each ICF category of the Extended ICF Core Set for Stroke was classified according to the results of Glaessel and colleagues (Glaessel A, Kirchberger I, Cieza A, Stucki G; 2007; unpublished research) as a "core competence ICF category" for physical therapists or as a "not-core competence ICF category." ICF categories that have been agreed on by a panel of international experts as areas being treated by physical therapists were classified as core competence ICF categories. For these ICF categories, at least 80% of the experts agreed that patients' problems in the areas considered were being treated by physical therapists.

The levels of agreement in the 2 classes of ICF categories (core competence versus not-core competence) were compared by use of the *t* test for independent samples after the data were examined for the requirement for a normal distribution with the Kolmogorov-Smirnov test. If the requirement for a normal distribution was not met, then the Mann-Whitney *U* test was applied. The variance in the level of rater agreement accounted for by the variable "core competence" was reported with the Spearman correlation coefficient.

Analyses were conducted with SPSS,^{*,38} except for the kappa and bias-corrected 95% bootstrap confi-

* SPSS Inc, 233 S Wacker Dr, Chicago, IL 60606.

Table 3.

Rater Agreement According to *International Classification of Functioning, Disability and Health* Component^a

Component	No. of:		% Observed Agreement (95% CI)	Kappa Value (95% CI)
	Valid Ratings	Missing Ratings		
Overall	4,936	44	51.4 (50.0, 52.8)	.41 (.39, .43)
Functioning	3,828	42	56.3 (54.7, 57.9)	.42 (.39, .44)
Body functions	1,749	21	54.6 (52.3, 56.9)	.37 (.34, .40)
Body structures	322	8	74.2 (69.4, 79.0)	.50 (.41, .57)
Activity and participation	1,757	13	54.7 (52.4, 57.0)	.43 (.40, .47)
Environmental factors	1,108	2	34.4 (31.6, 37.2)	.19 (.15, .24)

^a CI=confidence interval.

confidence interval calculations, which were conducted with SAS.^{†,39}

Results

The percentage of observed agreement, the kappa values and respective 95% confidence intervals, the number of valid pairs of ratings, and the number of missing ratings for all components of the ICF are shown in Table 3.

Figure 1 shows the distribution of the percentage of agreement and the kappa statistics for the ICF components. For the single categories of the Extended ICF Core Set for Stroke, the observed agreement of the physical therapists with regard to the severity ratings for participants' functioning problems ranged from 20% for the category "d930: Religion and spirituality" to 100% for the category "d940: Human rights" (Supplemental Tabs. 1, 2, and 3, available online only at www.ptjournal.org). For environmental

factors, the percentage of observed agreement ranged from 3% for the category "e440: Individual attitudes of personal care providers and personal assistants" to 80% for the category "e135: Products and technology for employment" (Supplemental Tab. 4, available online only at www.ptjournal.org).

According to the kappa coefficients and respective confidence intervals, 87 categories (52%) showed agreement beyond chance and 79 did not. Of the 87 categories that showed agreement, 80 categories were functioning categories and 7 were environmental factors. Thus, 62% of the functioning categories and 19% of the environmental factors were identified as categories that showed agreement.

Figure 2 provides information about the physical therapists' level of confidence in their ratings (for details, see Supplemental Tables 1, 2, 3, and 4). The medians of the confidence ratings were 90% (range=35%-100%) for rater 1 and 100% (range=70%-100%) for rater 2. For raters 1 and 2,

[†] SAS Institute Inc, PO Box 8000, Cary, NC 27513.

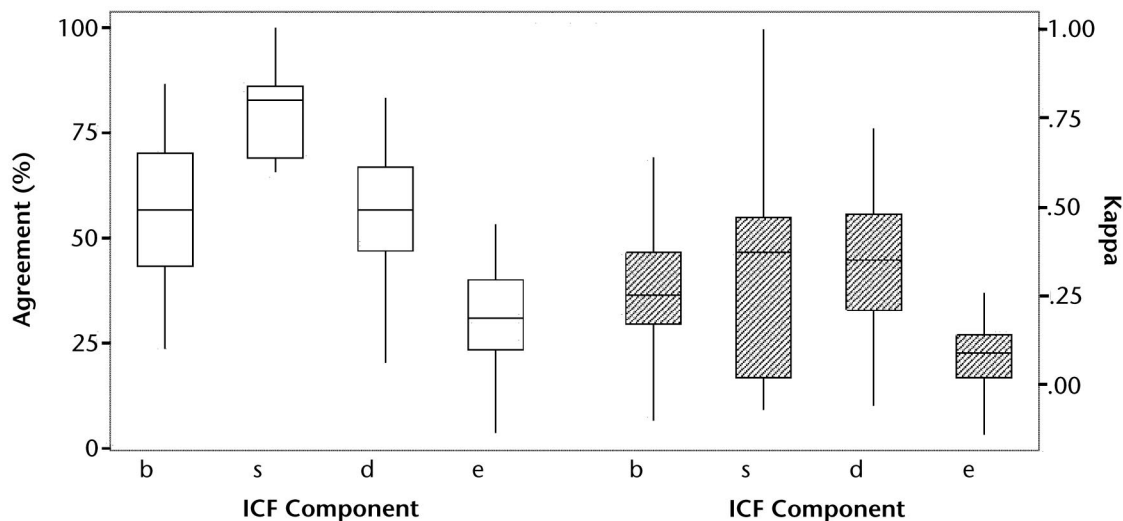


Figure 1.

Distribution of percentage of agreement (blank boxes) and kappa statistics (hatched boxes) by body functions (b), body structures (s), activity and participation (d), and environmental factors (e). ICF=*International Classification of Functioning, Disability and Health*.

the confidence ratings across all participants and all 166 categories of the Extended ICF Core Set for Stroke showed a normal distribution. The *t* test for independent samples confirmed a statistically significant difference between the 2 raters ($P < .00$) with regard to their levels of confidence. Thus, the relationship between confidence and agreement was examined separately for each rater. Pearson correlation coefficients were significantly different from 0 ($P < .01$) when calculated overall as well as for the 2 raters separately; however, the strength of the correlations was negligible ($r = .08$ overall, $r = .1$ for rater 1, and $r = .07$ for rater 2).

The Extended ICF Core Set for Stroke contains 56 categories identified as physical therapists' core competence categories and 110 not-core competence categories. Rater agreement was not found to differ between core competence and not-core competence categories by the Mann-Whitney test ($U = 2,982$; $P = .74$). The Mann-Whitney test was used because the percentage of observed agreement did not show a normal distribution (Kolmogorov-Smirnov test: $z = .97$, $P = .31$). The Spearman correlation coefficient was not significant ($\rho = .26$, $P = .74$).

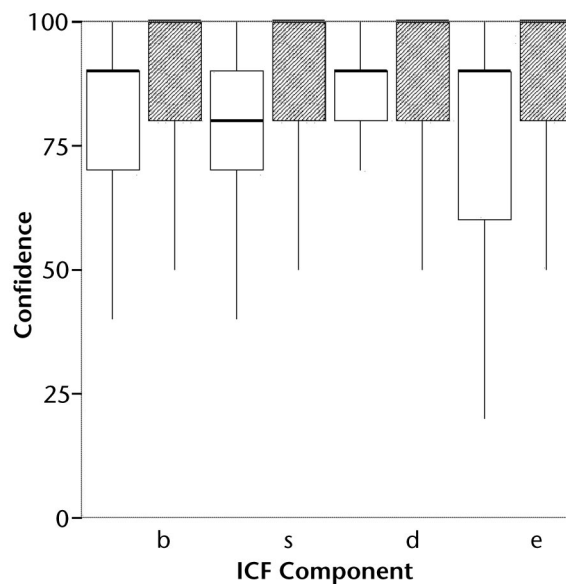
Discussion

The present study is the first to examine the interrater reliability of the ICF qualifiers used with the Extended ICF Core Set for Stroke. The present study demonstrated that the interrater reliability of the judgments was moderate overall. It showed that interrater reliability was not related to confidence or to the physical therapists' areas of core competence.

The overall result for interrater agreement was a kappa value of .41. This value can be interpreted as "moderate" according to Altman⁴⁰ but as "poor" according to El Emam.⁴¹ However, interpreting kappa values ac-

Figure 2.

Distribution of confidence ratings by body functions (b), body structures (s), activity and participation (d), and environmental factors (e). Blank boxes represent rater 1; hatched boxes represent rater 2. ICF = *International Classification of Functioning, Disability and Health*.



cording to recommendations in the literature is only an initial step. The acceptability of a certain reliability value depends on the intended uses and types of decisions based on the instrument. Instruments used for decisions about an individual should be more reliable than instruments used for decisions about a group or for research purposes. The interrater reliability of the qualifier scale used with the Extended ICF Core Set for Stroke should be enhanced in the future. However, for research purposes involving large samples, it might be acceptable. Furthermore, the usefulness of the Extended ICF Core Set for Stroke in current physical therapy practice essentially is based on its capability to integrate the results of different examinations, measurements, different professionals' clinical observations, and patients' self-reports into one and the same common framework. Decisions in current ICF-based practice rely on such an integrative per-

spective and use the complete ICF functioning profile of a patient to select and prioritize intervention goals, to decide on the course of interventions, and to manage the rehabilitation process.⁴² The 2 physical therapists in the present study gathered and analyzed a large amount of data in rating the 166 categories of the ICF Core Set for each participant. They did so largely without being able to rely on the results of standardized measurements and within a clinical environment that was not yet familiar with ICF-based practice. Under these conditions, the overall kappa value of .41 may be regarded as a promising starting point.

In a recent study, Uhlig et al⁴³ examined the interrater reliability of the ICF Core Set for Rheumatoid Arthritis in a sample of 25 people seen on an outpatient basis. The mean observed agreement across the 95 ICF categories was 47%, a value slightly lower than that in the present study

of the Extended ICF Core Set for Stroke (52%). However, the patterns of agreement across the components of the ICF were very similar in both studies (higher for functioning and lower for environmental factors). Uhlig et al⁴³ did not report an overall kappa value, but they indicated that 64% of the ICF Core Set for Rheumatoid Arthritis showed agreement beyond chance. Only 52% of the categories of the Extended ICF Core Set for Stroke demonstrated a significant kappa value. However, Uhlig et al used weighted kappa calculations instead of the nominal kappa calculations used in the present study—a more conservative approach regarding the characteristics of the ICF qualifier scale.

A recent methodological study by Grill et al⁴⁴ exemplified various approaches to the assessment of rater agreement, including observed agreement rates and kappa calculations. They used as an example the ICF category “d430: lifting and carrying objects” in a survey with a repeated-measurement design of a convenience sample of 25 patients in an acute-care hospital setting. For this single ICF category, Grill et al found results similar to those of the present study. The observed agreement rate was 52%; that in the present study was 57%. The weighted kappa value was .36; the nominal kappa value in the present study was .46 (95% confidence interval = .26–.69).

The observed rater agreement and the results of the chance-corrected kappa analyses suggested that interrater reliability was notably higher for the components of functioning than for the environmental factors. Thus, the level of rater agreement might be connected to content-related features in the various ICF categories. Categories with high agreement were likely to be narrow categories with clearly defined concepts and a less complex rating pro-

cess. The judgments in these categories could easily be obtained from observation (b215: function of structures adjoining the eye) and common knowledge (d940: human rights), taken from a medical chart (s120: spinal cord and related structures), or obtained by asking the participant (d850: remunerative employment).

Categories with low agreement were mainly broad and complex categories (eg, e440: individual attitudes of personal care providers and personal assistants) frequently encountered in environmental factors. However, for body functions and activity and participation as well, several categories subsumed broad contents and therefore might be difficult to agree on (eg, d455: moving around; and b240: sensations associated with hearing and vestibular function). For broad ICF categories, raters frequently face inconsistent or vague information because of conflicting statements from patients, proxies, and medical charts⁴⁵ or unawareness of the patient.³¹ For example, the raters often observed that participants were distracted during the interview or did not completely understand the questions. However, when asked specifically, the participants stated no subjective impairment in “listening” (d115) and “focusing attention” (d160). In the absence of neuropsychological records, the judgments were based on the personal impressions of the physical therapists.

Low agreement in some categories (eg, d640: doing housework; and e425: individual attitudes of acquaintances, peers, colleagues, neighbors, and community members) might have resulted from the fact that most inpatient participants were not yet confronted with their home environment and usual tasks. Thus, the ratings of their difficulties in these categories were based on inferences

rather than direct observation and factual information.

However, the difference in agreement between the components of functioning and the environmental factors might also have been influenced by the different numbers of response options for the ICF qualifier scale. For the environmental factors, the qualifier scale contains more steps (–4 to +4), the variability is higher, and it is more difficult to achieve exact agreement between 2 raters.

The ratings of the participants' functioning following stroke were based on information gathered by interviewing, observing, and reading medical records and were rarely based on measurement results. For many areas encompassed by the ICF Core Set, no measures exist or measures are not routinely used in the clinical setting. Developing standardized measurement methods and operationalizations, such as patient self-rating questionnaires, for these areas is a future step that needs to be based on, guided, and justified by the evaluation of the ICF Core Set and the qualifier scale as they are now, a goal that has been undertaken in the present study.

It is important to stress that the ICF Core Set is not a measurement instrument but rather a classification-based tool, with the qualifiers assigned being global severity judgments but not measures. Thus, common rater effects (such as the halo effect and severity-leniency errors), the raters' knowledge, attitudes, and beliefs, the interaction between the rater and the patient, and demographic characteristics may influence the rating process in a highly individual way.^{46–48} Within the ICF Core Set, these effects are not controlled for and most probably contribute to nonagreement in several categories. Severity ratings may have differed be-

tween the physical therapists because of their personal characteristics (if one would have been stricter while the other would have been more serene about the participants' state in general). Confidence ratings may have differed because one of the physical therapists was—as the leader of the present study—familiar and connected with ICF-related issues in a more in-depth way than the other, despite the ICF training completed before the study. The present study involved a simple design and a straightforward method of analysis of rater agreement that did not take into account the factors that influence the rating process in a systematic way. However, in future studies, systematic variations in the rating process could be examined and accounted for by use of a more comprehensive design and modeling methods or Rasch analysis techniques.⁴⁶

The present study revealed no relationship between agreement and confidence. From a statistical perspective, this result was attributable to the low variability of the confidence ratings of the 2 physical therapists. Two possible explanations might account for this low variability. Some studies on the cognitive processes of decision and judgment provide evidence that confidence in judgment can be understood in terms of stable interindividual differences or as a personality trait.^{49,50} Thus, in light of this evidence, it is not surprising that the confidence judgments in the present study showed only low variations across different rating situations, that is, different participants and categories, but remained relatively constant for a given individual or rater.

In addition, the ICF qualifier scale includes the option of choosing the qualifier “8: not specified,” which can be applied when the available information is not sufficient to make a judgment.

Thus, low confidence levels can be avoided by raters by use of the qualifier 8. Doing so reduces the variability of the confidence ratings and complicates any relationship between confidence and agreement. Therefore, future studies on the reliability of the ICF qualifier scale, including confidence ratings, should omit the qualifier “8: not specified.” However, in the present study, the effect of the use of the qualifier 8 on the confidence-agreement relationship might have only minor relevance, because only 5% of the judgments of rater 1 and 7% of those of rater 2 were assigned the qualifier “8: not specified” (data not shown).

The results of the present study further showed that rater agreement was independent of the raters' areas of core competence as physical therapists. The raters achieved a high level of agreement in several categories that were not among the physical therapists' areas of core competence (eg, d330: speaking; d710: basic interpersonal interactions; d850: remunerative employment; and d845: acquiring, keeping, and terminating a job). ICF categories defined as physical therapists' areas of core competence are aspects of functioning that are treated by physical therapists. However, physical therapists are trained to have comprehensive knowledge of stroke and are experienced in observing and detecting the full scope of patients' problems, for example, when taking their history. This means that physical therapists are well able to identify patients' problems, for example, in the category “d330: speaking,” even though these problems are typically treated by speech and language therapists. That is, although the ICF categories that cover the core competencies describe the areas in which physical therapists are usually trained, their experience, skills, and knowledge as health care professionals working in an interdisciplinary team surpass

the specified focus of these ICF categories.

At this point, the potential of the ICF Core Set as a basis for multidisciplinary communication and cooperation in rehabilitation practice and management arises. Still, the question of agreement and interrater reliability among health care professionals of different specializations remains open. Future studies involving various health care professions should be conducted to clarify this question.

However, the results also revealed several categories that addressed physical therapists' areas of core competence but that were rated differently by the 2 physical therapists (eg, b176: mental functions of sequencing complex movements; b755: involuntary movement reaction functions; and b260: proprioceptive functions). These results indicated that for the participants in the present study, the information available on problems was not based on measurements and in-depth examinations but relied on global impressions from the interviews.

The present study has several limitations. Because of the monocentric design, the small sample size, and the small number of participating raters, the generalizability of the results is limited, and the results need to be interpreted with caution. The results suggest next steps for future investigations. In addition, the present study addressed interrater reliability in terms of agreement between 2 raters. It did not address the quality or “truth” of the ratings. Currently, no gold standard exists against which a description of patients' problems across the categories of the ICF Core Set can be compared.

The present study suggests potential improvements to enhance the implementation of the Extended ICF Core

Set for Stroke in practice. To enhance interrater reliability, the training of health care professionals with regard to the ICF should be further developed and standardized. Implementing the ICF in rehabilitation practice at an institutional level may enhance the availability and accessibility of information about all aspects of functioning in individual patients, which in turn may enhance the reliability of ICF-based ratings.⁸ In addition, the metric characteristics of the ICF qualifier scale should be taken into account. In particular, the distance between the steps of the scale may be too narrow and therefore may lead to disagreements. Thus, examining and restructuring the rating scale, for example, with Rasch analyses, may also enhance its interrater reliability.⁴³ However, the results of the present study mainly hint at the importance of operationalization of the categories and standardization of the rating process to control for rater effects and to increase reliability.

In the future, 2 paths toward the operationalization of ICF categories can be followed, namely, the development of ICF-based measures and the development of detailed ICF manuals. Efforts in the latter direction are already being made, for example, by the American Psychological Association⁵¹ and the Australian Institute of Health and Welfare.⁵² Operationalizing the categories of the Extended ICF Core Set for Stroke could be an important next step to ease and to facilitate the application of the ICF in clinical practice and to use its full potential. Physical therapists can make valuable contributions to these developments to enhance professional, scientifically founded, multidisciplinary practice for the benefit of patients.

Mr Starrost, Dr Geyh, Dr Stucki, and Dr Cieza provided concept/idea/research design. Mr

Starrost, Dr Geyh, and Dr Cieza provided writing and project management. Mr Starrost, Ms Trautwein, and Ms Grunow provided data collection. Mr Starrost and Dr Geyh provided data analysis. Dr Ceballos-Baumann and Dr Prosiegel provided facilities/equipment and subjects. Dr Stucki provided institutional liaisons. Dr Stucki and Dr Cieza provided consultation (including review of manuscript before submission).

This study was approved by the Ethics Committee of the University of Munich.

This article was submitted July 27, 2007, and was accepted March 24, 2008.

DOI: 10.2522/ptj.20070211

References

- 1 *International Classification of Functioning, Disability and Health: ICF*. Geneva, Switzerland: World Health Organization; 2001.
- 2 Bartlett DJ, Macnab J, Macarthur C, et al. Advancing rehabilitation research: an interactionist perspective to guide question and design. *Disabil Rehabil*. 2006;28:1169-1176.
- 3 Jette AM. Toward a common language for function, disability, and health. *Phys Ther*. 2006;86:726-734.
- 4 Levack W. The *International Classification of Functioning, Disability and Health* (ICF): application to physiotherapy. *New Zealand Journal of Physiotherapy*. 2004;32:1-2.
- 5 Tempest S, McIntyre A. Using the ICF to clarify team roles and demonstrate clinical reasoning in stroke rehabilitation. *Disabil Rehabil*. 2006;28:663-667.
- 6 Perry A, Morris M, Unsworth C, et al. Therapy outcome measures for allied health practitioners in Australia: the AusTOMs. *Int J Qual Health Care*. 2004;16:285-291.
- 7 Steiner WA, Ryser L, Huber E, et al. Use of the ICF model as a clinical problem-solving tool in physical therapy and rehabilitation medicine. *Phys Ther*. 2002;82:1098-1107.
- 8 Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. *Disabil Rehabil*. 2002;24:932-938.
- 9 Stucki G, Sangha O. Principles of rehabilitation. In: Klippel JH, Dieppe PA, eds. *Rheumatology*. 2nd ed. London, United Kingdom: Mosby; 1997:11.11-11.14.
- 10 Bilbao A, Kennedy C, Chatterji S, et al. The ICF: applications of the WHO model of functioning, disability and health to brain injury rehabilitation. *NeuroRehabilitation*. 2003;18:239-250.
- 11 Rentsch HP, Bucher P, Dommen Nyffeler I, et al. The implementation of the *International Classification of Functioning, Disability and Health* (ICF) in daily practice of neurorehabilitation: an interdisciplinary project at the Kantonsspital of Lucerne, Switzerland. *Disabil Rehabil*. 2003;25:411-421.

- 12 Cieza A, Ewert T, Ustun TB, et al. Development of ICF Core Sets for patients with chronic conditions. *J Rehabil Med*. 2004;44(suppl):9-11.
- 13 Geyh S, Cieza A, Schouten J, et al. ICF Core Set for stroke. *J Rehabil Med*. 2004;44(suppl):135-141.
- 14 Ewert T, Grill E, Bartholomeyczik S, et al. ICF Core Set for patients with neurological conditions in the acute hospital. *Disabil Rehabil*. 2005;27:367-373.
- 15 Stier-Jarmer M, Grill E, Ewert T, et al. ICF Core Set for patients with neurological conditions in early post-acute rehabilitation facilities. *Disabil Rehabil*. 2005;27:389-395.
- 16 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.
- 17 van Triet EF, Dekker J, Kerrens JJ, Curfs EC. Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy. *Int Disabil Stud*. 1990;12:61-65.
- 18 *International Classification of Impairments, Disabilities, and Handicaps*. Geneva, Switzerland: World Health Organization; 1980.
- 19 Roberts CC, McDaniel NT, Krupinski EA, Eryl WK. Oblique reformation in cervical spine computed tomography: a new look at an old friend. *Spine*. 2003;28:167-170.
- 20 Stokking R, van Isselt JW, van Rijk PP, et al. Integrated visualization of functional and anatomic brain data: a validation study. *J Nucl Med*. 1999;40:311-316.
- 21 Denis D, Lortie M, Bruxelles M. Impact of observers' experience and training on reliability of observations for a manual handling task. *Ergonomics*. 2002;45:441-454.
- 22 Weiner D, Peterson B, Keefe F. Chronic pain-associated behaviors in the nursing home: resident versus caregiver perceptions. *Pain*. 1999;80:577-588.
- 23 Berry J, Kramer K, Binkley J, et al. Error estimates in novice and expert raters for the KT-1000 arthrometer. *J Orthop Sports Phys Ther*. 1999;29:49-55.
- 24 Brunnekreef JJ, van Uden CJ, van Moorsel S, Kooloos JG. Reliability of videotaped observational gait analysis in patients with orthopedic impairments. *BMC Musculoskelet Disord*. 2005;6:17.
- 25 Carr EK, Kenney FD, Wilson-Barnett J, Newham DJ. Inter-rater reliability of postural observation after stroke. *Clin Rehabil*. 1999;13:229-242.
- 26 Ellis B, Bruton A. A study to compare the reliability of composite finger flexion with goniometry for measurement of range of motion in the hand. *Clin Rehabil*. 2002;16:562-570.
- 27 Hermansson LM, Bodin L, Eliasson AC. Intra- and inter-rater reliability of the assessment of capacity for myoelectric control. *J Rehabil Med*. 2006;38:118-123.
- 28 Lennon S, Johnson L. The modified Rivermead mobility index: validity and reliability. *Disabil Rehabil*. 2000;22:833-839.

- 29 Pomeroy VM, Chambers SH, Giakas G, Bland M. Reliability of measurement of tempo-spatial parameters of gait after stroke using GaitMat II. *Clin Rehabil*. 2004;18:222-227.
- 30 Rankin J. Cerebral vascular accidents in patients over the age of 60, II: prognosis. *Scott Med J*. 1957;2:200-215.
- 31 McGlynn SM, Schacter DL. Unawareness of deficits in neuropsychological syndromes. *J Clin Exp Neuropsychol*. 1989;11:143-205.
- 32 Cohen J. A coefficient for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
- 33 Vierkant R. A SAS macro for calculating bootstrapped confidence intervals about a kappa coefficient. *SAS Users Group International Online Proceedings*. Available at: <http://www2.sas.com/proceedings/sugi22/STATS/PAPER295.PDF>. Accessed July 23, 2004.
- 34 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992;304:1491-1494.
- 35 Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol*. 1990;43:543-549.
- 36 Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol*. 1988;41:949-958.
- 37 Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clin Exp Pharmacol Physiol*. 2002;29:527-536.
- 38 *SPSS for Windows* [computer program]. Version 12.0.1. Chicago, Ill: SPSS Inc; 1989-2003.
- 39 *The SAS System for Windows* [computer program]. Version 8.2. Cary, NC: SAS Institute Inc; 2001.
- 40 Altman DG. *Practical Statistics for Medical Research*. London, United Kingdom: Chapman & Hall; 1999.
- 41 El Emam K. Benchmarking kappa: interrater agreement in software process assessments. *Empirical Software Engineering*. 1999;4:113-133.
- 42 Swiss Paraplegic Research. Implementation of the *International Classification of Functioning, Disability and Health* (ICF) in rehabilitation practice. Web site of the Swiss Paraplegic Research Case Study Project. 2007. Available at: <http://www.icf-casestudies.org>. Accessed April 11, 2008.
- 43 Uhlig T, Lillemo S, Moe RH, et al. Reliability of the ICF Core Set for rheumatoid arthritis. *Ann Rheum Dis*. 2007;66:1078-1084.
- 44 Grill E, Mansmann U, Cieza A, Stucki G. Assessing observer agreement when describing and classifying functioning with the International Classification of Functioning, Disability and Health. *J Rehabil Med*. 2007;39:71-76.
- 45 Sager MA, Dunham NC, Schwantes A, et al. Measurement of activities of daily living in hospitalized elderly: a comparison of self-report and performance-based methods. *J Am Geriatr Soc*. 1992;40:457-462.
- 46 Myford CM, Wolfe EW. Detecting and measuring rater effects using many-facet Rasch measurement: part I. *J Appl Meas*. 2003;4:386-422.
- 47 Elstein AS, Schwartz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*. 2002;324:729-732.
- 48 Landy FJ, Farr JL. Performance rating. *Psychol Bull*. 1980;87:72-107.
- 49 Blais AR, Thompson MM, Baranski JV. Individual differences in decision processing and confidence judgments in comparative judgment tasks: the role of cognitive styles—personality and individual differences. *Pers Individ Dif*. 2005;38:1701-1713.
- 50 Pallier G, Wilkinson R, Danthiir V, et al. The role of individual differences in the accuracy of confidence judgments. *J Gen Psychol*. 2002;129:257-299.
- 51 Reed GM, Lux JB, Bufka LF, et al. Operationalizing the International Classification of Functioning, Disability and Health in clinical settings. *Rehabil Psychol*. 2005;50:122-131.
- 52 *ICF Australian User Guide*. Version 1.0. Canberra, New South Wales, Australia: Australian Institute of Health and Welfare; 2003.

Invited Commentary

Alan M Jette

In 2001, the World Health Organization (WHO) released the *International Classification of Functioning, Disability and Health* (ICF), which provided a comprehensive framework of health states that encompassed a biological, personal, and social perspective.¹ Since the May 2001 World Health Assembly endorsement as a member of the WHO family of international classifications, all member states were asked to implement the ICF in their respective health sectors.² Toward that goal, several international efforts have been launched to develop classification approaches for the assessment and reporting of ICF functioning and health concepts in clinical studies or clinical encounters.³ The Extended ICF Core Set for Stroke represents

one such effort, and the excellent article by Starrost et al⁴ reports on the agreement between 2 physical therapists when rating patients' functioning using the Extended ICF Core Set for Stroke.

As I have written elsewhere,⁵ I believe the development of the ICF framework is an important milestone that can contribute importantly to the field of physical therapy and rehabilitation. One of the most exciting aspects of the ICF framework is its potential to provide a universal, standardized disablement framework that, if widely adopted, will promote a common, international language that will facilitate communication and scholarly discourse across disciplines as well as across

national boundaries, stimulate interdisciplinary research, improve clinical care, and ultimately better inform health policy. Recently, the ICF was endorsed by an Institute of Medicine (IOM) report, *The Future of Disability in America*.⁶ Nonetheless, as pointed out in the IOM report, challenges around the operationalization of the ICF's core concepts need to be resolved if the ICF framework is to become an international standard. In this regard, Starrost et al are to be commended for publishing the first study of the interrater reliability of applying the ICF classification qualifiers in the Extended ICF Core Set for Stroke.

The Extended ICF Core Set for Stroke uses the ICF classification ap-