Scand J Rehab Med 27: 253–256, 1995

LETTER TO THE EDITOR

INTERRATER RELIABILITY OF THE 7-LEVEL FUNCTIONAL INDEPENDENCE MEASURE (FIM)

Sir,

The interrater reliability study by Hamilton et al. (1) of the 7-level functional independence measure contains possible sources of bias in its clinical assessments, inadequate statistical descriptions of its results, and lacks evidence to support its conclusions. The study abstract contains a description of results that does not reflect the actual methodology.

Hamilton et al. reported on the interrater reliability in the clinical setting, asking two clinicians in each participating facility "to make their patient FIM assessments on the same day during the patient's first rehabilitation admission and not to discuss their findings with each other". They then state that "most of the items were assessed by the disciplines usually assigned to evaluating a given functional area. For example, occupational therapists were likely to assess eating.... physical therapists, ambulation and stairs; nursing, bowel and bladder". Therefore each patient has not necessarily been rated by two clinicians, but perhaps by two groups of clinicians. If the assessment was carried out by groups, mechanisms must have been necessary to ensure that consultation did not occur, but these are not described, apart from the instruction that the clinicians not discuss their findings with each other.

The method of assessment of the patients is not described. Ideally, an inter-rater reliability study should have the two assessors assessing the one performance by the subject at the same time. It is not known if the assessments were performed on one or two different occasions on the same day, raising the possibility of test-retest bias, especially important when considering the fluctuating daily functional performance in some patients with neurological conditions.

Because of the lack of information about the setting of assessments, the number of participants in the study at the various institutions involved, the times at which the assessments were made, and the information given to the patients undergoing assessment, we may assume that uniform conditions did not prevail across the study.

The study consists of 89 separate studies by two, or perhaps more assessors, of at least 10 different patients, with no two studies involving the same patients. This is a less than adequate methodology for a formal study of reliability, as the confidence interval for intraclass correlation coefficients based on 10 patients would be unacceptably wide. Eighty-nine separate studies of 10 different patients each is not the equivalent of one study of 890 patients (or 1018 patients).

While Hamilton et al. [quoting Bartko & Carpenter (2)] admit that intraclass correlation coefficients are not recommended for assessing reliability at the nominal or ordinal level, they present us with a group of averaged intraclass correlation coefficients for FIM total, domain and subscale scores. Ordinal data from FIM when summed do not become interval data, and these intraclass coefficients are, in our opinion, lacking in meaning, quite apart from lacking a confidence interval.

No range or variance of Kappa coefficients was presented. No test of significance of Kappa could be presented, presumably because of the small number of patients in each of the 89 studies. It is possible that the range of Kappa coefficients in this study included values less than 0.4. The fact that one of the criteria for inclusion in a "criterion facility", was "15 of 18 items with a Kappa greater than 0.45" indicates the strong likelihood of this situation. If any of these low values were taken as evidence of reliability, the FIM would not be an acceptable measure. There is no reason, of course, given the nature of this study, for a high Kappa to be more acceptable than a low Kappa, as neither may have been greater than a value expected by chance.

No information is given as to when the selection techniques used to establish the criterion facilities were decided. If the criteria were decided *post hoc*, it seems self-evident that those meeting the criteria had high inter-rater reliability scores, as setting lower limits for cut-off scores for inclusion in aggregated statistics results in a higher average for those

included. This does not change the fact that the overall agreement is not as good as the authors would like it to be. When one looks at the average Kappa coefficients of all respondents, they lie between 0.53 and 0.66, the majority being 0.59 or less. It would be interesting to know what processes were used which resulted in facilities changing from not meeting the criteria for inclusion to meeting those criteria at a subsequent assessment, and what incentives there were in order for individuals as well as institutions to meet those criteria, noting that those participating in the study subscribed to the Centre for Functional Assessment Research.

No information is given relating the amount or type of training in FIM assessments to either Kappa or ICC scores. Some Units did not become criterion facilities at their first attempt, but this may have been due to the vagaries of chance resulting from the unreliability of FIM rather than a failure of training.

REFERENCES

1. Hamilton, B.B., Laughlin, J.A., Fiedler, R.C. & Granger, C.V.: Interrater reliability of the 7-level functional independence measure. Scand J Rehabil Med *26*:115–119, 1994.
2. Bartko, J.J. & Carpenter, W.T.: On the methods and theory of reliability. J Nerv Mentl Dis *163*: 307–317, 1976.

Hugh G. Dickson, FACRM, FAFRM, and Friedbert Köhler FACRM, FAFRM
Department of Rehabilitation and Geriatrics, Liverpool Hospital, PO Box 103, Liverpool, NSW, 2170, Australia.

REPLY TO THE LETTER BY DICKSON & KÖHLER

Ikegami (1) recently identified four reasons to use functional assessment in health care. The first is the use of functional assessment as an indicator of outcome. A second use of functional assessment, is in planning care. The third, and closely related application of functional assessment is the development of a system to monitor payment and accountability of resources. The final use identified by Ikegami is the application of functional assessment for prediction in programs aimed at prevention and the reduction of risk factors. To accomplish each of the above purposes requires a reliable and valid measure of function. In a recent study published in the *Journal*, Hamilton and colleagues (2) reported on the reliability of the *Functional Independence Measure* (FIM)^{SM}. Dickson & Köhler (3) have suggested the methodology of this study is flawed and, as a result, the reliability of the FIM is suspect. The major criticisms of Dickson & Köhler (3) are discussed below.

Inadequate Procedures. Dickson & Köhler (3) are concerned that the methodology did not conform to the typical model used to assess interrater reliability. They state, "*Ideally, an interrater reliability study should have the two assessors assessing the one performance by the subject at the same time.*" This is the typical model for conducting an interrater reliability study. There is no empirical evidence, however, that this is the only valid model—it is simply the most common. Past reliability studies have been based on the use of the Pearson product moment correlation (r) which is a bivariate statistic and can be used only with two raters (variables) (4). The introduction of generalizability theory and the use of the intraclass correlation coefficient have dramatically improved the ability to conduct reliability investigations and expanded the available models (5–6). The limitations of reliability investigations using only two raters are now well documented and interested readers should consult one of the following references for more detailed information (5–7).

In the Hamilton et al. (2) investigation the FIM ratings were assessed by disciplines usually assigned to evaluate a given functional area. This meant that items assessing locomotion were usually recorded by physical therapists, while eating and dressing items were usually rated by occupational therapists, etc. The reason for using this model was simple—it reflects how the FIM is used in natural (real world) clinical environments. Dickson & Köhler (3) suggest that using different settings and raters, and not specifying the time of administration or the instructions to the patients introduced error and unreliability. This error variance is "*especially important when considering the fluctuating performance in some patients with neurological conditions.*" We agree with Dickson & Köhler (3) that "*uniform conditions did not persist across the study.*" This lack of uniformity was purposeful and reflects the real world application of the FIM. The important question is: What effect does this variability have on the reliability results? The introduction of error variance should produce a decrease in the consistency of ratings across settings, therapists, times, and impairment groups. The fact that the

reliability values for the FIM were uniformly high strongly suggests that the FIM produces reliable information when used in the normal clinical environment. To argue that the methodological differences noted above would result in increased reliability makes the statistically absurd assumption that the error variance across settings, therapists, times, and impairment groups, was systematic (in a positive way) rather than random.

Unit of Analysis. Dickson & Köhler (3) indicate that the study is really "*89 separate studies by two, or more assessors, of at least 10 different patients.*" They go on to note that "*Eighty-nine separate studies of 10 different patients each is not the equivalent of one study of 890 patients.*" While we agree that it is not the exact "equivalent," there is no methodological or statistical reason that data from "*89 separate studies*" examining the same research question cannot be synthesized. The aggregation of such studies is routinely accomplished in meta-analytic investigations where the data and methods are much less homogeneous than the data from the current study (8, 9). There is no statistical reason that reliability values from the different settings cannot be aggregated.

Statistical Procedures. Dickson & Köhler (3) are critical of the ICC for the summed FIM total, domain, and subscale scores. While there is some disagreement in the research literature (10) regarding the appropriateness of the ICC for use with ordinal data, several investigators have demonstrated that the ICC is appropriate for use with ordinal level scales (6, 11). For example, Tinsley & Weiss (11) state, "*The intraclass correlation is recommended as the best measure of interrater reliability available for ordinal and interval level measurement*"(p. 373). When the ICC and kappa are used interchangeably to examine ordinal level scales, they will not produce the same values because of the unequal marginal distributions for kappa (12). Several investigators (13, 14) have published formulas that allow kappa values to be converted to ICC. Dickson & Köhler (3) do not appear to be aware of this similarity and their comments regarding the interpretation of kappa are puzzling. Specifically, the statement that, "*there is no reason, of course, given the nature of this study, for a high kappa to be more acceptable than a low kappa, as neither may have been greater than a value expected by chance,*" makes no statistical sense.

There is no reason why confidence intervals cannot be computed for the summed reliability values. Dickson & Köhler's (3) argument that confidence intervals should be computed is a good one. While we agree that this is a useful suggestion it is important to note that other reliability studies recently published in this and other major rehabilitation journals have not provided confidence intervals for reported reliability values. The absence of confidence intervals cannot be considered a major flaw in the investigation.

Criterion Facilities. Dickson & Köhler state that "*no information is given as to when the selection techniques used to establish the criterion facilities were decided. If the criteria were decided post hoc, it seems self-evident that those meeting the criteria had high interrater reliability scores.*" The description of the requirements for the 24 "criterion facilities" is included on page 117. The criteria used to identify these facilities were introduced prior to the collection of data for this study using completely different rehabilitation centers and hospitals. The criteria were originally established to ensure the accuracy and reliability of data submitted to the Uniform Data System for Medical Rehabilitation.

We stand by the conclusion of our study that the 7-level FIM demonstrates high interrater reliability. This finding has been supported by other recent investigations (15, 16). Replication of the results by other researchers provides the most convincing evidence regarding the reliability of the FIM. Additional studies are certainly needed and welcome. We are confident that the results from future investigations will confirm that the findings we reported are correct.

REFERENCES

1. Ikegami, N.: Functional assessment and its place in health care. N Engl J Med *332:* 598–599, 1995.
2. Hamilton, B. B., Laughlin, J. A., Fiedler, R. & Granger, C.V.: Interrater reliability of the 7-level Functional Independence Measure (FIM). Scand J Rehabil Med *26:* 115–119, 1994.
3. Dickson, H. G., & Köhler, F.: Letter to the Editor. Scand J Rehabil Med *27:* 253–254, 1995.
4. Dunn, G.: Design and analysis of reliability studies: the statistical evaluation of measurement error. New York, Oxford University Press, 1989.
5. Cronbach, L. J., Glesser, G., Nanda, H., et al.: The dependability of behavioral measurements: theory of generalizability of scores and profiles. New York, Wiley, 1972.
6. Berk, R. A.: Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. Am J Ment Defic *83:* 460–466, 1979.

7. Fleiss, J. L.: The measurement of interrater agreement. In Statistical methods for rates and proportions. New York, Wiley, 1981.
8. Cooper, H. M.: Integrating research: a guide for literature reviews (2nd ed.). Newbury Park, CA, Sage, 1989.
9. Glass, G. V., McGraw, B. & Smith, M. L.: Meta-analysis in social research. Beverly Hills, CA, Sage, 1981.
10. Bartko, J. J., & Carpenter, W. T.: On the methods and theory of reliability. J Nerv Mentl Dis *163:* 307–317, 1976.
11. Tinsley, H. E. & Weiss, D. J.: Interrater reliability and agreement of subjective judgements. J Couns Psych *22:* 358–376, 1975.
12. Soeken, K. L. & Prescott, P. A.: Issues in the use of kappa to estimate reliability. Med Care *24:* 733–741, 1986.
13. Suen, H. K., Ary, D., & Ary, R. M.: A note on the relationship among eight indices of interobserver agreement. Behav Assess *8:* 301–303, 1986.
14. Rae, G.: The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. Educ Psych Measur *48:* 367–374, 1988.
15. Fricke, J., Unsworth, C., & Worrell, D.: Reliability of the functional independence measure with occupational therapists. Aust Occup Ther J *40:* 7–15, 1993.
16. Ottenbacher, K., Mann, W., Granger, C. V., Tomita, M., Hurren, D., & Charvat, B.: Interrater agreement and stability of functional assessment in the community-based elderly. Arch Phys Med Rehabil *75:* 1297–1301, 1994.

Byron B. Hamilton, M.D., Ph.D.
Judith A. Laughlin, R.N., Ph.D.
Roger C. Fiedler, Ph.D.
Carl V. Granger, M.D.
Kenneth J. Ottenbacher, Ph.D.