Scand J Rehab Med 29: 267–272, 1997

LETTER TO THE EDITOR

FIM MEASUREMENT PROPERTIES AND RASCH MODEL DETAILS

We appreciate the opportunity provided through the article by Drs. Dickson & Köhler (1) to clarify aspects of the Functional Independence Measure (FIM^{SM}) instrument's measurement properties and details of the Rasch model which are not widely appreciated.

Dickson & Köhler question the unidimensionality of the FIM

Dickson & Köhler (1) state "FIM mobility items are not measuring a unidimensional construct". We note that unidimensionality is an ideal. Empirical data always fall short of this ideal. For instance, a wooden ruler or a spring balance always falls short as ideal measuring instruments. We use the wooden ruler and the spring balance because they are good enough and useful enough. Certainly the FIM falls short of ideal unidimensionality, but our analysis (and also Dickson & Köhler's analysis) indicates that the FIM is close enough to perfect unidimensionality and is good enough and useful enough to treat as though it were unidimensional.

Dickson & Köhler ask whether "seeing people who could walk and not swallow" is a disproof of Rasch unidimensionality? We answer "No," as the reverse would be a disproof! If everyone who walks were always able to swallow, then the relationship between walking and swallowing would be deterministic, not probabilistic. The Rasch model is a probabilistic model. It does not assert that a person who walks must be able to swallow. It only asserts that it is highly probable that a person who walks is also able to swallow. Dickson & Köhler confuse the Rasch model with Guttman's deterministic model. In real life, there are always exceptions to theoretical models, but these exceptions do not necessarily disprove the general rule.

Reports by Heinemann et al. (3) and Linacre et al. (6), in particular, showed that the 18-item FIM was comprised of two unidimensional measures, physical and cognitive, based on a sample of thousands of patients. The items for expression, especially, and comprehension, to a lesser degree, would misfit if stroke patients with aphasia from left hemisphere lesions were not analyzed separately. Of the major diagnostic groups treated in comprehensive medical rehabilitation, three major motor patterns emerged: (a) orthopedic conditions, (b) neurologic conditions, and (c) spinal cord dysfunction. Within the orthopedic and neurologic conditions there was a tendency for misfitting of bowel, bladder and stair climbing items. This could be perceived as a threat to unidimensionality. Variation in scoring bladder and bowel function may be explained on the basis that they are not only under voluntary control but also autonomic control. Therefore, these functions are subject to pathologies that may be coincidental with the major pathology. Another explanation may lie in the wording of the scoring criteria. This is being monitored. The stair climbing item tends to fit better at the time of discharge than at admission. It is presumed that the rehabilitation clinicians are more familiar with and more willing to test the true ability of patients at discharge. Further, the stair climbing item is useful as an "anchor" at the difficult end of the measure.

No misfitting of items threatened the validity of the total measure.

Dickson & Köhler express problems with rasch analysis of FIM data

Dickson & Köhler ask whether the Rasch model is necessary and sufficient for constructing interval measures from ordinal data. We note that there are numerous proofs and demonstrations of this (4).

For example, the authors state "For a FIM motor function interval scale, the ability to climb stairs would imply necessarily an ability to eat normally." "We have seen people who could walk but not swallow, swallow but not walk, and even climb stairs but not swallow. These patient do not fit the interval scale." We agree that exceptions such as these do exist. However, that misapplies what the Rasch model reveals.

As background, Rasch analysis ("a part of a general system called Item Response Theory or Latent Trait Analysis") develops a model of expected responses based upon the interactions of persons of differing abilities with items of differing difficulties. Each person and item is compared in terms of fitting the

model of a unidimensional continuum. We understand and expect that some individuals may differ from the model.

Successful application of Rasch analysis requires that the items collaborate to define a single, dominant (unidimensional) construct (e.g. physical is different from mental). Also, there must be a hierarchical structure such that the items form a continuum in which they are easy at one end and difficult at the other. The spacing between items is marked by their calibrations. Item calibrations quantify the probabilities that persons taking the test can pass items in the test. Items that are calibrated lower have a higher probability of being passed than higher calibrated items. Items are calibrated and persons are measured on the same scale.

Dickson & Köhler state "the method of scoring FIM does not appear to us to allow a lower asymptote of 0, as the chance of producing a correct score is 1/7 for any FIM item". Our FIM analysis is performed with the Rasch rating scale model, which compares the probabilities of observing pairs of adjacent categories. The log ratio of the probabilities can have any value between minus infinity to plus infinity. Dickson & Köhler may be referring to someone else's paper (which we have not seen) that may have analyzed the FIM items as though they were 7-option, multiple-choice questions.

Dickson & Köhler state "No description of the sample distribution exists in Rasch analysis". We note that since measurement is the purpose of Rasch analysis, description of the person measures and item calibrations is fundamental to Rasch analysis, not just on a sample level, but also for each individual. But do they perhaps refer to the sample distribution of misfit? There are many papers discussing the sampling distributions of the fit statistics, including those by Smith (7–9).

Dickson & Köhler state "Any system of measurement based on probabilities must necessarily be imprecise". We note that all measurement is imprecise to some extent. It is the standard error of measurement (a concept founded in probability theory) that enables us to quantify the imprecision. Dickson & Köhler imply that there is some type of measurement that is perfectly precise. Whatever that type of measurement is, it is definitely not supported by the FIM instrument. No one has ever reported an FIM reliability of 1.0, i.e. perfect precision.

They also state "Rasch analysis does not allow a post hoc scaling to make the summed FIM scores more useful". We note that the analyst is free to perform whatever data manipulation is desired. But, in departing from the Rasch model, the analyst gives up equal-interval linearity and statistical independence from the sample distribution.

Dickson & Köhler appear to advocate rescaling the FIM observations, i.e. weighting the observations, in order to maximize the correlation between FIM summed scores and some diagnostic criteria. This approach capitalizes on idiosyncrasies in the local sample and sacrifices any chance there might be to construct a measurement system that goes beyond sample description. Under such circumstances, Dickson & Köhler recommend "careful matching for age, sex...", requirements that no one feels compelled to meet when comparing measures obtained with wooden rulers or spring balances.

Principal components analysis of the FIM

According to Dickson & Köhler the principal components analysis of the FIM indicates multiple factors. We note that the discovery of too many factors is a known problem of factor analysis (5). It has been noted that differences in difficulty are represented in the factorial configuration as additional factors (2).

Dickson & Köhler state that "To explain more than 80% of the variance three factors are required..." Does this mean that they consider secondary factors to be too large? We note that Smith & Miao (10) simulated data to fit the Rasch model and then performed principal components analysis. They report that, in their perfectly simulated data, the first factor is about 7 times larger than the second factor (10). Dickson & Köhler's first factor is 6.3 times as large as the second factor, indicating that, even with the well-known problems in bowel and bladder, FIM data are almost as unidimensional as "perfect" data.

Reproducibility of analysis and comparing disability groups

In the case of the FIM, across diagnostic impairment groups typically admitted to comprehensive medical rehabilitation programs, eating has low calibration (an easy item) and stair climbing has high calibration (a difficult item). Since they are on either end of the continuum of item difficulties, it is expected that eating would not correlate highly with stair climbing. Most patients with difficulty in eating find stair climbing impossible. Most patients attempting stair climbing

Table I. *Expected relationship of item scores for eating and stair climbing*

Eating	Stair climbing
1	1
2	1
3	1
4	1
5	1
6	2
6	3
7	4
7	5
7	6
7	7

have no trouble with eating. In fact, the expected relationship of FIM observations is shown in Table I.

However, each of the FIM items may correlate more highly with items that are nearer to that item's level of difficulty. On the other hand items that correlate too highly with each other may signal collinearity or redundancy of items. Nonetheless, correlation of items with that of the summed value is desired. Since Dickson & Köhler do not report the ability distribution of their patients, we cannot compute the expected correlation for their data.

Dickson & Köhler state "Caution should be exercised when performing comparisons between different groups for different rehabilitation units if careful matching for age, sex, impairment and socio-economic variables has not occurred, and the statistical methodologies employed should be appropriate to the data." The issues of reproducibility and comparability are large. Here, we will only address aspects that relate to functional assessment.

Reproducibility is the constancy in the order and spacing of the item calibrations along the continuum. Reproducibility is tested in two ways: by applying the item calibrations to other persons who share similar characteristics and by applying them to the same persons at different times. Reproducibility of item calibrations is essential for comparing results from different groups and results from the same persons over time.

Comparability is the consequence of reproducibility. It means that persons compared are performing through the same common pathway and are not responding to the items in a way that is at variance from the main group. It is the extent to which the same rules of probability are applicable from one comparison to another. However, in the nature of things, some variance is inevitable.

Reproducibility may not be apparent in the authors' report for two reasons. We are unable to know whether the patients in the groups analyzed for correlation and principal components analyses varied in their patterns of disability; in other words, they may not have shared similar characteristics. Studies were performed on admission but not discharge; therefore, reproducibility could not be estimated over time.

SUMMARY

To summarize, we take issue with the criticisms of Dickson & Köhler for two main reasons:

1. Rasch analysis provides a model from which to approach the analysis of the FIM, an ordinal scale, as an interval scale. The existence of examples of items or individuals which do not fit the model does not disprove the overall efficacy of the model; and
2. the principal components analysis of FIM motor items as presented by Dickson & Köhler tends to undermine rather than support their argument. Their own analyses produce a single major factor explaining between 58.5 and 67.1% of the variance, depending upon the sample, with secondary factors explaining much less variance.

Finally, analysis of item response, or latent trait, is a powerful method for understanding the meaning of a measure. However, it presumes that item scores are accurate. Another concern is that Dickson & Köhler do not address the issue of reliability of scoring the FIM items on which they report, a critical point in comparing results. The Uniform Data System for Medical Rehabilitation (UDSMR^{SM}) expends extensive effort in the training of clinicians of subscribing facilities to score items accurately. This is followed up with a credentialing process. Phase 1 involves the testing of individual clinicians who are submitting data to determine if they have achieved mastery over the use of the FIM instrument. Phase 2 involves examining the data for outlying values.

When Dickson & Köhler investigate more carefully the application of the Rasch model to their FIM data, they will discover that the results presented in their paper support rather than contradict their application of the Rasch model! This paper is typical of supposed refutations of Rasch model applications. Dickson & Köhler will find that idiosyncrasies in their data and misunderstandings of the Rasch model are the only basis for a claim to have disproven the relevance of the model to FIM data. The Rasch model is a mathematical theorem

(like Pythagoras') and so cannot be disproven by empirical data once it has been deduced on theoretical grounds. Sometimes empirical data are not suitable for construction of a measure. When this happens, the routine fit statistics indicate the unsuitable segments of the data. Most FIM data do conform closely enough to the Rasch model to support generalizable linear measures. Science can advance!

REFERENCES

1. Dickson, H. G. & Köhler, F.: The multi-dimensionality of the motor items precludes an interval scaling using Rasch analysis. Scand J Rehab Med *26:* 159–162, 1996.
2. Ferguson, G. A.: The factorial interpretation of test difficulty. Psychomet *6:* 323–329, 1941.
3. Heinemann, A. W., Wright, B. D., Hamilton, B. B. & Granger, C. V.: Relationships between impairment and physical disability as measured by the functional independence measure. Arch Phys Med Rehab *74:* 566–573, 1993.
4. Linacre, J. M. (ed.): Rasch measurement transactions, Vol. 1. MESA Press, Chicago, 1995.
5. Linacre, J. M. (ed.): Rasch measurement transactions, Vol. 2. MESA Press, Chicago, 1996.
6. Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V. & Hamilton, B. B.: The structure and stability of the functional independence measure. Arch Phys Med Rehab *75:* 127–132, 1994.
7. Smith, R. M.: The distributional properties of Rasch standardized residuals. Educ Psych Meas *48:* 657–667, 1988.
8. Smith, R. M.: IPARM: Item and Person Analysis with the Rasch Model. MESA Press, Chicago, 1991.
9. Smith, R. M.: The distributional properties of Rasch item fit statistics. Educ Psych Meas *51:* 541–565, 1991.
10. Smith, R. M. & Miao, C. Y.: Assessing unidemsionality for Rasch measurement. *In* Objective Measurement: Theory into Practice II (ed. M. Wilson). Ablex, Norwood, NJ, 1994.

B. D. Wright[1], J. M. Linacre[1], R. M. Smith[2], A. W. Heinemann[3] and C. V. Granger[4]

[1]MESA Lab, Chicago University, [2]Rehabilitation Foundation Inc, Wheaton, IL, [3]Rehabilitation Inst, Northwestern University, Chicago and [4]Department of Rehabilitation Medicine, University of Buffalo, School of Medicine, 232 Parker Hall, 3435 Main St., Buffalo, NY 14214-3007, U.S.A.

RESPONSE TO LETTER BY WRIGHT ET AL.

There is nothing in the letter by Wright et al. that defeats our proposition that the Functional Independence Measure (FIM) motor item scale is multidimensional. Heinemann et al. (3) and Linacre et al. (5), using Rasch analysis, have found a misfit of bladder, bowel and stair climbing items. Tsuji et al. (7) reported misfit in the motor items of bladder, bowel, stair climbing and bathing, and Chang & Chan (1) reported misfit of eating, bladder management and toilet transfer, while Pollak et al. (6) found a misfit of grooming, bowel and bladder items. Chang & Chan considered that the assumption of unidimensionality may have been violated by the FIM motor items as evidenced by the number of misfitting or redundant items found in their analysis.

We performed a principal component analysis (Statistics for Windows, minimum eigenvalue set at 1) on the admission FIM motor items of a sample of 295 patients with stroke admitted to our Rehabilitation Unit. When the bladder, bowel and stair climbing items are omitted from the analysis, the number of principal components falls from two to one, indicating a change to statistical unidimensionality. This single component explains 76% of the variance in the data set. Therefore, there is no question that the omission of the items identified as misfits by Wright et al. in their letter improves the statistical properties of the FIM motor item section. This finding again demonstrates that the current FIM motor items are multidimensional.

To turn to specific points raised in the letter by Wright et al., we use the conventional definition in behavioural science of an interval scale as providing a strict order of categories with an equal distance between each category. A Guttmann scale demands only a strict ordering of categories and unidimensionality, but not an equal interval. Strictly speaking, an interval scale should contain Guttmann properties. A Rasch probabilistic interval scale differs from an interval scale by providing probability intervals estimated from data obtained from a non-interval scale, rather than being a scale designed to be interval by reference to a standard external criterion. In Rasch analysis, the scale is formed entirely by reference to the sample from which it is generated. What this means, as Wright et al. indicate, is that there will be people who will be able to be scored on the FIM but who will not fit a Rasch generated scale. We contend that this inability of people to fit the scale argues against the adequacy of a Rasch generated interval scale of FIM motor items based on conventional usage of the terms, even assuming that use of the technique is valid in this instance. We also contend that swallowing and walking are different activities, but we accept the argument that, for the purposes of a statistical construct, they may form

part of a probabilistic scale, though we confess a preference for constructs that also make sense.

We state correctly in our paper that the claim that the Rasch model is necessary and sufficient for the construction of measures in any science was not referenced by Wright & Linacre (8). To assert that the Rasch model is necessary for construction of measures in any science is an excessive and unsubstantiated claim. Wooden rulers and spring balances obviously predate Rasch analysis.

We are grateful to Wright et al. for their correction regarding the issue of the lower asymptote of the Item Characteristic Curve being in fact zero, allowing the log ratio of the probabilities of observing pairs of adjacent categories to vary as they describe.

Wright et al. argue that since Rasch analysis provides descriptions of the person measures and item calibrations, a sample distribution exists. We refer to the fact that when Rasch analysis is performed on different samples of patients, different item calibrations are found, as evidenced by the differing results reported by Grimby et al. (2), Heinemann et al. (3), Linacre et al. (5), Pollak et al. (6), Chang & Chan (1), and Tsuji et al. (7). This variability needs to be accounted for in some way if there is to be a useful general application of measures derived from Rasch analysis. It is pointless to persist in the claim that Rasch analysis applied to FIM provides measures free from the sample when the practical application of the technique provides evidence to the contrary. As Holland (4) states, item parameters and subject abilities are always estimated relative to a population, even if this fact is obscured by the mathematical properties of the models used. The samples providing the FIM data on which Rasch analysis is performed are certainly not random samples of the population of people with disability on whom FIM may be applied, but groups of people selected according to differing sets of criteria. We agree with Holland that the goal of obtaining sample-free item analysis with Item Response Theory models is unattainable because the effect of the population from which the sample is drawn will always be present, and the best that can be hoped for is that the effect is small enough to be ignored.

Rasch analysis assumes the presence of a statistically unidimensional variable that underlies performance across all samples. Every clinician knows that, despite the manual for FIM, there are a number of different ways of performing the tasks, as well as differences in the settings where activities are performed. There may be a number of reasons why a person achieves a particular FIM score. A score of 6 on a FIM motor item, for instance, may mean a subjective judgement about the length of time taken to achieve a task, or the use of an aid, or a subjective concern for safety. As at least three different situations may be responsible for the one score, and as the setting is likely to be different from Rehabilitation Unit to Rehabilitation Unit, the assumption that items are measuring the same ability in the same way across all samples is unreasonable. Additionally, with the data used by Linacre et al. (5), Heinemann et al. (3), and Tsuji et al. (7), a score of 1 might represent missing data, rather than meaning that total assistance is required.

We deny that we imply in our paper that there is some type of measure that is perfectly precise. We agree that all measurement is imprecise to some extent. Rasch analysis provides scales that incorporate a sampling error, plus an error introduced by the application of the technique. A probabilistic scale cannot be more accurate than the scale from which it is produced. Wright et al. state that Rasch analysis assumes that the original classification in the sample is accurate, a condition that they state is unlikely with the FIM.

We agree with Wright et al. that linear equal interval measures of behaviour that are independent of the sample from which they were derived offer considerable advantages to ordinal and categorical measures. These measures in rehabilitation medicine are best constructed using physical measures such as time, length and so on, rather than by *post hoc* statistical manipulations based on a number of assumptions, many or all of which may be open to question.

Wright et al. believe that stair climbing is useful as an anchor at the difficult end of the measure, despite it being identified as a misfit item, despite it being identified by Heinemann et al. (3) and Tsuji et al. (7) as an item frequently not tested and hence scored as 1, and despite the fact that for patients in one hospital in the study performed by Grimby et al. (2), walking was more difficult than stair climbing.

Wright et al. argue that examples of items or individuals which do not fit the Rasch model do not disprove its efficacy. Misfits are expected with Rasch analysis and these may be interpreted as evidence of item or person inappropriateness rather than model inappropriateness. So far the list of misfits includes eating, bathing, toilet transfers, grooming, bladder, bowel, and stair climbing, representing 7 out of the 13 items in the motor item scale. All Rasch scales of FIM reported on different samples so far differ from each other not only in logit distance between items but also in order of item difficulty, again

even when the diagnostic group was the same. This state of affairs is indicative of model inappropriateness.

We do not use the data facility offered by the State University of New York, preferring to analyse our data in-house. We do use the Uniform Data System for Medical Rehabilitation form, video training and manual. Our staff profile is stable, and one of us (F.K.) collects, checks and enters the data. We think that our test–retest reliability would be high, and though we freely admit that the possibility of a systematic error in our data exists, we think that the probability of a clinically significant error is small.

The contention of Wright et al. that the Rasch model cannot be disproven by empirical data must lie with the many similar contentions about mathematical models of the world that have been made over the years, and which litter the historical accounts of the development of science. A model may be correct mathematically, but be unable to explain adequately the physical behaviour it supposedly represents. In science, as distinct from statistics, empirical data are important and do not disqualify themselves. Rather, someone chooses to ignore them. Ignoring data that do not fit a model is poor science. Misfitting data generally indicate that there is a problem with the adequacy of the model rather than with the data. Science has advanced, can advance, and will advance, regardless of Rasch analysis. Unthinking application of elegant statistical theory is no substitute for a clear hypothesis and a well-constructed experiment. The use of Rasch analysis on the FIM motor item scale is not indicated, as the FIM motor item scale is not a unidimensional construct.

REFERENCES

1. Chang, W. & Chan, C.: Rasch analysis for outcomes measures: some methodological considerations. Arch Phys Med Rehab *76:* 934–939, 1995.
2. Grimby, G., Gudjonsson, G., Rodhe, M., Sunnerhagen, K. S., Sundh, V. & Östensson, M.: The Functional Independence Measure in Sweden: experience for outcome measurement in rehabilitation medicine. Scand J Rehab Med *28:* 51–62, 1996.
3. Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B. & Granger, C.: Relationships between impairment and physical disability as measured by the functional independence measure. Arch Phys Med Rehab *74:* 566–573, 1993.
4. Holland, P. W.: On the sampling theory foundations of Item Response Theory models. Psychomet *55:* 577–601, 1990.
5. Linacre, J. M., Heinemann, J. M., Wright, B. D., Granger, C. & Hamilton, B. B.: The structure and stability of the functional independence measure. Arch Phys Med Rehab *75:* 127–132, 1994.
6. Pollak, N., Rheault, W. & Stoecker, J. L.: Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. Arch Phys Med Rehab *77:* 1056–1061, 1996.
7. Tsuji, T., Sonoda, S., Domen, K., Saitoh, E., Liu, M. & Chino, N.: ADL structure for stroke patients in Japan based on the Functional Independence Measure. Am J Phys Med Rehab *74:* 432–438, 1995.
8. Wright, B. D. & Linacre, J. M.: Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehab *70:* 857–860, 1989.

Hugh G. Dickson and Friedbert Köhler

Liverpool Health Service, Liverpool Hospital, Liverpool NSW 2170, Australia