Effect sizes can be misleading: Is it time to change the way we measure change?

Jeremy C Hobart, Stefan J Cano, a J Thompson

Effect sizes can be misleading:

Is it time to change the way we measure change?

Jeremy C Hobart PhD,[1,2] Stefan J Cano PhD,[1,2] Alan J Thompson MD[2]

The Neurological Outcome Measures Unit,

Peninsula College of Medicine and Dentistry,[1] Devon and UCL Institute of Neurology,[2] London UK.

Correspondence to Dr. Jeremy Hobart, Reader and Honorary Consultant Neurologist, Department of Clinical Neuroscience, Peninsula College of Medicine and Dentistry, Room N16 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK

T: +44 (0) 1752 315272; F: +44 (0) 1752 315254; E: Jeremy.Hobart@pms.ac.uk

Word count: 3415

Keywords: rating scales, responsiveness, psychometrics, Rasch analysis, measurement

ABSTRACT

Objectives: Previous comparisons of the ability to detect change of the Barthel Index (BI) and Functional Independence Measure motor scale (FIMm) have implied these two scales are equally responsive when examined using traditional effect size statistics. Clinically, this is counter-intuitive, as the FIMm has greater ***potential*** to detect change than the BI, and raises concerns about the validity of effect size statistics as indicators of rating scale responsiveness. To examine these concerns in this study we applied a sophisticated psychometric analysis, Rasch measurement to BI and FIMm data.

Methods: BI and FIMm data were examined from 976 people at a single neurorehabilitation unit. Rasch analysis was used to compare the responsiveness of the BI and FIMm at the group comparison level (effect sizes, relative efficiency, relative precision) and for each individual person in the sample by computing the significance of their change.

Results: ***Group-level*** analyses from both interval measurements and ordinal scores implied the BI and FIMm had equivalent responsiveness (BI and FIMm effect size ranges -0.82 to -1.12 and -0.77 to -1.05, respectively). However, ***individual person-level*** analyses indicated that the FIMm detected significant improvement in almost twice as many people as the BI (50%, n=496 versus 31%, n = 298), and recorded less people as unchanged on discharge (FIMm= 4% n=38; BI=12% n=115). This difference was found to be statistically significant (Chi-square, 273.81; $p<0.000$).

Conclusions: Findings demonstrate that effect size calculations are limited and potentially misleading indicators of rating scale responsiveness at the group comparison-level. Rasch analysis at the individual person-level showed the FIMm's superior responsiveness, supporting clinical expectation, and its added value as a method for examining and comparing rating scale responsiveness.

INTRODUCTION

Rating scales must be able to detect clinically important change if they are to be used as outcome measures in clinical trials.[1-3] The relative responsiveness of competing rating scales is a critical factor in the selection of scales for studies.[4, 5]

This study examines the responsiveness of two widely used activity limitation rating scales, the Barthel Index (BI) [6] and Functional Independence Measure motor scale (FIMm),[7] in 1400 people who have undergone neurorehabilitation. Previously,[8] we demonstrated problems with the BI (substantial item and scale ceiling / floor effects), which cautioned against its appropriateness in evaluating neurorehabilitation. This led us to hypothesize that the BI would be more responsive if its items had more response categories. We tested this hypothesis by comparing the BI with the FIMm, a scale that uses the same items but has more item response categories. [19] Results showed that the FIMm had greater ***potential*** to detect change (smaller item and total score floor and ceiling effects than the BI), and detected change in more people undergoing rehabilitation. Despite this evidence of better ***potential*** to detect change, the FIMm and BI had almost identical effect size calculations implying the same ***ability*** to detect change at the group comparison level. This finding is counter intuitive clinically, and questions the validity of effect size statistics as indicators of rating scale responsiveness.

To explore this issue, we examined the relative responsiveness of the BI and FIMm in the same dataset using a more sophisticated psychometric method, Rasch measurement,[10-12]. This method advances the analysis of rating scale responsiveness in three specific ways. First, Rasch analysis enables interval-level (linear) measurements of activity limitation to be estimated from ordinal-level BI and FIMm total scores. This is valuable because fixed changes in ordinal total scores (eg 10 points) imply variable changes in interval-level measurements across the scale range.[2, 3, 13]. Thus, analysing total scores may hide responsiveness differences between scales. Second, Rasch analysis enables a legitimate examination of changes in activity limitation at the ***individual person-level***, in addition to comparisons at the ***group-level***. In contrast, traditional

psychometric analyses are **not** recommended for individual person decision making.[3, 14] The third benefit is that Rasch analysis enables scales measuring the same construct, as the BI and FIMm purport, to be equated on a common metric. [2] This enables people's measurements on the BI and FIMm to be compared on an identical "ruler" of activity limitation.[15]

METHODS

Sample

Data were available for 1495 people who underwent neurorehabilitation at a single UK unit. In our analyses we included cases with complete admission and discharge data, and excluded all people who had the minimum possible score or the maximum possible score on either scale at either admission or discharge. This was to ensure that results and inferences were not confounded by floor and ceiling effects. This study was approved by the local ethics committee of the National Hospital for Neurology and Neurosurgery, London, UK. Details of the sample have been reported elsewhere.[8, 9]

BI and FIMm

Table 1 shows the BI and FIMm. The BI has 10 items. Two items have two response categories, six items have three response categories, and two items have four response categories.[6] The FIMm has 13 items. All items have seven response categories.[7] The BI and FIMm share eight identical items. The two remaining BI items (dressing, transferring) are represented in the FIMm by five items (dressing upper body, dressing lower body, bed transfer, toilet transfer, bath transfer).

Table 1: The items and item response categories of the BI and FIMm

Barthel Index		FIMm motor scale	
Item	**No. response categories**	**Item**	**No. response categories**
Feeding	3	Feeding	7
Grooming	2	Grooming	7
Bathing	2	Bathing	7
Dressing	3	Dressing upper body	7
-		Dressing lower body	7
Toileting	3	Toileting	7
Bladder	3	Bladder	7
Bowels	3	Bowels	7
Transfer	4	Bed transfer	7
-		Toilet transfer	7
-		Bath transfer	7
Mobility	4	Walk/wheelchair use	7
Stairs	3	Stairs	7

Analysis

Rasch analysis is a method of analysing rating scale data. In brief, the analysis examines the extent to which the data satisfy the requirements of a mathematical model - the Rasch measurement model.[10-12] This model articulates a theory of how rating scales must perform if the values they generate are to be considered scientific measurements.[3] Thus, when the data fit the requirements of the Rasch model, within reason, there is evidence that scales (here the BI and FIMm) are measurement instruments. Under these circumstances the analysis is able to transform scale scores for people, which are by necessity ordinal, into interval level measurements. These estimates, termed “person locations” to distinguish them from ordinal scale scores, are in log-odds units (logits). For each individual person’s location the analysis also generates a bespoke standard error. Rasch analysis is explained elsewhere.[3, 11, 15-18]

Rasch analyses were performed using RUMM2020.[19] We analysed BI and FIMm data together as a co-calibrated pool of items, organised in a racked (by scale) and stacked (by time point) format. We compared the responsiveness of the BI and FIMm scale at both *group* and *individual person-level*.

Group-level comparison

The relative responsiveness of the BI and FIMm was examined at the group-level by comparing admission and discharge person locations using four standard indicators: two effect size calculations (Kazis' effect size- ES,[20] standardized response mean – SRM[21]), relative efficiency (RE; pair–wise squared t-values from paired samples t-tests[22]), relative precision (RP; ratio of pair-wise F-values from one-way ANOVA).[5] We compared the results of these analyses, which are derived from person locations and are interval level measurements, with the results of the same analyses undertaken on BI and FIMm total scores, which are generated by summing item scores and are ordinal level data. This was to determine if estimates of responsiveness based on interval level measurements differed, in magnitude or inference, from those based on ordinal level scores.

Individual person-level comparison

The relative responsiveness of the BI and FIMm was compared at the individual person level. This was achieved by computing, for each and every person, the significance of ***their own*** change in activity limitation measurement ('Sig Change'). First, we computed the size of the change score for each individual person (discharge location –admission location). Second, we computed that size of the error associated with their change (standard error of the difference) for each individual person as the square root, of the sum, of the squared standard error values at admission and discharge. Third, we computed the significance of the change for each individual by dividing their change score by their standard error of the difference (ie how big is their change in standard error units). Finally, we

categorised the significance of each person's change into one of 5 groups according to the size and direction of the significance of change value. The formulae are as follows:

$$\text{Significance of change (Sig Change)} = \frac{\text{Discharge location} - \text{Admission location}}{\text{Standard error of the difference (SEdiff)}}$$

Where, SEdiff for a person = $\sqrt{\{(\text{SE admission location})^2 + (\text{SE discharge location})^2}$.

Significance of change values obtained from this formula were categorized into 5 groups:

Significant improvement = Sig Change $\geq$ +1.96;

Non significant improvement = 0 < Sig Change $\leq$ +1.95;

No change = Sig Change = 0;

Non significant worsening = -1.95 $\leq$ Sig Change < 0;

Significant worsening = Sig Change $\leq$ -1.96.

Now, we can simply count the numbers of people achieving each level of significance of change, and compare the distributions for the FIMm and BI using a Chi-square test and relative risk statistics.

RESULTS

Sample

Data were available for 1495 people. Complete data at admission and discharge were available for 1396 (93% of sample), of which n= 976 (70%) did not score at either the floor or ceiling of either scale at both time points. In the total sample, at both admission and discharge, total score floor and ceiling effects were lower for the FIMm than the BI. This indicates that the FIMm provides an

extended range of measurement.* As predicted for a sample of people undergoing an intervention aimed to improve function, the floor effects were smaller on discharge than admission, and the ceiling effects were larger on discharge than admission. These values were: FIMm admission (floor = 0.8%, ceiling = 0.3%), discharge (floor = 0.2%, ceiling = 1.7); BI admission (floor = 1.1%, ceiling = 5.3%), discharge (floor = 0.1%, ceiling = 27.9%). Overall 519 people were at either the floor or the ceiling, on either scale. Of these, only 30 people (5.9%) scored were at the floor or ceiling on both scales.

The mean age and length of rehabilitation were 49 years (sd: 15) and 36 days (sd: 26) respectively, and 56% of the cohort was female. The main diagnostic groups were multiple sclerosis (46%), Stroke (18%), spinal cord syndromes (17%). Fuller details of the samples have been reported before.[9]

Group-level comparison of BI and FIMm relative responsiveness

The responsiveness data (Table 2) generated by the analysis of both interval measurements (BI and FIMm person locations) and ordinal scores (BI and FIMm scale scores) shows that both scales quantified significant changes at the group level, and that both scales had near identical similar responsiveness according to the four analyses. Conclusions reached about the relative responsiveness of the BI and FIMm were essentially the same for both interval measurements and ordinal scores.

* At admission and discharge, total score ceiling effects were lower for the FIMm than the BI. Thus, a significant proportion of patients who scored at the maximum of the BI were within the floor and ceiling of the FIM, implying that latter does in fact have an extended range of measurement and thus a better *potential* to detect change.

Table 2a: BI with FIMm: admission, discharge, change and relative responsiveness (n=976)

	Ordinal Scores		Interval Measurements (locations)	
	BI	**FIMm**	**BI**	**FIMm**
Possible range	0 - 20	13 - 91	-3.15 to +2.86*	-2.63 to +4.98*
Admission				
mean	10.5	53.5	-0.086	0.017
sd	4.9	18.2	1.372	0.930
Discharge				
mean	14.5	67.7	1.152	0.932
sd	4.7	17.3	1.455	1.154
Change				
Mean	-4.0	-14.1	-1.238	-0.915
sd	3.7	13.5	1.109	0.868
Indicators of group-level responsiveness				
t-test				
t	-33.52	-32.76	-34.877	-32.932
p	<0.000	<0.000	<0.000	<0.000
RE	100%	96%	100%	89%
1-way ANOVA				
F	337.3	308.8	373.50	372.11
P	<0.000	<0.000	<0.000	<0.000
RP	100%	91%	100%	99%
Effect size				
Kazis	-0.82	-0.77	-0.90	-0.98
SRM	-1.08	-1.04	-1.12	-1.05

*These range from 'minus' to 'plus' values as the person locations are transformed log-odds units (logits) centered around a mean of zero. These possible range values presented here are extrapolated estimates as extreme locations cannot be accurately estimated.
RE = relative efficiency = $(\text{t-scale})^2 / (\text{t-BI})^2$
RP = relative measurement precision = (F-scale) / (F-BI)
(RE and RP use "best" scale as arbitrary denominator)
Kazis effect size = mean change / sd admission
SRM = standardized response mean = mean change / sd change

Table 2b: BI with FIMm: relative responsiveness, indicators of individual person-level responsiveness, Chi-ssquare and relative risk statistics (n=976)

	BI	FIM	
Significance of change	**%(n)**	**%(n)**	**RR (upper, lower 95% CI, p-value)***
Significant improvement	30.5 (298)	49.8 (486)	1.6 (1.5, 1.8 p<0.000)
Non-significantly improvement	52.7 (514)	39.8 (389)	0.8 (0.7, 0.8 p<0.000)
No change	11.8 (115)	3.9 (38)	0.3 (0.2, 0.5 p<0.000)
Non-significant worsening	4.8 (47)	5.8 (57)	1.2 (0.8, 1.8 p<0.314)
Significant worsening	0.2 (2)	0.7 (6)	3.0 (0.6, 14.8 p<0.178)
Chi square (p)	273.81 (p<0.000)		

Significance of change (Sig Change) = Discharge location – Admission location / Standard error of the difference (SEdiff)
- = Calculation not possible
*RR, relative risk, is a ratio of the probability of an event occurring in one group versus another group (eg RR=pA/pB, where p=probability of event, A = Sample A; B = Sample B). In terms of comparing the statistical significance of significance of change comparing the FIMm and BI, RR = (FIMm p/N) / (BI p/N), where p=proportion, N=total sample. Also shown are associated lower and upper bound confidence intervals and p-values. The RR highlights the statistically significant difference between the BI and FIMm detecting individual-level significant improvement post-rehabilitation in this sample.

Individual person-level comparison of BI and FIMm relative responsiveness

Tables 2a and 2b show that the FIMm detected significant improvements in activity limitation in nearly 200 more people than the BI (50%, n=486 versus 31%, n = 298), and also recorded less people as unchanged on discharge (FIM= 4% n=38; BI= 12% n=115). Importantly, these analyses cannot be undertaken legitimately on ordinal rating scale data.

DISCUSSION

The aim of this study was to explore the consistent [9, 23-26] but counter-intuitive finding that the BI and FIMm are equally able to detect change in activity limitation; counter-intuitive because every FIMm item has 7 response categories, whereas corresponding BI items have between two and four categories. As such, changes in activity limitation ***should*** be more easily detected by FIMm items than BI items. This greater ***capacity*** of the FIMm to detect change should result in superior responsiveness.

This study had three major findings. The first is that the FIMm was more responsive than the BI. It detected significant improvements in very many more people (n=486 v 298), and detected

change in 67% of those considered unchanged by the BI. However, this clear demonstration of the FIMm's superiority was only possible through ***individual person-level*** analyses. These are only legitimately achieved using sophisticated methods, such as Rasch analysis.[3, 10, 11, 16]

The explanation for the different responsiveness of the FIMm and BI can be seen by plotting the standard error of measurement (y-axis) for every level of activity limitation defined by the FIMm and BI (x-axis; see Figure). At every activity limitation level, the standard error associated with a FIMm measurement is smaller than the standard error associated with the corresponding BI measurement. This is mainly because the FIMm has more item response categories. As a consequence, measurements made by the FIMm have narrower confidence intervals than those made by the BI. Thus, statistical significance is achieved with smaller changes in the FIMm than the BI.

The second important finding from this study is that the group-level indicators of responsiveness (ES, SRM, RE, RP) did not detect the superiority of the FIMm, even when the analyses were conducted on interval measurements derived from the BI and FIMm. This finding provides further support for our suggestion [9] that standard group-level indicators of rating scale responsiveness are limited and may be positively misleading.

The third important finding of this study concerns the similarity of measurements generated by the BI and FIMm. One feature of Rasch analysis is that it enables rating scales measuring the same construct to be equated on an identical metric. A close look at the results in Table 2 shows three things: on admission, the mean FIMm location is ***higher*** than the mean BI location; at discharge, the mean FIMm location is ***lower*** than the mean BI location; and the mean change measured by the FIMm (0.915 logits) is ***less*** than that measured by the BI (1.238 logits).*

These three findings raise two questions. Why do the FIMm and BI produce different measurements of the same people on admission and discharge? Why does the FIMm register ***less*** mean change than the BI given that it has the greater capacity to detect change? There are a number

* These inferences are legitimate because Rasch analysis enables scales measuring the same construct to be equated on a common interval level metric.

of possible explanations. First, these could occur if the FIMm and BI measured somewhat different constructs. This is unlikely as the FIMm was developed, in part, to improve on the limitations of the BI, [7] all items are common, and Rasch analysis supports them as measures of the same construct.

A second explanation is that inherent psychometric limitations in each scale account for the findings. This is possible as Rasch analysis identifies limitations in both scales (misfitting items, disordered thresholds). A summary of the results of the co-calibrated data analysis (essentially the 10 item BI and 13 item FIMm analysed as if they were a single 23 item scale) are shown in the supplementary Appendix. This table shows that 10 items have disordered thresholds, most items have statistically significant misfit (examination of the item characteristic curves confirmed this misfit, revealing over and under discrimination for the items with highest negative and positive fit, respectively), 14 items demonstrate statistically significant DIF (a combination of items with uniform and non-uniform DIF). When taken together the items with most concerning psychometric properties were the BI and FIMm Bowels and Bladder, Stairs, and FIMm Feeding.

Thus, at face value, the requirements of the Rasch model are not well met by the co-calibrated data, which reflect and build on the findings of others [27-29] who have demonstrated a range of psychometric problems including misfit for the BI and FIMm (largely because of the mixing of clinically different constructs, eg activities, mobility, sphincter function), and DIF for the FIMm. However, we explored the impact of the psychometric problems by modifying the data (albeit post hoc) to overcome the weaknesses (focusing specifically on items with disordered thresholds and exclusion of the items with poorest fit), and repeating our analyses of relative responsiveness. The same conclusion was reached; that effect sizes appear to be misleading when seeking to understand the relative ability of scales to detect change (data available on request).

A third explanation is that the results were biased by the therapists who rated the patients. It is conceivable that the clinically crude response categories of the BI might encourage therapists to overestimate people's activity limitation on admission and their activity limitation change at discharge. Our data do not allow us to investigate this further. Another explanation is that our

findings reflect the different precisions of the two scales. If this is the case, measures of the same construct, but with different precisions, may come to different conclusions about change associated with an intervention. This warrants further interrogation not possible within our BI and FIMm data.

Although the psychometric concerns outlined above are important considerations, we believe our examination of the two scales in this study bring to the fore some key clinical issues. We had the rare opportunity to directly compare two firmly established, widely used, highly clinically related instruments, one of which (the FIMm) was developed to improve upon the perceived insensitivity of the other (BI). Clinical experience suggests that the FIMm is more responsive than the BI. Thus, we would expect our study to find the FIMm better able to measure change. However, inferences based on the widely used traditional responsiveness indicators would lead us to believe the FIMm and BI are equally responsive. What we hope we have achieved here is that we have shown that using the more sophisticated analysis techniques (afforded by Rasch measurement methods) indicated that the FIMm is indeed more sensitive to change than the BI. This is in line with clinical expectation, and has important ramifications for the use of the tools in clinical research and trials, and the methods we use to determine and compare scale responsiveness.

At present we do not have a full explanation for our findings. However, from a clinical perspective we would expect the FIMm to be more sensitive to change than the BI. So, we believe that our inference that group-based statistics are misleading has credence. Despite this, at the current time, we cannot square the circle of the issues identified by this study. There is a clear need for further work using scales that better fit the Rasch model requirements, to elaborate upon what we have uncovered here and to ultimately pin down its root cause

Rasch measurement is not the only psychometric method available to analysing change in individual person level data. The other main new method is called Item Response Theory (IRT), [30] which in contrast to Rasch measurement, takes into account other sample related parameters such as item discrimination. Despite being mathematically similar, Rasch measurement and IRT have different research agendas. [18] In essence, albeit a simplification, IRT models are statistical

models used to explain data. When the observed data do not fit the chosen IRT model another model is sought to better explain the data (ie taking into account other sample dependent parameters as described above). In contrast, Rasch analysis provides a mathematical model for guiding the construction of stable linear measures from rating scale data.

The aim of a Rasch measurement analysis is to determine the extent to which observed rating scale data satisfy (fit) the measurement model.[*] This is vital for measuring change as the most important measurement *axiom* is the ability to test for invariance (stability). [11] This is achievable with Rasch Measurement, but not with IRT models as the presence of other *parameters* renders the estimates sample dependent. [3, 16, 18] It follows that Rasch measurement enabled us to obtain interval level activity limitation measurements to be estimated from ordinal BI and FIMm scores, legitimately examinechange at the individual person-level rather than just the group comparison level, and direct comparison of the BI and FIMm on the same activity limitation metric. We chose Rasch measurement rather than IRT for these very specific reasons.

We feel that it is vital that neurologists are aware of the key issues surrounding the use and analysis of rating scale data because rating scales have an increasingly crucial role in the determination of patient care, the guidance of clinical research directions, the evaluation of advances in basic science, and the evaluation of clinician professionalism.[31,32] Each of these eventually impacts on patients, clinical practice, and clinicians.

One limitation of this study is that responsiveness of the BI and FIMm was evaluated in a sample of neuro-rehabilitation patients, which included a large sub-group of people with MS, from one tertiary referral hospital in the south-east region of the UK. Importantly, when we analysed the main clinical sub-groups within our sample, the findings remained the same (data available from authors). However, to examine generalisability, it is important that others seek to replicate our analyses.

[*] In contrast the aim of an IRT analysis is to determine the extent to which the measurement models fit the rating scale data. This fundamental different is poorly appreciated.

Our findings have three important implications for clinicians, clinical practice and clinical trials. First, they demonstrate that group based statistics can be misleading, not of their own volition, when representing the ability, and relative ability, of scales to detect change. As such, they demonstrate the added value of using Rasch analysis and indicate that group based analyses should be complemented by legitimate analyses at the individual person level. The second implication, a consequence of the first, is that clinical investigators need to become familiar with, and apply, modern psychometric methods that enable legitimate comparisons at the individual person level. Traditional psychometric analyses, using raw scores, are not suitable for that purpose. Fourth, although Rasch analysis does not confirm clinical change, it helps to take us further than existing approaches, because the information provides us with a firm quantitative base upon which qualitative explorations of the differences between those people who report change and those who do not. We believe it is these sorts of explorations can move us towards a better understanding the nuances of what constitutes clinical change. When considered together, the findings demonstrate the added value that Rasch analysis brings to examining and understanding measuring change in activity limitation.

ACKNOWLEDGEMENTS

We thank Professor David Andrich (University of Western Australia, Perth, Western Australia) and Dr Barry Sheridan (RUMM Laboratory, Perth, Western Australia) for their contributions towards this work, and staff at the Neurorehabilitation Unit, National Hospital for Neurology and Neurosurgery London UK who routinely collect audit data.

COMPETING INTERESTS

There are no competing interests.

FUNDING

None

COPYRIGHT LICENCE STATEMENT

REFERENCES

1. United States Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labeling claims, draft guidance 2006. www.fda.gov/cber/gdlns/ prolbl.pdf.
2. Hobart JC, Cano SJ. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. Health Technology Assessment Monograph 2009; 13:12 1-200.
3. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007; 7: 1094-105.
4. Cano SJ, Hobart JC, Edwards, M., et al. CDIP-58 can measure the impact of botulinum toxin treatment in cervical dystonia. Neurology 2006; 67: 2230-32.
5. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with other self-report scales. J Neurol Neurosurg Psychiatr 2005; 76: 1539-43
6. Mahoney FI, Barthel DW. Functional evaluation: the Barthel index. Md Med J 1965; 16: 61-5.

7. Granger CV, Hamilton BB, Keith RA, Zielezny M, Sherwin FS. Advances in functional assessment for medical rehabilitation. Top Geriatr Rehabil 1986; 1: 59-74.
8. O'Connor RJ, Cano SJ, Thompson AJ, Hobart JC. Exploring rating scale responsiveness: Does the total score reflect the sum of its parts? Neurology 2004; 62:1842-44.
9. Cano SJ, O'Connor RJ, Thompson AJ, Hobart JC. Exploring disability rating scale responsiveness II: Do more response options help? Neurology 2006; 67: 2056-2059.
10. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen Chicago: Danish Institute for Education Research, 1960.
11. Andrich D. Rasch models for measurement. Sage Publications, Newbury Park, 1988
12. Wright BD. Solving measurement problems with the Rasch model. J Educ Measure 1977; 14: 97-116.
13. Wright BD, Linacre JM. Observations are always ordinal: measurements, however must be interval. Arch Phys Med Rehabil 1989; 70: 857-60.
14. McHorney C, Tarlov A. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995; 4: 293-307.
15. Wright BD, Stone MH. Best test design: Rasch measurement. Chicago: MESA, 1979.
16. Massof R. The measurement of vision disability. Optom Vis Sci 2002; 79: 516-52.
17. Hobart JC. Rating scales for neurologists. J Neurol Neurosurg Psychiatr 2003; 74: iv22-iv26.
18. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004; 42: I7-I16.
19. RUMM 2020 [program]. 4.0 version. Perth: RUMM Laboratory Pty Ltd, 2006. (www.rummlab.com)
20. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989; 27: S178-89.
21. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990; 28: 632-38.

22. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28: 542-47.

23. Van der Putten, J.J.M.F., Hobart, J.C., Freeman, J.A., Thompson, A.J. Measuring change in rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. J Neurol Neurosurg Psychiatr, 1999; 66: 480-84.

24. Wallace, D., Duncan, P.W., Lai, S.M. Comparison of the responsiveness of the Barthel Index and the Motor Component of the Functional Independence Measure in stroke: The impact of using different methods for measuring responsiveness. J Clin Epidemiol 2002; 55: 922-28

25. Hsueh, I.P., Lin, J.H., Jeng, J.S., Hsieh, C.L. Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. Neurol Practice 2002; 73: 188-90.

26. Houlden H, Edwards M, McNeil J, Greenwood R. Use of the Barthel Index and the Functional Independence Measure during inpatient rehabilitation after single brain injury. Clin Rehabil 2006; 20: 153-59.

27. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation* 1994;**75**(2)**:**127-132.

28. Tennant A, Geddes J, Chamberlain M. The Barthel Index: an ordinal score or interval measure. *Clinical Rehabilitation* 1996;**10:**301-308.

29. Lawton G, Lundgren-Nilsson A, Biering-Srensen F, et al. Cross-cultural validity of FIM in spinal cord injury. *Spinal Cord* 2006;**44:**746–752

30. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

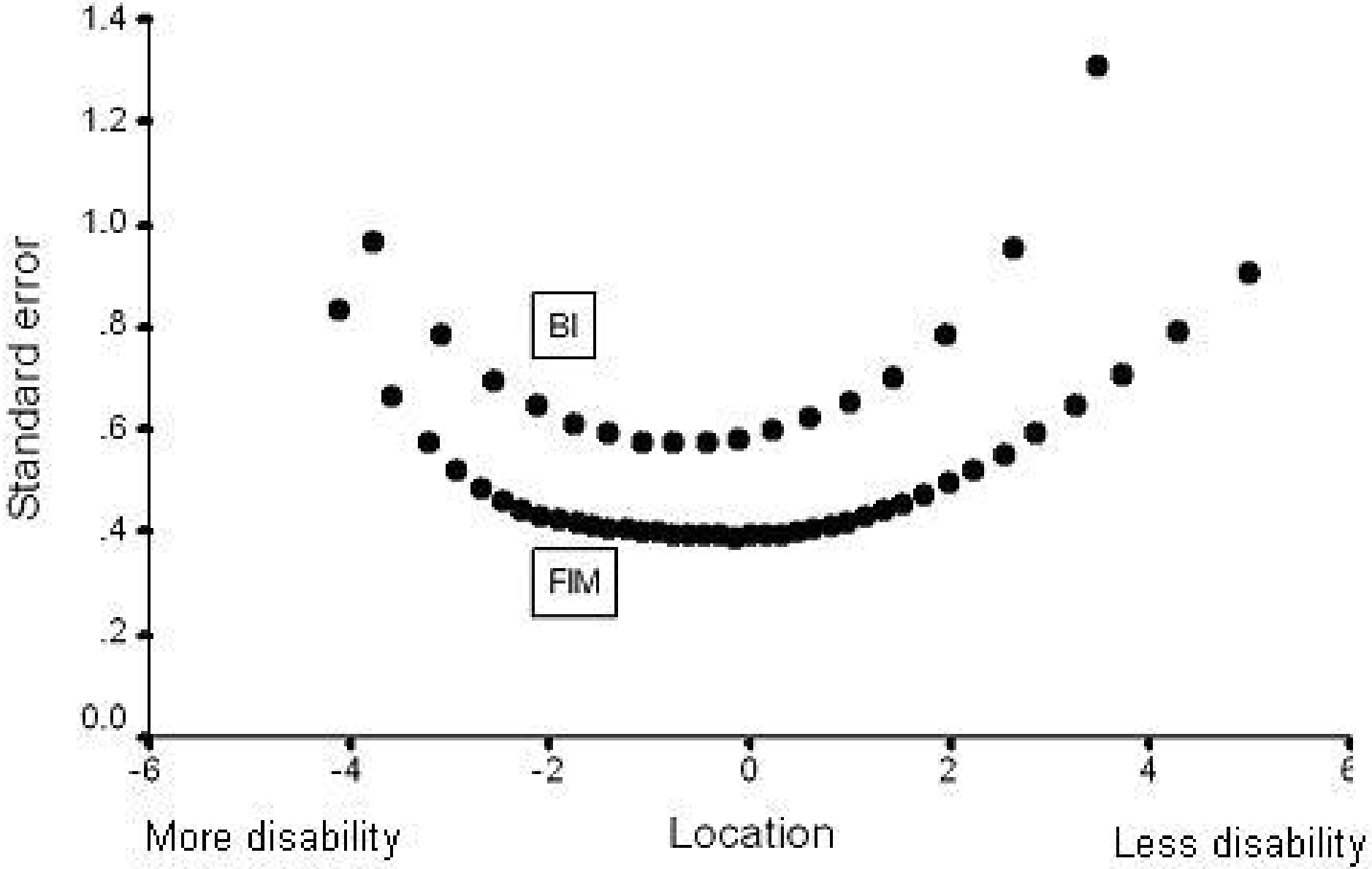
31. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ 2009;43:757-66.

32. Roland M, Elliott M, Lyratzopoulos G, et al. Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England. BMJ 2009;339:b3851.

Figure: BI vs FIMm – Comparison of standard errors across locations

Figure Legend:

This Figure shows that the standard errors for all FIMm measurements (person locations) are lower than for all BI locations on the activity limitation continuum. Thus, for a location of '0', the standard error for the FIMm is 0.4, and the BI is 0.6. This equates to 95% confidence intervals of 0.8 for the FIMm and 1.2 for the BI, showing how statistical significance is achieved with smaller changes in the FIMm


1.4
1.2
1.0
.8
.6
.4
.2
0.0
Standard error
BI
FIM
-6
-4
-2
0
2
4
More disability
Location
Less disability