

Rasch Analysis of the Gross Motor Function Measure: Validating the Assumptions of the Rasch Model to Create an Interval-Level Measure

Lisa M. Avery, BEng, Dianne J. Russell, MSc, Parminder S. Raina, PhD, Stephen D. Walter, PhD, Peter L. Rosenbaum, MD, FRCP(C)

ABSTRACT: Avery LM, Russell DJ, Raina PS, Walter SD, Rosenbaum PL. Rasch analysis of the Gross Motor Function Measure: validating the assumptions of the Rasch model to create an interval-level measure. *Arch Phys Med Rehabil* 2003; 84:697-705.

Objectives: To describe the Rasch analysis of the Gross Motor Function Measure (GMFM-88) and to demonstrate how the assumptions of unidimensionality, sample-free measurement, and test-free measurement were validated to create an interval level measure.

Design: Cross-sectional and longitudinal (12-mo) data from a prospective study of motor development in children with cerebral palsy (CP) were used for the analysis.

Setting: Motor assessments were completed at 18 children's ambulatory rehabilitation centers in Ontario, Canada, by pediatric physical therapists trained in the use of the GMFM-88.

Participants: The first 537 of 682 children enrolled into a longitudinal study of motor development in children with CP. Children had a mean age of 6.43 ± 2.75 years (range, 11mo–12y) with varying types and severity of CP.

Interventions: Not applicable.

Main Outcome Measure: The GMFM-88.

Results: The Rasch analysis, in conjunction with clinical decisions, identified 66 items from the GMFM-88 that formed a unidimensional measure (GMFM-66). Assumptions of sample-free and test-free measurement were confirmed, and a user-friendly scoring program was developed.

Conclusions: The GMFM-66 is an interval-level measure of gross motor function for children with CP; it should improve the scoring, interpretation, and overall clinical and research utility over the original GMFM.

Key Words: Cerebral palsy; Developmental disabilities; Motor skills; Rehabilitation.

© 2003 by the American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation

HEALTH CARE PROVIDERS are being challenged to evaluate the efficacy of their programs. A move toward evidence-based practice has produced the need for clinically relevant, reliable, valid, and responsive outcome measures to evaluate the effects of treatment. One such measure used in childhood disability is the criterion-referenced Gross Motor Function Measure (GMFM).¹ The GMFM was designed and validated for children with cerebral palsy (CP) by using principles of classical test theory and is used widely as a clinical and research outcome measure. Although the GMFM has been useful to document gross motor function in a systematic way, a limitation of the measure is that the scoring (and thus interpretation) is based on ordinal-level data. The purpose of this study was to describe the Rasch analysis of the GMFM and to show how the assumptions of unidimensionality and sample-free and test-free measurement were validated to create an interval-level measure (GMFM-66) with improved interpretability of scores.

The original GMFM, now referred to as the GMFM-88, is composed of 88 items grouped into 5 functional dimensions: lying and rolling (17 items); sitting (20 items); crawling and kneeling (14 items); standing (13 items); and walking, running, and jumping (24 items). The items are arranged within dimensions by difficulty, using the literature and clinical judgment, but the ordering is done primarily for ease of administration. It was acknowledged that the original scale was not necessarily hierarchical, and therefore all items needed to be administered and scored for a child. Each item is scored on a 4-point rating scale from 0 to 3, with 0 indicating that the child cannot initiate the item and 3 indicating that the child can complete the item (as defined in the GMFM manual). Each of the scoring options within the 88 items is explicitly defined, in order to describe clearly the motor behavior to be observed and scored. If a therapist cannot elicit a response from the child or if an item is inadvertently missed, then the child receives a score of 0 for that item. Percentage scores for each dimension are summed and averaged to obtain a total GMFM-88 score. There is considerable evidence of the reliability, validity, and responsiveness of the GMFM-88 for children with CP.¹⁻³

Rasch analysis has been used to develop new measures⁴⁻⁶ and to test existing measures to see whether the assumptions of the Rasch model can be met.⁷ Rasch analysis is based on a probabilistic model that uses maximum likelihood estimation to order items and subjects simultaneously, thereby arranging the items along a difficulty continuum and subjects along an ability continuum.⁸ The interest in Rasch analysis by measurement developers is based on the ability of this method to aid in the construction of measures with interval-level measurement.

There are several advantages of interval (vs ordinal) measurement. First, placing items on a difficulty continuum reveals the hierarchical structure of the items and their relative difficulties and gives information about the underlying trait being examined. Second, parametric statistics can be computed from

From the *CanChild* Centre for Childhood Disability Research, Institute for Applied Health Sciences (Avery, Russell, Rosenbaum, Walter), and the Department of Clinical Epidemiology and Biostatistics (Raina, Walter), McMaster University, Hamilton, ON, Canada.

Supported in part by the National Institutes of Health (grant no. R01-HD 34947). A commercial party with a direct financial interest in the results of the research supporting this article has conferred or will confer a financial benefit upon the authors or one or more of the authors. The authors will receive royalties from the sale of a book published by Mac Keith Press that describes the development of the GMFM-66.

Reprint requests to Dianne J. Russell, MSc, CanChild Centre for Childhood Disability Research, Rm 408, Institute for Applied Health Sciences, McMaster University, 1400 Main St W, Hamilton, ON L8S 1C7, Canada, e-mail: russell@mcmaster.ca.

0003-9993/03/8405-7404\$30.00/0

doi:10.1016/S0003-9993(03)04896-7

the scores on an interval continuum, thereby making possible reliable and accurate comparisons of change among subjects, as well as changes over time for a given subject. In addition, a benefit of Rasch analysis is its ability to estimate a total score for a client even when not all items have been administered. This is particularly useful when working with children who may not comply with all the constraints of a standardized testing situation.

As with any model, the Rasch model sets out certain specifications, which must be met before the benefits of the model can be realized. The main specification is that the items are based on a single underlying trait (referred to as unidimensionality). Once this specification is met, it follows that sample-free and test-free measurements may be made. Sample-free calibration refers to the ability of the scale to produce stable item estimates regardless of the sample of subjects used to calibrate the items. Similarly, test-free measurement implies that a subject's expected score is independent of the items administered. A full review of the statistical aspects of Rasch models is beyond the scope of this article but is available elsewhere.^{9,10}

Our group was interested in examining whether the assumptions of the Rasch model could be met and used to improve the scoring and interpretability of the GMFM-88 for children with CP. Briefly, Rasch analysis was done (1) to identify those items in the GMFM-88 that contribute to a unidimensional measure of gross motor function; (2) to determine the hierarchical ordering of these items and to devise an interval scoring system based on the item calibrations; (3) to examine the assumptions of Rasch analysis as they pertain to the revised GMFM, including unidimensionality, sample-free measurement, and test-free measurement; and (4) to create a computerized data entry and scoring program. The psychometric properties of reliability, validity, and responsiveness of the Rasch version of the GMFM-88 (known as the GMFM-66) were assessed and are reported elsewhere.¹¹

METHODS

Subject Selection

Data from a large prospective study of motor development in children with CP using the GMFM were used. A detailed description of the sampling procedure is presented elsewhere.¹² In brief, 18 of the 19 children's treatment centers in Ontario, Canada, and 1 additional pediatric treatment facility participated in this study. The vast majority of children with CP in Ontario attend 1 of these 19 treatment centers. Sampling was random within centers, and the overall sample is thought to be representative of the population and hence generalizable. Subjects were randomly sampled from lists of eligible children, stratified by age and severity, by using the Gross Motor Function Classification System (GMFCS).¹³ The GMFCS is a system of severity classification of motor involvement that places children into 1 of 5 levels of function, with level I being the most functional and level V the most limited. Data from 537 children in the study were used for these analyses (table 1).

All children were assessed at baseline with the GMFM-88. In addition to the traditional scoring, a reported score was also collected from parents.¹⁴ A subsample of 228 children with repeat GMFM-88 assessments 12 months after baseline was used to examine change over time. A second subsample of 115 children with repeat measurements, who were randomly selected and stratified by GMFCS level, was used to compare the ability among different measures to detect change over time. All children had a diagnosis of CP. The mean age was 6.43 ± 2.75 years (range, 11mo–12y) at the time of initial

Table 1: Demographics of the 537 Children Used in the Rasch Analysis of the GMFM

Variable	Frequency	%		
GMFCS level				
I	155	28.9		
II	70	13.0		
III	104	19.4		
IV	105	19.6		
V	103	19.2		
Type of CP				
Spastic	411	76.5		
Dystonic/athetotic	32	6.0		
Ataxic	14	2.6		
Low tone hypotonic	27	5.0		
Mixed	53	9.9		
Distribution of CP				
Leg dominant	183	34.1		
Three-limb dominant	53	9.9		
Four-limb dominant	215	40.0		
Right hemiplegia	43	8.0		
Left hemiplegia	42	7.8		
Missing	1	0.2		
Gender				
Male	299	55.7		
Female	238	44.3		
	Min	Max	Mean	SD
Age (y)	11mo	12	6.43	2.75

Abbreviations: Min, minimum; Max, maximum; SD, standard deviation.

GMFM-88 assessment, with slightly more than half of the sample being boys (55.7%).

All 102 therapists administering the GMFM-88 were trained in its administration and scoring and were tested to ensure they reached an acceptable level of agreement with an expert-scored videotape.¹⁵

Model Selection

There are numerous Rasch models, each with different assumptions about the scaling of response options. Figure 1 illustrates the decision process that was used to decide the appropriate model. The partial credit model was chosen for 2 reasons. First, this model assumes nothing about the relative difficulty of response options or steps within items. Second, the more item parameters a model has, the greater the sample size needed, and the sample size available was thought to be large enough to obtain reliable estimates of the item step difficulties.

Analysis

Rasch analysis is an iterative process, and decisions are made based on a combination of statistical and clinical considerations. The first important step in the analysis after selecting the appropriate model was to identify a unidimensional set of items that would (1) meet the assumptions of the Rasch model, (2) retain the ability of the measure to detect change over time, (3) span the spectrum of ability commonly seen in children with CP, and (4) be judged by therapists as clinically meaningful. Once the appropriate items were identified, it was important to demonstrate that both sample-free and test-free measurement had been attained. Finally, when all the assump-

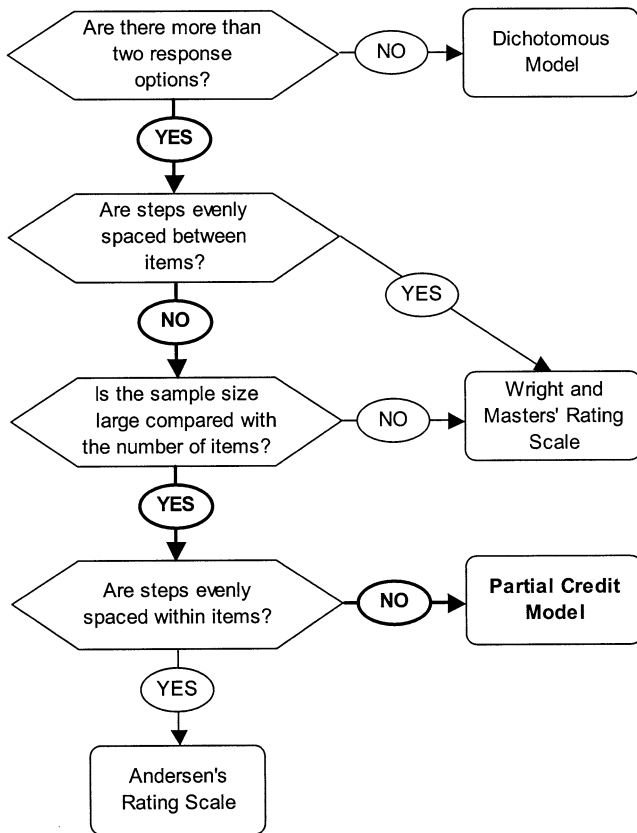


Fig 1. The decision process used to select the most appropriate Rasch model for the GMFM. Note that all models are discussed elsewhere.¹⁰

tions of the Rasch model had been met, a method of scoring the new measure was needed.

Rasch analysis of the data set was completed using the Bigsteps program, version 2.71.^a Among other things, this analysis provided 2 estimates of item difficulty (item step measures, Thurstone thresholds), the standardized infit statistic for each item, and the person separation statistic. Rasch analysis uses logits to express item difficulty and person ability.¹⁶ A linear transformation was used to convert the person abilities from the original logit scale (theoretical range, $-\infty$ to $+\infty$) to a scale ranging from 0 to 100. The same transformation was also applied to the item and step difficulty estimates.

Step 1: defining a unidimensional set of items. Once the partial credit model had been chosen, it was necessary to determine which of the items comprised a unidimensional measure of gross motor function. The criterion for assessing unidimensionality was that less than 5% of items could misfit the Rasch model.^{17,18} The fit of individual items was examined by using goodness-of-fit (GIF) statistics and principal component analysis. The GIF statistics were used as the primary indicator of item fit because of the prevalence of this approach in the literature. The weighted infit statistic (which describes how well an item assesses the gross motor function of children whose abilities are near to the difficulty of the item) was used to assess fit because it was felt to be more important than the weighted outfit statistic (which describes how well an item assesses the gross motor function of children whose abilities

are farther away from the difficulty of the item, for example, those children for whom the item was likely to be very hard or very easy). The standardized statistic was chosen over the mean square in light of the work done by Smith et al,¹⁹ who concluded that the mean square is less sensitive to large sample sizes.

The choice of an appropriate critical value for determining misfit with the standardized infit statistic was somewhat complicated and required both analytic and clinical considerations. The goal of the analysis was to identify, from the original 88 GMFM items, a subset of items that was unidimensional in its ability to measure gross motor function. Furthermore, the items should be able to place children along an ability continuum and be at least as responsive to change as the original 88-item test. Bearing this in mind, it was decided to remove items that did not fit the Rasch model but to keep possibly redundant items (as identified by negative GIF statistics) so to retain as much clinically useful information as possible. Because of the relatively large sample size and test length and because the infit values are sensitive to the sample being used, it was decided to test 2 critical values. A choice between the 2 values would then be made based on the ability of the resulting measures to meet the above requirements. The standard critical value of 2.0 was examined, as well as a larger critical value of 3.0.

For each of the critical values, a Rasch analysis with the Bigsteps program^a was done, starting with all 88 items. Items with infit values above the critical value were progressively removed until the total number of misfitting items was less than 5% (this required repeated analysis with the Bigsteps program). This process resulted in 2 groups of items: the items that were judged to fit the model with a critical value of 2.0 and the items that were judged to fit with a critical value of 3.0. Once these 2 groups of items had been identified as possible measures, a final decision between the 2 measures was made. This decision was based on each measure's ability to detect change over time, the capacity of the items to spread the subjects out along the ability continuum, and clinical judgment as to the usefulness of the identified items to the measure as a whole. In other words, principles of classical test theory were blended with the techniques of Rasch analysis to ensure clinical sensitivity as well as statistical soundness.

The ability of each measure (the 2 unidimensional measures and the original 88-item GMFM) to detect change over time was examined by assessing the change in subjects' scores with a repeated-measures analysis of variance (ANOVA). A random sample of cases from the original 537 with repeat assessments, stratified by GMFCS level, was used so as not to bias the choice of threshold toward any 1 functional level. The sample size for this analysis was limited by the least populated GMFCS level. In all, 23 subjects in each of the 5 levels were selected for the analysis (total, 115 children). To assess the ability of the items to spread the sample along the ability continuum, the person separation statistic (a measure of the variation in the scores) was used.¹⁶ Four clinicians experienced with the GMFM also judged the groups of items and gave input as to the clinical implications of removing various items.

Once the measure was determined, a principal component analysis was performed on the standardized score residuals to confirm that there were no signs of obvious multidimensionality within the new measure.

Step 2: testing for sample-free measurement. Sample-free measurement can be evaluated by assessing the stability or reliability of items as calibrated by different samples. A number of tests were done to investigate the stability of the item and step difficulty estimates. It should be noted that because there

is no assumption about the position of step difficulties across items, the partial credit model requires that the estimates of each of the step difficulties (as opposed to merely the item difficulties) be compared. Reliability was assessed by (1) testing the stability of the item difficulties over time, between baseline and after 12 months ($n=228$); (2) testing the item step difficulties between 2 randomly selected, independent samples ($n=536$ or 268/group); and (3) testing the item step difficulties between distinct groups of children with known differences in ability ($n=536$ or 268/group). All agreements between step estimates were measured by using an intraclass correlation coefficient (ICC), based on a 1-way ANOVA.²⁰

To select 2 distinct groups with known differences in ability, it was decided that the target mean of the lower-ability group would be 40, roughly corresponding to the mean ability for children in GMFCS level IV, and the target mean of the higher ability group would be 60, approximately the mean ability of children in GMFCS level II. The target standard deviation (SD) of both groups was 15. A modified version of the rejection method was used to select the groups from the available data.²¹ The actual means \pm SDs of the 2 groups were $\bar{x}_1=38.1\pm 18.9$ and $\bar{x}_2=64.9\pm 18.7$, respectively.

Step 3: testing for test-free measurement. Test-free measurement means that a subject's expected score is independent of which items have been administered. Two different methods were used to examine the assumption of test-free measurement in the GMFM-66: (1) using true data from the children in our study and (2) using simulated response patterns of hypothetical children that fit the Rasch model. The first exercise used data from all 537 children used in the Rasch analysis. The GMFM-66 items were ordered according to difficulty and then alternately assigned to 1 of 2 subgroups. For each subject, 2 ability estimates were computed, 1 from each subgroup of items. The agreement between the 2 ability estimates was analyzed with an ICC.

For the second exercise, ability levels for 100 children were simulated. Item responses that "fit" the Rasch model for children with CP were then simulated from these seed abilities, according to methods proposed by Smith.²² A random selection of items was drawn, and the ability of each of the 100 subjects was computed based on these items alone. The number of items selected was fixed for each simulation, and all options (from 1 item to 66 items) were examined. The ability computed with a random subset of items was then compared with the "true" ability (the seed ability used to simulate the response strings) with an ICC. One hundred simulations were used to calculate the mean agreement between the true and estimated ability for each possible number of items.

Step 4: scoring the GMFM. Once the unidimensionality of the selected items was verified, those items were ordered according to difficulty and became the GMFM-66. This interval measure allowed the determination of the gross motor ability of the children in the sample. A scoring algorithm was developed that allows for missing data and assigns the best-fitting estimate of ability to each subject (fig 2).

RESULTS

Defining a Unidimensional Set of Items

The choice between critical values for the infit statistic was made based on the properties of the items selected with each threshold. A critical value of 2.0 resulted in a group of 52 items, and a critical value of 3.0 resulted in a group of 68 items. The properties of these 2 groups, plus those of the original 88 items, are summarized in table 2. The result of the repeated-

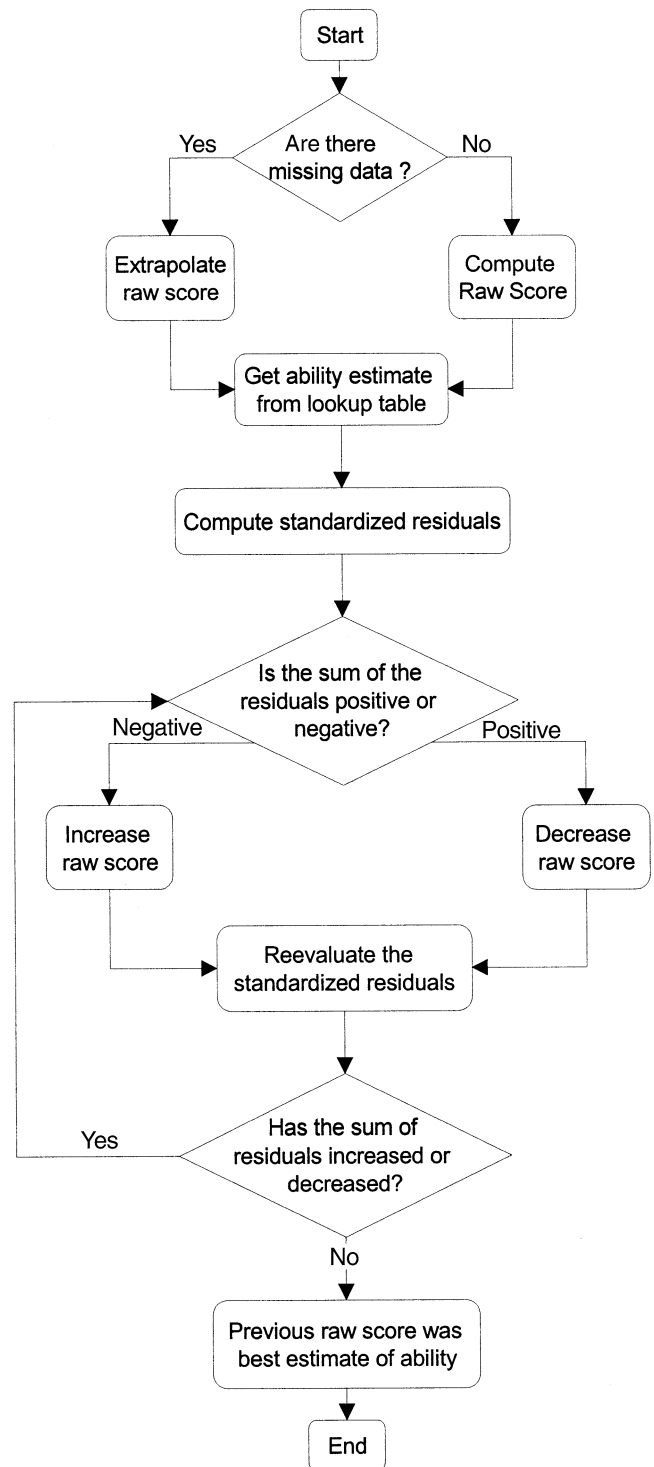


Fig 2. Illustration of the algorithm used to score the GMFM-66.

measures ANOVA was not statistically significant for the effect of item group, but the mean change score over 12 months was largest for the group of 68 items. In addition, using a critical value of 3.0 retained some of the items in the lying and rolling dimension, which was considered clinically important.

Table 2: Comparison of the Properties of Different Groups of GMFM Items

Variable	No. of Items Retained	Mean Change Score (logits)	Person Separation (calibration error units)
All items	88	.257	9.47
Critical value=2.0	52	.175	8.55
Critical value=3.0	68	.285	10.19

In comparison, a critical value of 2.0 resulted in the elimination of the entire lying and rolling dimension. The person separation statistic (a measure of the ability of items to discriminate children) was also greatest for the group of 68 items (table 2). Based on these results, the items identified with the critical value of 3.0 were candidates for the new measure. Closer examination of the group of 68 items revealed 2 unilateral items whose counterparts had been removed. Items 8 (a right-sided item) and 50 (a left-sided item) were removed due to high infit values. We removed the contralateral items (which also had high infit values) to ensure that the scoring system would be equally responsive for children with hemiplegia. The result was a group of 66 items that was unidimensional with respect

to gross motor function. This measure is now referred to as the GMFM-66.

The principal component analysis of the standardized residuals revealed that only 3 items had factor loadings greater than |0.5| and that there was no obvious clustering of items. This was accepted as sufficient evidence that the measure is unidimensional.

The step measures and the standardized infit statistic for each the items of the GMFM-66 are found in table 3.

Sample-Free Measurement

Table 4 presents the results of tests of the reliability of item difficulty. All correlations were greater than .96, indicating that

Table 3: Item Step Measures and Standardized Infit Values for the Items in the GMFM-66

Item*	Step Measures			Infit Value	Item	Step Measures			Infit Value
	1	2	3			1	2	3	
2	20.19	26.07	22.31	2.9	56	56.09	55.44	52.15	-1.8
6	21.31	22.54	30.08	0.9	57	61.45	78.05	84.93	-2.0
7	23.72	21.95	27.90	3.0	58	61.92	78.28	83.64	-0.9
10	14.30	15.89	21.48	1.7	59	46.67	55.68	53.91	-3.2
18	21.84	21.66	29.49	1.8	60	57.15	56.33	69.63	0.3
21	14.54	13.89	10.77	-0.3	61	56.68	56.50	71.51	1.0
22	9.48	22.01	22.95	2.3	62	55.80	51.44	64.98	-0.8
23	27.49	21.90	19.89	2.2	63	57.92	51.97	64.51	-0.6
24	31.96	33.67	24.66	0.5	64	59.80	51.21	54.15	-2.6
25	32.14	31.37	38.08	-1.1	65	47.50	47.62	41.32	-2.9
26	36.90	42.44	33.55	2.4	66	47.62	48.20	40.79	-2.9
27	36.14	45.26	29.84	-0.2	67	44.32	43.97	33.78	3.4
30	38.61	37.73	37.67	2.0	68	50.56	53.21	43.73	0.0
31	47.73	48.03	36.79	0.6	69	67.10	58.09	41.20	-1.6
32	47.56	48.85	38.43	2.8	70	71.39	58.92	41.91	-1.8
34	36.49	37.20	35.90	2.7	71	69.69	59.15	55.03	-2.6
35	52.62	49.03	41.14	-1.4	72	74.63	54.50	43.97	-1.8
36	43.14	48.79	43.03	-3.6	73	64.74	64.16	69.57	0.4
37	46.56	46.32	50.62	0.6	74	67.39	73.16	78.52	-1.6
39	40.20	41.02	35.14	-0.2	75	69.98	65.39	66.51	-3.3
40	50.85	38.73	39.91	-1.8	76	69.69	66.63	65.21	-2.5
41	41.08	43.08	34.08	-2.5	77	65.27	72.10	57.80	0.6
42	51.15	39.55	42.32	1.3	78	72.98	61.39	44.56	-0.9
43	49.56	41.85	42.55	1.7	79	73.28	58.98	48.20	-0.5
44	51.85	47.44	28.08	-3.6	80	68.98	67.39	87.93	-3.9
45	49.91	51.56	38.20	-0.5	81	71.63	65.80	70.98	-2.0
46	54.15	49.21	38.61	0.9	82	79.34	88.82	83.76	2.0
48	45.03	44.08	47.20	-0.3	83	80.11	87.58	83.58	0.7
51	52.97	52.68	53.44	3.4	84	61.15	67.16	59.98	-1.2
52	43.91	50.15	35.31	-2.6	85	60.51	70.75	68.45	0.1
53	40.85	48.09	51.91	-1.1	86	75.69	74.10	67.39	-2.4
54	47.50	51.27	53.33	1.8	87	75.22	79.52	77.10	-2.9
55	49.68	49.91	53.33	1.7	88	77.99	73.69	58.39	-0.5

* Item numbers refer to the original GMFM-88 numbering.

Table 4: Results of Sample-Free Measurement Tests of Reliability for Item Step Difficulty Estimates for the GMFM-66 With 3 Different Samples

Description of Sample	ICC	n	No. of Steps*
Single sample with 2 assessments 12mo apart	.966	228	193
Distinct, randomly selected samples	.976	115	198
Distinct, randomly selected samples, biased with respect to ability	.975	536 [†]	190

*The number of steps in the GMFM-66 is (66 items×3 response options [steps] per item=198 steps). The reduced number of steps in the first and third analyses is due to the fact that not all response options were used in all samples.

[†] One child was excluded to ensure equal sample sizes in the 2 groups.

the difficulty estimates for the GMFM-66 are stable over time (ICC=.966, n=228), between random samples (ICC=.976, n=115), and between samples intentionally biased with respect to ability (ICC=.975, n=536).

Test-Free Measurement

The correlation between the child ability estimates by the 2 item subgroups for our sample was high (ICC=.98, n=537).

From the simulation study, we found that the greater the number of items tested, the more accurately the true ability could be estimated (fig 3). Furthermore, it was shown that, on average, testing 13 items was sufficient to estimate accurately the true ability of the subject (ICC>.95).

Scoring Program

A user-friendly computerized scoring program²³ was developed to allow clinicians and researchers to enter GMFM-66 scores either individually or as an ASCII text file. For individual data entry, the program calculates a child's GMFM-66 score and plots it on an item map along with the 95% confidence intervals (CIs) around the score (fig 4). The program also allows users to track a child's progress over time.

DISCUSSION

The GMFM-88 is a widely used clinical and research instrument developed with principles of classical test theory. The decision to apply Rasch analysis to the GMFM-88 was made to determine whether we could improve the scoring and interpretability of this already established measure without sacrificing its reliability, validity, and responsiveness to change over time. Unlike new measures created with this method, working with

an existing health outcome measure posed some additional challenges. We acquired the technical resources that could do the analytic work; however, we found that there was less information available to guide us through either the statistical or clinical decisions we faced.

Decisions Regarding Model Selection

The use of the partial credit model was based on our knowledge of the GMFM and the belief that the response options were not equal between or within items, rendering the rating scale models inappropriate. The tests for sample-free measurement confirmed that the choice of this model was justified. The high agreement between the step difficulty estimates, when compared between distinct groups, indicates that there were sufficient data to estimate the parameters of the partial credit model.

Defining a Unidimensional Set of Items

There are 2 decision points involved in determining the unidimensionality of a model: (1) deciding how many (if any) items may misfit and (2) deciding how to determine if individual items fit the Rasch model. Based on previous studies, we decided that a 5% misfit level was acceptable. The more difficult decision was the choice of GIF statistic and a critical threshold for determining when individual items misfit the model. It is difficult to assess these choices directly, because there are many possible combinations of fit statistics and critical values to evaluate. However, by examining the results and implications of these choices empirically, it was possible to justify these decisions and to make an indirect assessment of their suitability. In this study, repeated analyses were performed, wherein misfit items were removed until less than 5% of remaining items had infit values greater than the critical value. That the remaining group of items met the expectations of the Rasch model provides evidence that the choice was reasonable.

Decisions About Sample-Free and Test-Free Measurement

A search of the literature did not reveal any previous examination of the assumptions of sample-free or test-free measurement, so no a priori judgment was made as to what would constitute sufficiently high ICCs. However, because all of the ICCs were very high (>.96), it was assumed that both sample-free and test-free measurement had been attained. The strongest evidence to support the assumption of sample-free measurement comes from the high agreement of the step difficulties as estimated by 2 distinct groups of children with CP with different ability distributions. The assumption of test-free measurement was validated by both the simulation analysis and the test on the real data. Using real data from the children in our study, we illustrated that the ability estimates were similar

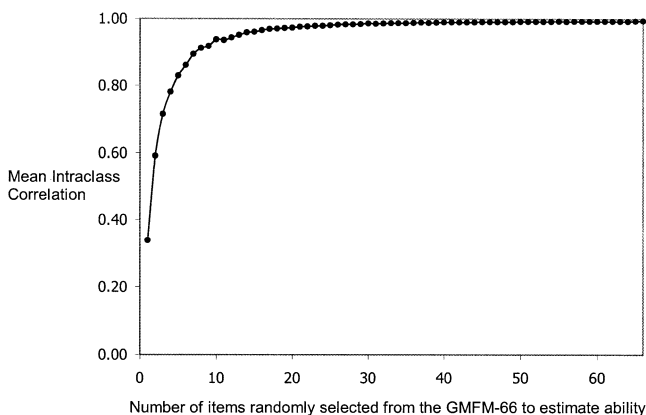


Fig 3. Illustration of the agreement between true and estimated ability over 100 simulations for varying numbers of items tested.

Item Map by Difficulty Order

Client ID: 12025
 Name: K E
 Assessment Date: 26 November 1997
 Date of Birth: 30 March 1993
 Age: 4y 7m

Gross Motor Function Measure
 GMFM-66

GMFM-66 Score: 61.21
 Standard Error: 1.29
 95% Confidence Interval: 58.68 to 63.74

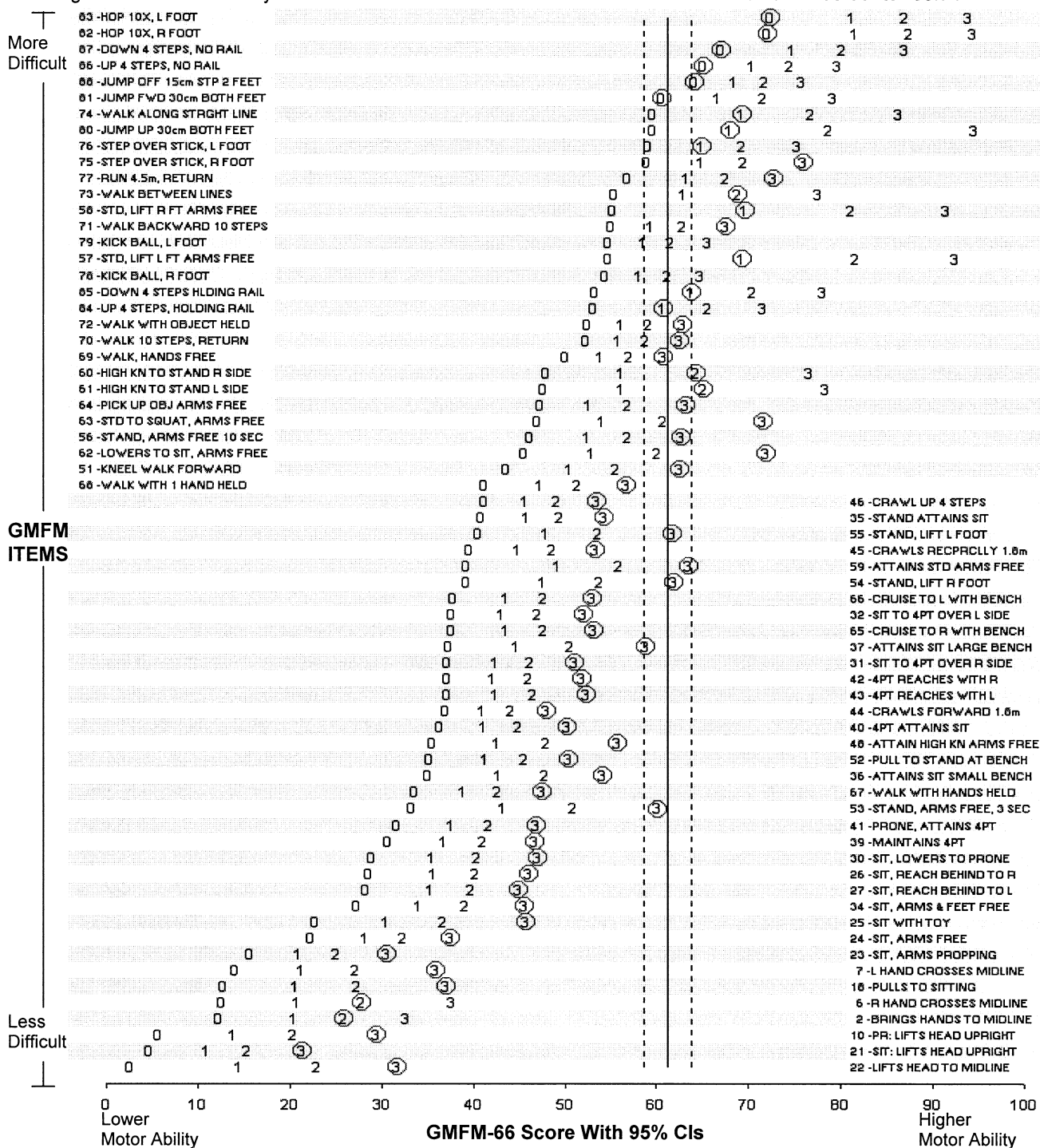


Fig 4. Example of an item map by difficulty order for the GMFM-66.

when estimated with completely different subsets of items. In the simulation exercise, we showed that testing a minimum of 13 items will accurately estimate the ability of a child. However, this was a simulation analysis, and all items were randomly selected. At present, no guidelines exist for choosing an appropriate subset of items with which to test subjects, however, it makes sense to test as many items as possible around a child's ability level and where there is variation in scores. Therefore, although the property of test-free measurement has been validated and the use of incomplete data sets is defensible, the testing of all items is still preferred.

Decisions About Scoring the GMFM-66

After a unidimensional group of items was obtained, a method of scoring subjects based on the responses to those items was needed. The requirements of the new scoring system were (1) that it retain the responsiveness of the original measure, (2) that it be reliable, (3) that it be on an interval scale, and (4) that it allow for missing responses.

One of the selection criteria for the items in the GMFM-66 was that the items maintain the responsiveness of the original measure. It was therefore assumed that any new scoring method based on these items would produce a measure that was at least as responsive and reliable as the original measure. An examination¹¹ of the responsiveness and reliability of the GMFM-66 has shown that indeed the GMFM-66 is a reliable and responsive measure.

One output of the Bigsteps program is estimates of person ability for each possible raw score in units of logits. Because the raw score is a sufficient statistic for the person ability, the program assigns all subjects with the same raw score to the same ability estimate. In contrast, the new scoring algorithm created to score the GMFM-66 uses all the item scores available for each child and assigns the value of gross motor ability that most closely fits the Rasch model by minimizing the standardized score residuals. This procedure is equivalent to determining the optimum raw score for each person, based on all the available responses, and then assigning the corresponding person ability, for which this optimum raw score is sufficient. There are advantages and disadvantages of this method. The scoring algorithm allows missing data on children to be assigned a person ability and reduces the effects of noncompliance in the population. However, the use of conversion tables to score children is not possible, and a computer program is needed.

It was decided that a computer program would be used to score the GMFM-66, to allow for missing responses and to give the best estimate of ability for each child. Missing responses occur when a therapist cannot elicit a response for an item, for instance, when a child capable of walking refuses to perform an activity in the prone position. The conventional GMFM-88 scoring guidelines require that children who refuse to perform an item be assigned a score of 0, though they may be assigned a reported score if a caregiver reports that the activity is part of the child's usual functional repertoire. If a conversion table were used to score the GMFM, noncompliant children who could perform some of the more difficult activities (walking, running, jumping) but refused to perform the easier tasks during the testing might be assigned a lower ability estimate than truly reflected their ability. For this study, the reported scores were used to supplement the observed scores (taking the higher of the 2), and the Rasch analysis was performed on these combined scores in an attempt to minimize the number of false-zero scores. There was some concern that the practice of recording scores of 0 for all activities not

observed would impact the results of the Rasch analysis. Although this concern cannot be definitively addressed, the high agreement of the item difficulties between different samples suggests that this scoring practice had negligible effects on the Rasch analysis for the sample used in this study.

The scoring algorithm was incorporated into the Gross Motor Ability Estimator (GMAE), a program that computes the best estimate of gross motor ability of a child with CP based on the scores on the GMFM-66.²³ This program reports an estimate of gross motor ability and the standard error of the score. In addition, a graphical approach has been used to examine the fit of the children to the Rasch model. A child's scores on the GMFM are circled on an item difficulty map (a map that illustrates the relationship between a child's ability and the difficulty of the GMFM-66 items) to give clinicians a sense of whether the child's behavior fits his/her estimated ability (fig 4). The use of a computerized scoring program was felt to be the most rigorous method of obtaining a total score—however, there were concerns about whether this would limit its clinical usefulness. Do clinical therapists generally have access to a computer, and, if they do, would the information gained from the program convince them to use it? Results from a pilot study²⁴ of over 50 therapists in Ontario suggest that they would be interested in using it.

Any linear transformation can be applied to person abilities and item difficulties to make the scales easier to interpret, as long as the same linear transformation is used for both domains. The GMAE uses a transformation that rescales the person abilities from an interval scale centered at 0 to a ratio scale that runs from 0 to 100. The decision to use such a transformation was based solely on clinicians' preferences. The item step measures have been rescaled with the same transformation and are reported in table 3.

CONCLUSION

The application of Rasch analysis is by no means a pure science. Because decisions must be made by using a balance between statistical and clinical reasoning, there is no single correct approach. It is therefore important to make decisions with care, considering a number of possible options before deciding on an appropriate course of action. A review of the procedures after their completion can further ensure the suitability of the decisions. Validation of the assumptions of the Rasch model, specifically those of sample-free and test-free measurement, is desirable to ensure that a unidimensional measure has in fact been achieved.

The adaptation of the new interval-level scoring system for the GMFM-66, for children with CP, is an improvement over the GMFM-88 percentage scores. Of the 88 original items, 66 have been found to contribute to a unidimensional group of items that measure gross motor function. A computer program, the GMAE, has been developed to compute reliable person ability estimates based on the responses to these 66 items. Because the assumption of test-free measurement has been validated, not all of these items need to be tested to estimate a child's gross motor ability; however, the more data available for a subject, the more accurate the estimate of gross motor function.

Acknowledgments: We thank the project coordinator, Barbara Galuppi, the assessing therapists, and all the children and families who participated in the Ontario Motor Growth Study.

References

1. Russell D, Rosenbaum P, Cadman D, Gowland C, Hardy S, Jarvis S. The Gross Motor Function Measure: a means to evaluate the

- effects of physical therapy. *Dev Med Child Neurol* 1989;31:341-52.
2. Bjornson KF, Graubert CS, Burford VL, McLaughlin JF. Validity of the Gross Motor Function Measure. *Pediatr Phys Ther* 1998; 10:43-7.
 3. Bjornson KF, Graubert CS, McLaughlin JF, Kerfeld CI, Clark EM. Test-retest reliability of the Gross Motor Function Measure in children with cerebral palsy. *Phys Occup Ther Pediatr* 1998; 18:51-61.
 4. Haley S, Coster W, Ludlow L, Haltiwanger JT, Andrellos PJ. Pediatric Evaluation of Disability Inventory (PEDI). Version 1.0. Boston: New England Medical Center Hospitals Inc; 1992.
 5. Campbell SK, Osten ET, Kolobe TA, Fisher AG. Development of the Test of Infant Motor Performance. In: Granger CV, Gresham GE, editors. *New developments in functional assessment; Physical Medicine and Rehabilitation Clinics of North America*. Philadelphia: WB Saunders; 1993. p 541-50.
 6. Coster W, Beeney T, Haltiwanger J, Haley SM. *School function assessment*. San Antonio (TX): Psychological Corp/Therapy Skill Builders; 1998.
 7. Haley SM, McHorney CA, Ware JR Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10). I: Unidimensionality and reproducibility of the Rasch scale. *J Clin Epidemiol* 1994;47: 671-84.
 8. Rasch G. *Probabilistic models for some intelligent and attainment tests*. Chicago: MESA Pr; 1980.
 9. Patrick DL, Chiang Y, editors. *Health outcomes methodology: symposium proceedings*. *Med Care* 2000;38(9 Suppl).
 10. van der Linden WJ, Hambleton RK, editors. *Handbook of modern item response theory*. New York: Springer; 1997.
 11. Russell D, Avery L, Rosenbaum P, Raina P, Walter S, Palisano R. Improved scaling of the Gross Motor Function Measure for children with cerebral palsy: evidence of reliability and validity. *Phys Ther* 2000;80:873-85.
 12. Palisano RJ, Hanna SE, Rosenbaum PL, et al. Validation of a model of gross motor function for children with cerebral palsy. *Phys Ther* 2000;80:974-85.
 13. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev Med Child Neurol* 1997;39:214-23.
 14. Russell D, Palisano R, Walter S, et al. Evaluating motor function in children with Down syndrome: validity of the GMFM. *Dev Med Child Neurol* 1998;40:693-701.
 15. Russell DJ, Rosenbaum PL, Lane M, et al. Training users in the Gross Motor Function Measure: methodological and practical issues. *Phys Ther* 1994;74:630-6.
 16. Wright B, Masters G. *Rating scale analysis*. Chicago: MESA Pr; 1982.
 17. Magalhaes L, Fisher A, Bernspang B, Linacre M. Cross-cultural assessment of functional ability. *Occup Ther J Res* 1996;16:45-62.
 18. Handlesman D. *The construct validity of the worker role interview for the chronic mentally ill [thesis]*. Chicago: Univ Illinois; 1994.
 19. Smith R, Schumaker R, Bush M. Using item mean squares to evaluate fit to the Rasch model. *J Outcome Meas* 1998;2:66-78.
 20. ShROUT P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8.
 21. Bevington P, Robinson D. *Data reduction and error analysis for the physical sciences*. Toronto (ON): McGraw-Hill; 1992.
 22. Smith R. *Applications of Rasch measurement*. Chicago: MESA Pr; 1992.
 23. *Gross Motor Ability Estimator [computer program]*. Version 1.0. In: *The Gross Motor Function Measure (GMFM-66 and GMFM-88) user's manual*. Clinics in Developmental Medicine. No. 159. London: Mac Keith Pr; 2002.
 24. Russell DJ, Leung KM, Rosenbaum PL. Accessibility and perceived clinical utility of the GMFM-66: evaluating therapists' judgements of a computer-based scoring program. *Phys Occup Ther Pediatr*. In press.

Supplier

- a. Winsteps, PO Box 811322, Chicago, IL 60681-1322.