한국전문물리치료학회지 제12권 제4호
PTK Vol. 12 No. 4 2005.

Scaling of the Korean Version of the GMFM

So-yeon Park, Ph.D., P.T.
Seiyon Child Development Center
Chung-hwi Yi, Ph.D., P.T.
Dept. of Physical Therapy, College of Health Science, Yonsei University
Institute of Health Science, Yonsei University

Abstract

The Gross Motor Function Measure (GMFM) is an internationally widely used outcome measure. The aim of this study was to evaluate the structural properties of the Korean version of GMFM using the Rasch Model, with regard to scoring within rehabilitation centers in Korea. GMFM data for 206 children with cerebral palsy were collected from 11 outpatient rehabilitation facilities by 29 pediatric therapists. The Winsteps software was used to refine the rating scale. This study suggests that the scoring categories of the Korean version of the GMFM should be collapsed from 0 (subject does not initiate task), 1 (subject initiates task), 2 (subject partially completes task), 3 (subject completes task) to 0 (subject does not initiate task), 1 (subject initiates or partially completes task), 2 (subject completes task) for better accuracy in estimating the gross motor function of children with cerebral palsy.

Key Words: Cerebral palsy; GMFM; Rasch analysis.

Introduction

Health care providers recently have renewed the outcome measures used to detect clinical changes during rehabilitation (Avery et al, 2003; Chang and Chan, 1995 Wright and Linacre, 1989). To evaluate clinical changes in a patient, the outcome measures must be clinically relevant, reliable, valid, and adjustable. The criterion-referenced Gross Motor Function Measure (GMFM) has been widely used to evaluate childhood motor function. The GMFM was designed and validate for children with cerebral palsy. The original GMFM (GMFM-88), is composed of 88 items grouped into five functional dimensions: lying and rolling (17 items), sitting (20 items), crawling and kneeling (14items), standing (13 items), and walking, running, and jumping (24 items) (Russell et al, 1989; Russell et al, 2002). Each item is scored on a four-point ordinal 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.

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 (Rasch, 1980). Rasch analysis is used to transform ordinal-scaled measures into interval-scaled measures that provide good clinical precision (reliability) and acceptable fit characteristics (quantitative validity) (Wright and Mok, 2004). A representative advantage of an interval-scaled measure is that it establishes standardized distances between points, allowing for more accurate interpretation of the levels measured. Probabilities (expressed in logits: log-odds probability unit) of the occurrence of each rating are determined by distributing items according to their difficulty and distributing subjects according to their abilities. This results in the development of a single linear scale that is representative of the underlying construct (Wright

Corresponding author: So-yeon Park *soyeonii@dreamwiz.com*

and Linacre, 1989).

Fit statistics are used to identify items that do not fit the Rasch model criterion of unidimensionality, which compromises the scale construct validity (Bond, 2003; Velozo et al, 1995). Using fit statistics makes it possible to improve scaling characteristics and has led to the introduction of shortened versions of GMFM (Avery et al, 2003). For example, Rasch analyses of gross motor function measures resulted in the elimination of misfit items that did not contribute to the measurement of the intended constructs, thus shortening the scales appreciably (Avery et al, 2003). Consequently, several researchers have supported the application of Rasch analysis to refine rehabilitation outcome measures (Andresen, 2000; Page et al, 2002).

The rating scale diagnostics includes category frequencies, average measures, threshold estimates, probability curves, and category fit. These diagnostics should be used in combination with each other. Combining the various diagnostics is very useful for pointing out where we might begin to revise the rating scale to increase the reliability and validity of the measure (Bond and Fox, 2001; Linacre, 1999; Wright and Linacre, 1992).

The strategy for determining the optimal number of response categories requires examination of Rasch measurement diagnosis. Statistics guide us in assessing how the categories function to create an interpretable measure. The simplest way to assess category functioning is to examine category use statistics (category frequencies and average measures) for each response option (Linacre, 1999). Category frequencies indicate how many respondents chose a particular response category, summed for each category across all items. These category frequencies provide the distribution of responses across all categories, providing a very quick and basic examination of rating scale use.

Categories with low frequencies are also problematic because they do not provide enough observations for an estimation of stable threshold values. Infrequently used categories are often unnecessary or redundant. They increase monotonically, indicating that on average, those with higher ability/stronger attitudes endorse the higher categories, whereas those with lower abilities/weaker attitudes endorse the lower categories. When this pattern is violated, as indicated by a lack of monotonicity in the average measures, collapsing categories is recommended.

The magnitude of the distances between the threshold estimates is also important. Guidelines indicate that the threshold should increase by at least 1.4 logits, to show distinction between categories, but not by more than 5 logits, so as to avoid large gaps in the variables (Linacre, 1999).

Fit statistics provide another criterion for assessing the quality of rating scales. Outfit mean squares greater than 2 indicate more misinformation than information (Linacre, 1999), meaning that the particular category is introducing noise into the measurement process. Such categories warrant further empirical investigation, and might be good candidates for collapsing with adjacent categories.

Many researchers have been demonstrated the validity and reliability of GMFM and the Korean version of GMFM applying Rasch analyses (Avery et al, 2003; Palisano et al, 2000; Park, 2005; Russell et al, 2002; Yi and Park, 2004). However, the scaling characteristics of the GMFM have not been studied adequately. The purpose of this study was to analyses the Korean version of the GMFM items and response categories to determine how well they measure the construct of gross motor function.

Methods

Subjects

The subjects were a sample of 206 children with cerebral palsy diagnosed by physicians. There were 125 males, 74 females and 7 unidentified by gender, their mean age was 4.9 years (range: 8 months~14.5 years). The GMFM data used in this cross-sectional

study were collected from 11 outpatient rehabilitation facilities in Korea between August 2004 and April 2005.

Statistical Analysis

Rasch analysis was performed using Winsteps (Winsteps, Chicago, IL, USA) version 3.57.1. Rasch analysis allows the calibration of item responses to construct a scale on which linear measures underlying the observations are defined. It produces an estimate of a person's ability and item difficulty along a shared continuum, in addition to estimates of the fit of each item and person to the measurement model (Bond, 2003; Page et al, 2002). All 206 cases were used in a single calibration to ease interpretation and to provide a stable measure of person and item reliability for the Korean version of GMFM. The results of this calibration were examined to determine the quality of the rating scale and the psychometric characteristics of the instrument, specifically its reliability and validity.

Results

A Rasch analysis examined overall fit of all the Korean version of GMFM items initially. Rasch analysis showed that the original rating scale had adequate separation levels for person (10.93) and item (17.51), as well as good reliabilities for person (.00) and item (1.00). However, #3 item infit value was 2.33, #4 item infit value was 2.36, and #5 item infit value was 2.40. Their 3 items were considered noisy and not contributing measurement of gross motor function. The effect of misfitting persons on the item difficulty estimates was also examined. Eighteen children were excluded because the standard infit value exceeded 2.0.

In the second calibration, 3 items and 18 children were deleted, person and item reliabilities remained the same (Table 1). However, the separation levels for person and item increased (Table 1). Person separation level increased 10.93 to 12.69. Item separation level increased 17.51 to 19.26. Low coherence observation percentages were observed for item response categories 1 and 2 (6%, 7% respectively), meaning that they were inferentially insecure (Table 2).

Table 1. Summary of rating scale analysis calibrations

Calibration	Person separation	Item separation	Person reliability	Item reliability	Modification to next calibration
1	10.93	17.51	.99	1.00	Delete item #3, #4, #5 Delete 18 persons
2	12.69	19.26	.99	1.00	Collapse categories 1 and 2
3	12.09	17.77	.99	1.00	None

Table 2. Summary of the measured steps for second calibration

Category label	Observed count (%)	Average measure	Infit MnSq[a]	Outfit MnSq	Step calibration
0	6,557 (41)	-4.12	.88	2.31	None
1	963 (6)	-.37	1.13	.73	-.06
2	1,179 (7)	.85	.98	3.35	-.18
3	7,097 (44)	4.19	1.02	.98	.24

[a]MnSq: Mean Squares

Examination of the possibility Table confirmed that there was overlap between rating scale categories 1 and 2 (Figure 1). Step calibrations also showed disordered between category 1 and 2 (-.06 and -.18, respectively). The rating scale diagnostics are shown in Table 2 and Figure 1. The average measure values are ordered, but the 1- and 2-step calibrations are disordered. The outfit mean squares of categories 0 and 2 are greater than 2. Outfit mean squares greater than 2 indicate more misinformation than information (Linacre 1999), meaning that the particular category is introducing noise into the measurement process.

In the third (last) calibration, category 1 and 2 were collapsed. The person separation and item separation level were decreased (12.09, and 17.77, respectively) (Table 3). However, the level of person and item reliability were maintained the same. Table 3 and Figure 2 present the results of recategorization of a four-point scale (i.e. 1 and 2 treated as the same response). With three categories instead of four, the problems of the rating scales disappeared. The average measures and step calibrations are ordered, and the probability curves show that each category represents a distinct portion of the underlying variable (Table 4). Consequently, collapsing categories 1 and 2 improves the GMFM rating scale diagnostics.

Discussion

Rasch analysis allowed refinement of the rating scale. In this study, we suggest, after using Rasch analyses, that a three-category scale would be more precise in measuring level of gross motor function in children with cerebral palsy.

The scoring key of the GMFM is provided as a general guideline. "Does not initiate" (0) applies to a child who is requested to attempt an item and is unable to commence any part of the activity. "Initiates" (1) refers to less than 10% task completion. "Partially completes" (2) refers to a child performing from 10% to less than 100% of the task. "Completes" (3) describes 100% task completion (Russell et al. 2002). To determine whether the rating scale for the GMFM items was being used in the expected manner, the probability of each rating (0~3) was examined. The average measure values were ordered, but 1- and 2-step calibrations were disordered. The outfit mean squares of categories 0 and 2 were greater than 2. Outfit mean squares greater than 2 indicate more misinformation than information (Linacre, 1999), meaning that a particular category is introducing noise into the measurement process. When this information was not logical, we

Table 3. Summary of the measured steps for third calibration (diagnostics for 0112 collapsing)

Category label	Observed count (%)	Average measure	Infit MnSq[a]	Outfit MnSq	Step calibration
0	6,557 (41)	-5.25	.81	1.02	None
1	2,142 (13)	.33	.98	.00	-.84
2	7,097 (44)	5.31	.97	.97	.84

[a]MnSq: Mean Squares

Table 4. Comparison of the two categorizations

Category label	Average measure	Fit	Step calibration	Person separation	Item separation
0123	ordered	>2.0	disordered	12.69	19.26
0112	ordered	<2.0	ordered	12.09	17.77

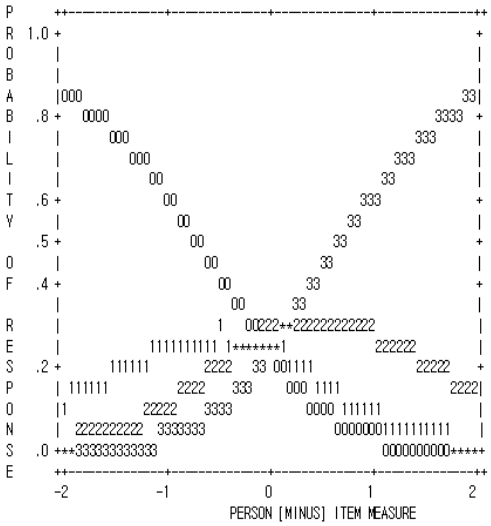

Figure 1. Model probabilities for each response item category, second calibration. Probability curve for the 0123 rating scale

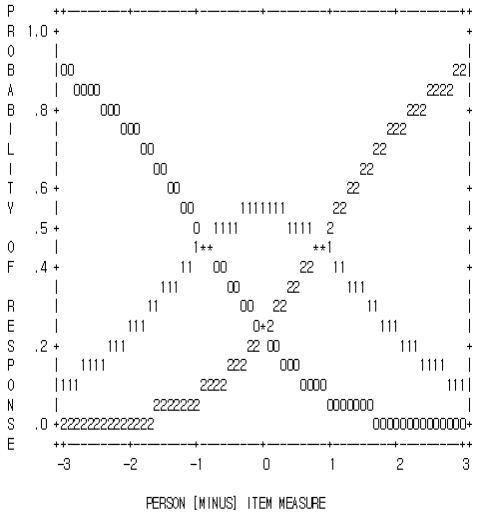

Figure 2. Model probabilities for each response item category, third calibration. Probability curves for 0112 collapsing

combined rating categories (i.e. 1 and 2 were treated as the same response) and reanalyzed the rating scales. With three categories (0, 1, 1, 2) instead of four (0, 1, 2, 3), the problems of the rating scale were solved. The average measures, step calibration order, and probability curves showed that each category represented a distinct portion of the underlying variable. Therefore, collapsing the categories "Initiates" and "Partially completes" may improve the Korean version of the GMFM rating scale diagnostics and make this measurement easier to use clinically.

The revised scale applying Rasch analysis makes it possible for individuals to arrange tests according to the order of difficulty, with the items that are easiest for a sample being tested first, and the most difficult items being assessed last. If patients cannot perform the easiest items, testing can be terminated because patients will most likely be unable to perform the more difficult items.

Conclusion

This study suggests that the scoring categories of the Korean version of the GMFM should be collapsed from 0 (subject does not initiate task), 1 (subject initiates task), 2 (subject partially completes task), 3 (subject completes task) to 0 (subject does not initiate task), 1 (subject initiates, or partially completes task), 2 (subject completes task) for more accurate estimates of gross motor function of children with cerebral palsy.

References

Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000;81(Suppl 2):S15-S20.

Avery LM, Russell DJ, Raina PS, et al. 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.

Bond TG. Validity and assessment: A Rasch measurement perspective. Metodologia de las Ciencias del Comportamiento. 2003;5(2):179-194.

Bond TG, Fox CM. Appling the Rasch Model: Fundamental Measurement in the Human

Sciences. New Jersey, Lawrence Erlbaum Associates, Inc., 2001.

Chang W, Chan C. Rasch analysis for outcomes measures: Some methodological considerations. Arch Phys Med Rehabil. 1995;76:934-939.

Linacre JM. Investigating rating scale category utility. J Outcome Meas. 1999;3(2):103-122.

Page SJ, Shawaryn MA, Cernich AN, et al. Scaling of the revised Oswestry low back pain questionnaire. Arch Phys Med Rehabil. 2002;83(11):1579-1584.

Park S. Application of Rasch analysis to the Korean version of the Gross Motor Function Measure. Doctoral Dissertation, Seoul, Yonsei University, 2005.

Rasch G. Probabilistic Models for Some Intelligent and Attainment Test. Chicago, MESA Press, 1980.

Russell DJ, Avery LM, Rosenbaum PL, et al. Improved scaling of the Gross Motor Function Measure for children with cerebral palsy: Evidence of reliability and validity. Phys Ther. 2000;80(9):873-885.

Russell DJ, Rosenbaum PL, Avery LM, et al. Gross Motor Function Measure (GMFM-66 & GMFM-88) User's Manual. Ontario, Canada, MacKeith Press, 2002.

Russell DJ, Rosenbaum PL, Cadman DT, et al. The gross motor function measure: A means to evaluate the effects of physical therapy. Dev Med Child Neurol. 1989;31:341-352.

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(10):974-985.

Velozo CA, Magalhaes LC, Pan AW, et al. Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. Arch Phys Med Rehabil. 1995;76(8):705-712.

Wright BD, Linacre JM. Observations are always ordinal; Measurement, however, must be interval. Arch Phys Med Rehabil. 1989;70(12):857-860.

Wright BD, Linacre JM. Combining and splitting categories. Rasch Measurement Transactions. 1992;6(3):233-235.

Wright BD, Mok M. An overview of the family of Rasch measurement models. In: Smith EV Jr, Smith RM, eds. Introduction to Rasch Measurement: Theory, models, and application. MN, USA, JAM press, 2004.

Yi C, Park S. Application of Rasch analysis to the Gross Motor Function Measure: A preliminary study. Journal of Korean Academy of University Trained Physical Therapists. 2004;11(2):9-16.

This article was received August 26, 2005, and was accepted October 30, 2005.