J. Phys. Ther. Sci.
23: 327–331, 2011

Development and Validation of the Korean Version of Gross Motor Function Measure

SO-YEON PARK, PT, PhD[1)], CHUNG-HWI YI, PT, PhD[2)], CRAIG A. VELOZO, OTR, PhD[3)]

[1)]*Department of Physical Therapy, College of Alternative Medicine, Jeonju University*
[2)]*Department of Physical Therapy, College of Health Sciences, Yonsei University: #1 Yonseidae-gil, Wonju, Gangwon-do 220-710, Republic of Korea. TEL: +82 33-760-2429, FAX: +82 33-760-2496, Email: pteagle@yonsei.ac.kr*
[3)]*Rehabilitation Outcomes Research Center, Department of Veterans Affairs Medical Center, Department of Occupational Therapy, University of Florida*

Abstract. [Purpose] The purpose of this study was to investigate the item-level psychometrics of a Korean translation of the Gross Motor Function Measure (GMFM) using Rasch analysis. [Methods] GMFM data for 206 children (mean age: 4.9 years) with cerebral palsy were collected in Korea. Focusing on the psychometric properties of the item measures, person measures, and rating scale categories, Rasch analysis was performed. Principal components analysis in combination with fit statistics was used to verify the unidimensional construct of GMFM. [Results] The original 4-point rating scale revealed that it was not being used in the expected fashion and could be improved by collapsing into a 3-point version. Eighty-eight items showed a logical item difficulty hierarchy. Furthermore this revision instrument appears to have good internal construct validity showing a person separation reliability of 0.99. The GMFM has a person separation index of 10.67 which indicates that the instrument separates the sample into approximately 15 statistically significant strata. [Conclusion] The rescaled version of the GMFM shows outstanding item-level psychometric qualities. This study suggests that the Korean version of the GMFM produces valid measures of gross motor ability in children with CP and furthermore appears to be sensitive in differentiating children with different gross motor abilities.
Key words: Cerebral palsy, Rehabilitation, Treatment outcome

(This article was submitted Sep. 27, 2010, and was accepted Oct. 26, 2010)

INTRODUCTION

Physical therapists have used assessment tools to measure patients' clinical status or to identify the effects of rehabilitative treatment for decades. While it is important to have evaluation tools for monitoring the development of gross motor function of children with cerebral palsy, unfortunately, most of the outcome measures for assessing motor function were developed outside of Korea and are intended for use in English speaking countries. Many non-English speaking countries, including Korea, have a desperate need for outcome measures.

Park and colleagues[1)] surveyed 130 Korean pediatric physical therapists regarding their use of evaluation instruments in their practice. Virtually all physical therapists (99.9%) reported the need to evaluate children with cerebral palsy, but they also reported that few standardized tools were available to assess function in this population. The standardized tools reportedly used include a subjective description format (47.1%), the Gross Motor Function Measure (GMFM) (21.3%), a facility-generated tool (18.8%), the Denver Developmental Screening Test (5.0%), the Test of Infant Motor Performance (2.6%), the Peabody Developmental Motor Scales (1.8%), the Alberta Infant Motor Scale (1.1%), the Pediatric Evaluation of Disability Inventory (1.1%), and the Bruininks-Oseretsky Test of Motor Performance (0.7%)[1)]. Although there is a great need for functional assessments for Korean children, a large percentage of these assessments are subjective descriptions or facility-generated tools.

English versions of the criterion-referenced GMFM have widely been used to evaluate childhood motor function[2–4)]. It shows good reliability and validity with respect to responsiveness to change of functional gross motor abilities of children with cerebral palsy[2,3,5)]. The English version of the original GMFM is composed of 88 items grouped into five 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)[6)]. 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 to 3 indicating that the child can complete the item. Each of the scoring options within the 88 items is defined explicitly, describing the motor behavior to be

observed and scored. The percentage scores for each dimension are summed and averaged to obtain the total GMFM score.

In the rehabilitation field, Item Response Theory (IRT) has become more and more frequently used to develop outcome measures and to evaluate the validity of translated instruments[7]. IRT psychometric approaches, in particular Rasch analysis, may be useful for evaluating translated instruments, since they provide psychometric information at the item level. IRT models estimate the probability of item responses and the level of each person on the trait that the items are intended to measure[8]. The simplest of the models, the Rasch, one-parameter model, 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[9]. Rasch analysis can be used to assess multiple psychometric characteristics: fit statistics, item difficulty (hierarchy), and the relation of item difficulty to person ability[10,11]. Furthermore, Rasch analysis transforms ordinal scale measures into interval scale measures and provides indicators of precision (reliability) and fit characteristics (quantitative validity). An advantage of interval scale measures is that they establish 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 a single linear scale that is representative of the underlying construct[12].

Rasch analysis has been used to aid in the construction and validation of rehabilitation outcome measures. It has also been used to validate translated measurement tools. Kjellberg et al.[15] translated the Assessment of Communication and Interaction Skills into Swedish and verified its validity and reliability using Rasch analysis. A Turkish version of the Stanford Health Assessment Questionnaire, the Modified Barthel Index, and the Nottingham Health Profile have been studied using Rasch analysis[16–18]. The Japanese version of the WeeFIM instrument was also validated using Rasch analysis[19].

In Korea, the need for a gross motor function measure is recognized and a Korean version of the GMFM-88 has been used to evaluate gross motor function more widely than any other outcome measure. The purpose of this study was to carry out a cross-cultural adaptation of the original GMFM for use with Korean-speaking patients and to investigate the psychometric properties of the Korean version of GMFM using Rasch analysis.

SUBJECTS AND METHODS

Subjects

The health center institutional review board of the University of Florida approved this protocol, and all participants or their parents gave informed consent. The participants 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 (SD: 2.5). The GMFM data used in this cross-sectional study were collected from 11 outpatient rehabilitation facilities in Korea.

Methods

This study used the extensive revised version of the GMFM in 2002[6]. The GMFM user's manual was translated into Korean. After the initial translation, the participants, twenty-nine Korean pediatric physical therapists and three Korean pediatric occupational therapists, attended a one-day GMFM workshop, and reviewed every sentence of the Korean translated GMFM manual. Changes were communicated to participants through a website and the principal researcher of this study made minor modification until all workshop participants agreed that the meaning of each interpretation was clear. One-day GMFM workshops were held between August, 2004 and February, 2005, to prepare the therapists in the use of the Korean manual for the GMFM. In addition, a GMFM self-instructional training CD-ROM was provided[6]. All thirty-two therapists (mean clinical experience 5.2 years) assessed children with cerebral palsy using the Korean version of the GMFM.

Focusing on the psychometric properties of the item measures, person measures, and rating scale categories, a Rasch rating scale analysis was performed using Winsteps (Chicago, IL, USA) version 3.57.1. Rating scale diagnostics include category frequencies, average measures, probability curves, and category fit11,[20,21]. Each rating scale response category was examined for frequency and average measure[20]. Unidimensionality indicates that the items should assess a single dimension or construct of a gross motor function measure. PCA in combination with fit statistics was used to assess the unidimensionality of the Korean version of GMFM[22]. Infit and outfit mean square standardized residuals (Mnsq) and standardized Z values (ZSTD, or z-score) were used to check the compatibility of the data with the Rasch model[11]. The infit statistics are more affected by unexpected responses close to each person's ability, while the outfit statistics are more sensitive to unexpected responses far from the person's ability, so the infit mean square statistics were used to assess fit[6,11]. Items with infit values > 1.4 and infit z values > 2.0 were considered misfits[14].

Items that are used to represent a construct should form a hierarchy of difficulty, ranging from the least difficult for the subject to perform to the most difficult to perform. In Rasch analysis, both personal ability and item difficulty are expressed as a logit, the natural logarithm of the odds of a person being able to perform a particular task. Logits of greater magnitude represent increasing item difficulty. When instruments are developed using a conceptual hierarchy of items, the empirical ordering produced by the Rasch analysis can be compared with the theoretical ordering, and the result can be treated as evidence of construct validity[13,23].

Winsteps software also produces several psychometric indices. Person separation reliability, analogous to Cronbach's alpha, provides an indication of the internal

Table 1. The scale of the Korean version of the GMFM. The column marked "original" refers to the original 0123 scale. The column marked "revised" refers to the revised 0112 scale based on the Rasch analysis

Original Scale				Revised Scale			
Category Label	Average Measure	Outfit MnSq[a]	Step Calibration	Category Label	Average Measure	Outfit MnSq[a]	Step Calibration
0	–3.12	4.91	None	0	– 4.15	5.60	None
1	–0.25	0.97	0.22	1	0.35	1.97	–0.63
2	0.75	1.52	–0.25				
3	3.28	1.07	0.02	2	4.30	0.97	0.63

[a] MnSq: mean squares, Original scale, 0 = does not initiate, 1 = initiates (completion of less than 10% of item), 2 = partially completes (completion of 10% to less than 100%), 3 = completes; Revised scale, 0 = does not initiate, 1 = completion of less than 100% of item, 2 = completes.

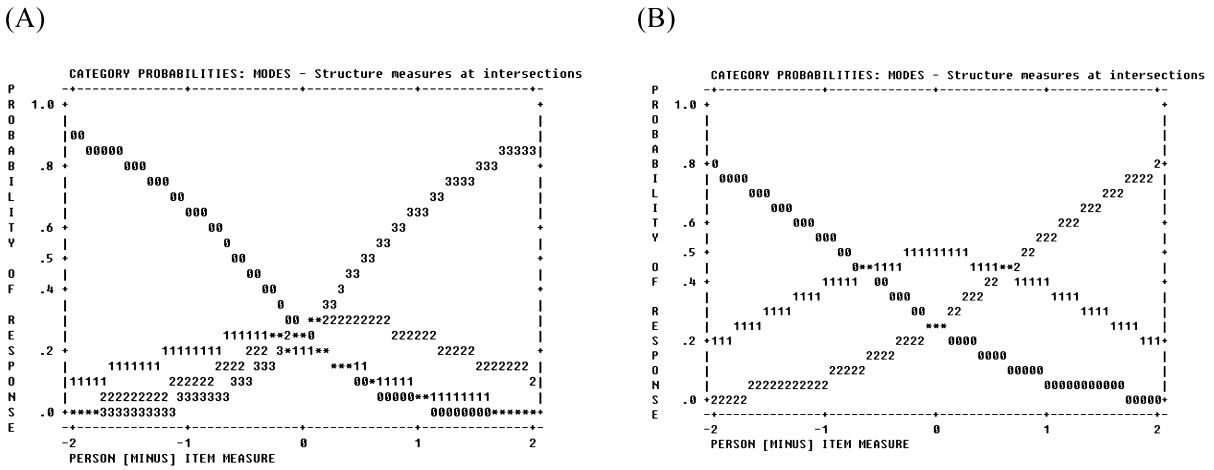

Fig. 1. (A) Original probability curve for the each of the 4 ratings (0123). (B) Revised Probability curve for 0112 collapsing.

consistency of the instrument. Separation ratio (SR) provides an indication of the number of statistically significant strata to which the sample is divided [SR=(4Gp+1)/3, where Gp=person separation reliability][24].

RESULTS

Before proceeding to determine whether the rating scale for the Korean version of the GMFM items was being used in the expected manner, the probability of each rating (0, 1, 2, 3) was examined. While most of the items have specific descriptions for each of their ratings, generally, 0 is "does not initiate", 1 is "initiates", 2 is "partially complete", and 3 is "completes".

The rating scale diagnostics are shown in Table 1 and Figure 1. The average measure values follow the expected order (low-to-high), but the step calibrations are disordered. The outfit mean squares of category 0 was greater than 2 suggesting that this particular category is introducing more noise into the measurement process than expected.[20] Figure 1(A) further demonstrates the limitations of the rating scale. As one proceeds along the x-axis from low values to high values of person-minus-item, we would expect each part of the rating scale to emerge as more probable than the others. As can be seen from the figure, only the lowest (0) and highest (3) parts of the rating scale emerged as more probable than the other ratings. That is, ratings of 1 and 2 never become more probable than the ratings of 0 and 3.

Based on the above results, the ratings of 1 and 2 were combined in this scale. As one proceeds from along the x-axis from low values to high values of person-minus-item, we would expect each part of the rating scale to emerge as more probable than the others and ability curves show that each category represents a distinct portion of the underlying variable (Table 1). While the outfit Mnsq of category of 0 continue to be greater than 2, since the collapsed scale was an improvement over the four-point scale, it was used for all further analyses.

After the rescale of the rating scale, PCA and fit statistics were performed to define how well all 88 items of the Korean version of GMFM contributed to unidimensionality. PCA results revealed that sixteen components had eigenvalues >1: the first component explained 38.19% of the covariance among the 88 items, the second component explained 13.34%, and the third component explained 5.51%.

Fit statistics analysis was used to determine how well the 88 items fit the Rasch measurement model (infit MnSq < 1.4 and infit z-values < 2.0). Six items were identified as misfits: Supine: Lift head 45 degrees; Supine: Flexes right hip and knee through full range; Supine: Flexes left hip and knee through full range; Prone: Rolls to supine over right

side; Supine: Roll to left side, attains sitting; and Right side sitting: Maintains, arms free, 5 seconds. While some investigators consider removing misfitting items from instruments, for the purposes of this study all items were retained for all further analyses.

On the item difficulty maps generated from the Korean version of the GMFM data, the item difficulty was ordered hierarchically on the same linear continuum as person ability. Figure 2 illustrates item difficulty hierarchical order of the 88 items on the same linear continuum as person ability. The maps generated from the Korean version of the GMFM data include a vertical logit scale in the center with the participants ordered by ability level in the left-hand column (lowest ability at the bottom; highest ability at the top) and the items ordered by difficulty level in the right-hand column (easiest items at the bottom; hardest items at the top). Items that involved hopping, and jumping had a tendency to be the most difficult items. Items that involved lying down and head control in supported sitting had a tendency to be the easiest. Items that involved kneeling and standing had a tendency to be of "medium difficulty". The distribution of personal ability and item difficulty were virtually identical showing similar means (M) and spread (S=1 standard deviation, T=2 standard deviations). Person ability showed a normal distribution with no ceiling or floor effects.

Person separation reliability, analogous to Cronbach's alpha was 0.99. The person separation index, an indication of how well the instrument separates the children in the sample, was very high at 10.67. The instrument separated the sample into approximately 15 statistically distinct strata.

DISCUSSION

The purpose of this study was to evaluate the Rasch measurement properties of a version of the GMFM translated into Korean in the evaluation of the gross motor function of children with cerebral palsy in Korea. Following modification of the rating scale, the translated instrument showed many good-to-excellent psychometric qualities (e.g., unidimentionalityunidimensionality, few misfitting items, excellent person/item separation and good person-item match).

The original four-point rating scale of the GMFM was not supported by Linacre's essential rating scale criteria[25]. While Linacre[25] presents 8 criteria for an "optimal" rating scale, three of those criteria are essential: (1) at least 10 cases per category, (2) monotonically increasing average measures across category, and (3) a category outfit mean square value less than 2. The original four-point rating scale showed high outfit statistics for the rating category "0". According to Linacre, this finding suggests that the "0" category is introducing more noise into the measurement process than expected. Furthermore, the probability plot of the rating scale categories against the person-minus-item difficulty, showed that the middle ratings (1 and 2) never became more probable than the ratings of 0 and 3. These findings suggest that the two middle-rated categories may have been difficult for therapists to reliably rate. Collapsing the middle two categories did not improve the outfit statistics of the lowest rating category, however it did result in the three rating categories showing a clear logical ordering.

The Korean version of the GMFM indicated a unidimensional construct supported by both the fit statistics and PCA. After modifying the rating scale, PCA and fit statistics were performed. PCA of the residuals indicated that less than 14% of the residual variance was accounted for in the first principal component[20]. There were sixteen components with eigenvalues >1: the first component explained 38.19% of the covariance among the 88 items, the second component explained 13.34%, and the third component explained 5.51%. These results, as well as the lack of a meaningful pattern in the assignment of items to components, suggest that the residuals are randomly distributed and support the unidimensionality assumption of the GMFM.

To determine the fit of the GMFM items to the Rasch model, items with infit mean-squares greater than 1.4 and infit z-values greater than 2 were considered misfits[11]. In this study, six items involving lying down and rolling were identified as misfits, and since three of these items, represented unilateral movement, the counterparts of these items could also be considered for removal. The lying down and rolling items had a tendency to be among the least difficult items and anecdotally, therapists reported difficulty in scoring these items. In their study of the GMFM-88, Avery and colleagues[26] used an infit mean square criterion of 3.0. Using this criterion, a Rasch analysis of the remaining 66 items was performed and the results of this analysis were used to develop the scoring for the GMFM-66[26]. In the present study, applying a critical value of 3.0, all 88 items would have fit the Rasch model. Because there was a considerable difference in the sample size between the study of Avery et al. and the present study, the results of the fit statistics may differ[11,26].

To obtain meaningful measures of gross motor function, it is essential to determine the item difficulty hierarchy defining the measure. The hierarchical arrangement found in this study is generally supported by previous studies. For example, in this analysis, "Standing: Hops on right foot 10 times within a 24 inch circle" was the most difficult and "Standing: Hops on left foot 10 times within a 24 inch circle" was the second most difficult, and "Sitting on mat, supported at thorax by therapist: Lift head upright, maintain 3 seconds" was the easiest to pass. According to the study of Avery et al.[26] "Standing: Hops on left foot 10 times within a 24 inch circle" was the most difficult, "Standing: Hops on right foot 10 times within a 24 inch circle" was the second most difficult, and "Sitting on mat, supported at thorax by therapist: Lift head to midline, maintain 3 seconds" was the easiest. That is, the lying and head control in supported sitting items were the easiest, whereas the hopping and jumping items were the most difficult.

The separation index of Rasch analysis is an estimation of the number of distinct strata of persons and items[11]. With a larger index, more distinct levels of functioning can be distinguished in the measure. Duncan et al.[13] suggested that

in general, a person separation index of 2.0 represents a good level of separation, and 3.0 represents an excellent separation. The person separation index in the present study was 10.67 which indicates that the instrument separates the sample into about 15 statistically significant strata[24], a level far exceeding Duncan and colleaguescolleagues' suggestion of excellent separation[13]. The person separation reliability is an estimation of the replicability of person placement, when the error is small and the ratio is high. The Korean version of the GMFM demonstrated an excellent person separation reliability of 0.99 (analogous to Cronbach's alpha), and it appears to be highly sensitive in differentiating children with different gross motor abilities.

This study used Rasch analysis to assess the validity of the Korean version of the GMFM. In Korea, most therapists realize the need for objective outcome measures for assessing gross motor function, but have been using subjective tools or translated tools without verified objectivity. This is the first study to validate a Korean-translated rehabilitation measure using Rasch analysis. We refined the rating scale, identified which items formed a unidimensional construct of the gross motor function measure, and verified the difficulty hierarchy of the 88 items retained. Compared to Avery's previous study[26], our items showed few misfits and no ceiling or floor effects when assessing children with cerebral palsy. Based on these results the Korean version of the GMFM should be useful for clinical practice and research in Korea.

ACKNOWLEDGEMENTS

We would like to thank Dr. Hyukcheol Kwon, Dr. Ohyun Kwon, Dr. Duckhyun Ahn, and Dr. Hyeseon Jeon for their support during the study. We also would like to thank the pediatric therapists who collected the data for this study.

REFERENCES

1) Park H, Yi C, Cho S, Kwon H, et al.: Physical therapists' understanding and the usage of assessment tools for children with delayed development and cerebral palsy. Journal of the Korean Academy of University Trained Physical Therapists, 2000, 7: 1–21.
2) Russell D, Rosenbaum P, Cadman D, et al.: The gross motor function measure: a means to evaluate the effects of physical therapy. Dev Med Child Neurol, 1989, 31: 341–352.
3) Ketelaar M, Vermeer A, Helders P: Functional motor abilities of children with cerebral palsy: A systematic literature review of assessment measures. Clin Rehabil, 1998, 12: 369–380.
4) Vos-Vromans D, Ketelaar M, Gorter J: Responsiveness of evaluative measures for children with cerebral palsy: The gross motor function measure and the pediatric evaluation of disability inventory. Disabil Rehabil, 2005, 27: 1245–1252.
5) Rosenbaum P, Russell D, Cadman D, et al.: Issues in measuring change in motor function in children with cerebral palsy: A special communication. Phys Ther, 1990, 70: 125–131.
6) Russell D, Rosenbaum P, Avery L, et al.: Gross Motor Function Measure (GMFM-66 & GMFM-88) User's Manual. Ontario: MacKeith Press, 2002.
7) Andresen E: Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil, 2000, 81: S15–S20.
8) Fortinsky RH, Garcia RI, Joseph Sheehan T, et al.: Measuring disability in Medicare home care patients: Application of Rasch modeling to the outcome and assessment information set. Med Care, 2003, 41: 601–615.
9) Rasch G: Probabilistic Models for Some Intelligent and Attainment Test. Chicago: MESA Press, 1980.
10) Wright B, Mok M: An Overview of the Family of Rasch Measurement Models. Minnesota, JAM press, 2004.
11) Bond T, Fox C: Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd ed). New Jersey: Lawrence Erlbaum Associates Inc., 2007.
12) Wright B, Linacre J: Observations are always ordinal; Measurement, however, must be interval. Arch Phys Med Rehabil, 1989, 70: 857–860.
13) Duncan PW, Bode RK, Min Lai S, et al.: Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. Arch Phys Med Rehabil, 2003, 84: 950–963.
14) Velozo CA, Peterson EW: Developing meaningful fear of falling measures for community dwelling elderly. Am J Phys Med Rehabil, 2001, 80: 662–673.
15) Kjellberg A, Haglund L, Forsyth K, et al.: The measurement properties of the swedish version of the assessment of communication and interaction skills. Scand J Caring Sci, 2003, 17: 271–277.
16) Kucukdeveci AA, McKenna SP, Kutlay S, et al.: The development and psychometric assessment of the Turkish version of the Nottingham Health Profile. Int J Rehabil Res. 2000, 23: 31–38.
17) Kucukdeveci AA, Sahin H, Ataman S, et al.: Issues in cross-cultural validity: Example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. Arthritis Rheum, 2004, 51: 14–19.
18) Kucukdeveci AA, Yavuzer G, Tennant A, et al.: Adaptation of the modified Barthel Index for use in physical medicine and rehabilitation in Turkey. Scand J Rehabil Med, 2000, 32: 87–92.
19) Tsuji T, Liu M, Toikawa H, et al.: ADL structure for nondisabled Japanese children based on the Functional Independence Measure for Children (WeeFIM). Am J Phys Med Rehabil, 1999, 78: 208–212.
20) Linacre JM: Investigating rating scale category utility. J Outcome Meas, 1999, 3: 103–122.
21) Wright B, Linacre JM: Combining and splitting categories. Rasch Measurement Transactions, 1992, 6: 233–235.
22) Smith R: A comparison of methods for determining dimensionality in Rasch measurement. Struct Equ Modeling, 1996, 3: 25–40.
23) Velozo CA, Wang Y, Lehman L, et al.: Utilizing Rasch measurement models to develop a computer adaptive self-report of walking, climbing, and running. Disabil Rehabil, 2008, 30: 458–467.
24) Wright B, Masters G: Number of Person or Item Strata (4G+1)/3. Rasch Measurement Transactions, 2002, 16: 888.
25) Linacre JM: . Optimizing rating scale category effectiveness. J Appl Meas, 2002, 3: 85–106.
26) 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.