Authors:

Chi-Wen Chien, MEd
Trevor G. Bond, PhD

Affiliations:

From the School of Education (CWC, TGB), James Cook University, Townsville, Australia.

Correspondence:

All correspondence and requests for reprints should be addressed to Chi-Wen Chien, MEd, Department of Occupational Therapy, Monash University-Peninsula Campus, Building G, McMahons Road, Frankston, Victoria 3199, Australia.

Disclosures:

This study was supported by research grants from the School of Education (Internal Research Grant) and the Faculty of Arts, Education and Social Science (Graduate Research Scheme) at James Cook University. Parts of the materials contained in the manuscript were presented at the Annual Meeting of the Pacific Rim Objective Measurement Symposium, Hong Kong, June 27–29, 2006, and the Asia Pacific Occupational Therapy Congress, Hong Kong, June 23–26, 2007.

0894-9115/09/8805-0376/0
American Journal of Physical Medicine & Rehabilitation
Copyright © 2009 by Lippincott Williams & Wilkins

DOI: 10.1097/PHM.0b013e318198a7c9

ORIGINAL RESEARCH ARTICLE

Measurement Properties of Fine Motor Scale of Peabody Developmental Motor Scales-Second Edition

A Rasch Analysis

ABSTRACT

Chien CW, Bond TG: Measurement properties of fine motor scale of peabody developmental motor scales-second edition: A Rasch analysis. Am J Phys Med Rehabil 2009;88:376–386.

Objective: To investigate the measurement properties (including rating scale performance, unidimensionality, and differential item functioning) of the fine motor scale of the Peabody Developmental Motor Scales-Second Edition in children, by using the Rasch analysis.

Design: A total of 419 children (including 342 typically developing children and 77 children with fine motor delays or difficulties) were recruited in Taiwan for this prospective study. Each child was evaluated with the Peabody Developmental Motor Scales-Second Edition that consists of 26-item grasping and 72-item visual-motor integration subtests. Partial credit Rasch analysis was used for all analyses.

Results: The Rasch analysis indicated that middle rating category for 19 grasping and 52 visual-motor integration items could be collapsed to allow only dichotomous response categories. Item fit analysis and principal component analysis suggested that the unidimensionality of the grasping and visual-motor integration subtests could be achieved after removal of two grasping and eight visual-motor integration misfitting items. All but 13 items in the composite scale could form a unidimensional construct of overall fine motor ability. Furthermore, only a few items were found to show differential item functioning across sex (ten items) or fine motor status (seven items). However, significant ceiling effects were found in the Peabody Developmental Motor Scales-Second Edition subtests and composite scale when applied to these typically developing children.

Conclusions: Our results suggest grounds for the revision of the Peabody Developmental Motor Scales-Second Edition in a subsequent edition. Simplifying the rating scales and reducing the misfitting items in the subtests and composite scales might result in a unidimensional assessment of children's fine motor ability. Clinicians and researchers could use the reduced Peabody Developmental Motor Scales-Second Edition as a criterion-referenced outcome measure to document changes; however, further work is needed to reduce the ceiling effects.

Key Words: Motor Skills, Outcome Assessment (Health Care), Rehabilitation

Fine motor (FM) development is an important component for children's general growth and successful participation in daily activities.[1,2] Studies found that children who present with FM problems can benefit from early intervention to optimize their development and to prevent further complications.[3,4] Therefore, measuring FM development in children at an early stage can assist clinicians in identifying at risk children, achieving timely interventions, selecting suitable therapy programs, monitoring progress, and determining the effectiveness of early intervention.[1,5]

A number of published assessment tools are available to assess children's FM development in rehabilitation and early intervention settings.[5] One of the most commonly used standardized tests is the FM scale of the Peabody Developmental Motor Scales-Second Edition (PDMS-2-FM).[6,7] The PDMS-2-FM can be administered to children across a wide age range from birth to 72 mos. The PDMS-2-FM provides two distinct subtests as indicators of children's grasping and visual-motor integration (VMI) abilities, respectively. A combination of the two subtests forms the FM composite scale, which gives an indication of children's overall FM ability. Furthermore, each PDMS-2-FM item employs a three-point rating scale, which allows partial credit for developing FM skills and allows for evaluation of the progress of children with whose skill acquisition develops at a slower rate. Specific scoring criteria for each item were provided in the administration manual to assist in making objective evaluations with the three-point rating scale.[6]

With regard to the psychometric properties of the PDMS-2-FM, Folio and Fewell,[6] the PDMS-2-FM developers, used an item response theory model to investigate the item characteristics (including item difficulty and discrimination statistics). Accordingly, unsatisfactory items were eliminated from the PDMS-2-FM. A procedure of logistic regression was used to identify items showing differential item functioning (DIF) or item bias (i.e., items performed differently from one group to another) across sex and ethnicity. Confirmatory factor analysis was used to assess the construct validity (i.e., dimensional structure) of both the PDMS-2-FM subtests and composite scale.[6] In addition, the PDMS-2-FM seemed to demonstrate reasonable interrater and test-retest reliability,[8,9] convergent validity,[6,9–11] and responsiveness.[8]

The measurement properties of a test must be investigated repeatedly until a conclusive body of scientific evidence has been accumulated.[12] Although the reliability and validity of the PDMS-2-FM have been examined as described earlier, at least three limitations require further scientific investigation. First, the appropriateness of three-point rating scales of the PDMS-2-FM items has little empirical support. It is thus difficult for users to conclude whether the three-point scales can differentiate children's FM abilities as clearly as the test developers propose. Second, factor analysis operates on interval scale scores,[13] whereas the PDMS-2-FM scores are ordinal. Test items, when factor analyzed, are assumed to have the same level of difficulty,[14] but the PDMS-2-FM items are instead ordered from easy to difficult. Therefore, the results of Folio and Fewell[6] using factor analysis might be inadequate to ascertain whether the PDMS-2-FM items could be combined and shown to work together to define single constructs of the subtests or composite scale. Third, although DIF analysis was conducted with the PDMS-2-FM, the biased items were not reported and their retention was not specifically justified.[6] Consequently, clinicians and researchers might inadvertently misuse the PDMS-2-FM to make invalid interpretation, without full awareness of the tool's potential weaknesses.

The Rasch model,[15] is another item response theory model that has been increasingly applied to aid in the validation of assessment tools within rehabilitation and early intervention fields.[16–19] The reason for increasing use of the Rasch model is that it conforms with the idea of fundamental measurement.[17] Unlike other item response theory models that add additional parameters to find the best model to explain observed rating scale data, the Rasch model aims to determine the extent to which the observed data satisfy the model's stringent requirements for interval level measurement. The Rasch model is based on the principle that only two attributes determine participants' responses to test items: the ability of the participant and the difficulty of the test item, each expressed as estimates on the underlying latent trait. It is a probabilistic model that converts ordinal scores obtained from rating scales into interval level measures. Residuals derived from measures that achieve interval level scaling can be further used to assess test unidimensionality (i.e., the extent to which items in a test measure a single construct), as construct validity evidence. By placing participants and items along an equal interval metric in hierarchical order, the Rasch model offers the advantage of examining test targeting (or the extent to which items are of appropriate difficulty for the sample). Furthermore, the Rasch model allows for the determination of whether the rating scale of a test is used in an appropriate manner as well as providing for the detection of DIF items by comparing item difficulties across different demographical variables. As a result of these important advantages, several children's FM instruments have been recently validated using the Rasch model, including the Assisting Hand Assessment[20] and ABILHAND-Kids questionnaire.[21]

This study aimed at using the Rasch model to investigate the measurement properties of the PDMS-2-FM, including (1) the appropriateness of the three-point rating scales, (2) the unidimensionality of the individual subtests and composite scale, and (3) possible DIF across sex and FM status (delayed or not delayed), in a group of Taiwanese children. In addition, the PDMS-2-FM items were developed originally in the United States and have not yet been validated for use in Asia. Previous studies[22,23] have found that cultural influences might be associated with developmental differences in FM skills between East Asian and Western children and might further affect the suitability of the FM tests in some cross-cultural contexts. Therefore, this study also investigated whether the American-developed PDMS-2-FM items would be appropriate (or could be tailored) to assess the ability levels of Taiwanese children.

TABLE 1 Demographic characteristics of the 419 Taiwanese children

Demographic Variables	Normative Sample (n = 342)	Clinical Sample (n = 77)
Gender, n (%)		
Boys	183 (53.5)	55 (71.4)
Girls	159 (46.5)	22 (28.6)
Average age in months, mean ± SD	37.6 ± 21.2	44.2 ± 16.5
Age in months, n (%)		
0–11	48 (14.0)	3 (3.9)
12–23	61 (17.8)	7 (9.1)
24–35	48 (14.0)	15 (19.5)
36–47	61 (17.8)	12 (15.5)
48–59	54 (15.9)	30 (39)
≥60	70 (20.5)	10 (13)
Diagnosis, n (%)		
Development delay	—	38 (49.4)
Sensorimotor disorders	—	17 (22.1)
Cerebral palsy	—	10 (13)
Autism	—	5 (6.5)
Down syndrome	—	4 (5.2)
Mental retardation	—	2 (2.5)
Spina bifida	—	1 (1.3)

METHODS

Participants

A total of 419 children were recruited from 6 day care centers/kindergartens and 4 rehabilitation units throughout central and northern Taiwan, between January 1, 2006 and August 31, 2006. Of those, 342 typically developing children from birth to 6 yrs of age comprised the normative sample, meeting the following criteria: (1) full-term infants (born between 36 and 42 wks), (2) birth weight of 2500 g or more, and (3) absence of any known sensorimotor deficits, major diseases (e.g., cancer or heart disease), or body impairments (e.g., amputations or fractures) that would limit the child's ability to perform movement tasks, according to the parents' reports.

In addition, 77 children who presented with FM delays or problems were recruited to constitute the clinical sample. Children were included in the clinical sample if they had a formal medical diagnosis or rehabilitation-related disorder such as cerebral palsy, developmental delay, Down syndrome, autism, and sensorimotor disorders, and presented with FM delays or difficulties (according to the referral therapists' assessments/observations), but other major diseases (e.g., cancer or heart disease) or bodily impairments (e.g., amputations or fractures) were absent. The inclusion of these children in the clinical sample was to provide a range of FM delays that were caused by diverse problems/diagnoses. Demographic characteristics of the 419 children in the 2 samples are presented in Table 1.

Procedure

The ethics committee of the university as well as the institutional review board of the participating day care centers/kindergartens and hospitals, approved the study protocols. All participants' parents/caregivers signed informed consent forms before participation in the study.

Each participant was administered the PDMS-2-FM individually in a quiet setting at the day care center/kindergarten facility or at the participant's home. Parents and caregivers were allowed to be involved in the PDMS-2-FM administration, if the child was aged <2 yrs or seemed uncomfortable when left alone with the examiner. One occupational therapist who was previously trained in administering the PDMS-2 conducted the assessments of all participating children.

Measure

The PDMS-2-FM, a norm-referenced test, comprises 98 items that are further divided into 2 subtests: Grasping (26 items) and VMI (72 items). Each item is scored on a three-point rating scale (0-1-2). General criteria for scoring the items were described as follows: 0 indicating that the child cannot perform the item, 1 indicating that the child performs the item but cannot fully meet the criteria, and 2 indicating that the child can complete the item according to the criteria specified for mastery.

A modified PDMS-2-FM administration procedure in this study was adapted to enable maximal administration of the items appropriate to the children's ability. The modified administration procedure was used because the original involves a standard procedure that is limited to administration of the items between the basal and ceiling levels.[6] The items below the basal level (or those above the ceiling level) were not administered, instead they were replaced with full (or zero) scores. This standard procedure is based on robust and orderly FM developmental progress, but previous studies have not supported the existence of the robust developmental hierarchy, especially in children who presented with FM problems.[5,24] Therefore, the current study divided all of the PDMS-2-FM items into six age bands (i.e., 0–11, 12–23, 24–35, 36–47, 48–59, and 60–71 mos), according to the possible age in which children might reasonably be expected to achieve the items.[6] Children in the normative group were assessed using items across three age bands (i.e., the child's current age band as well as one age band below and one above the child's current age). Whereas children in the clinical group, because of the potential impact of FM delays or difficulties, were assessed using items across wider age bands (i.e., the child's current age band plus two age bands below but only half above the child's current age). The items outside the stipulated age bands in either the normative or clinical groups were neither administered nor replaced with full (or zero) scores.

In addition, the order of presenting the PDMS-2-FM items was reorganized for the modified administration. Items requiring the same test equipment/position (e.g., all items involving the use of pen and paper) were grouped to help the administration flow smoothly and maintain the child's level of concentration and motivation. Thus, the testing of the PDMS-2-FM commenced with the first item corresponding to the child's age or interest or both, based on the examiner's judgment. Then, the group of the items requiring the same test equipment/position was assessed. The PDMS-2-FM administration was concluded once the child had completed all of the items across the stipulated age bands. If the PDMS-2-FM administration could not be completed in one session, the outstanding tests were completed within a 5-day interval period.[6]

Because no Chinese version of the PDMS-2-FM was currently available in Taiwan, the instructions for each PDMS-2-FM items were translated by the authors, before the study's commencement. Furthermore, the authors extracted the illustrations from the manual for each item administration and the summarized notes for the corresponding scoring criteria. These two additional elements together with translated instructions were incorporated in the reorganized PDMS-2-FM evaluation sheets to guide administration and scoring.

Data Analysis

Partial credit model Rasch analysis was performed using Winsteps,[25] version 3.61.1. The partial credit model[26] was selected because the PDMS-2-FM items use a three-point rating scale and the scoring criteria of each rating category vary between items. The Rasch partial credit model yields an ability estimate and an item difficulty estimate associated with accomplishing each step in an ordered sequence of rating categories.[26]

The Rasch analysis of the PDMS-2-FM consists of four parts. First, the appropriateness of the three-point rating scales in each item was investigated using a combined sample (children in the normative and clinical groups together). The rating categories of the items were reorganized (i.e., some were collapsed) if they were found to be inadequate according to Linacre's criteria for optimizing rating scale category effectiveness,[27] as follows: (1) at least ten cases per category, (2) monotonically increasing average measures across categories, (3) category outfit mean square (MnSq) values <2, and (4) monotonically increasing step calibrations.

Second, the extent to which PDMS-2-FM items with appropriate rating categories contributed to a unidimensional construct was examined using the normative sample only. Children in the clinical group were not included because their FM impairments might lead to misfit in certain items which might conform to typical FM development. With the normative sample, therefore, we analyzed the PDMS-2-FM items in each subtest (i.e., the grasping and VMI subtests) separately and in the FM composite scale. Fit statistics were used to monitor the compatibility of the raw item data with Rasch model expectations.[17] They include two types of fit statistics: the infit statistic (weighted) is most sensitive to ratings on the items located close to the children's ability, whereas the outfit statistic (unweighted) is more influenced by the ratings on the off-target items (i.e., those much easier or harder than the children's ability).[17] Fit statistics are routinely reported as the MnSq (mean of the squared residuals) as well as standardized Z values (Zstd). In general, infit/outfit MnSq statistics in the range of 0.6–1.4 and their Zstd values ranging from -2 to 2 are regarded as acceptable when evaluating whether items measure the same underlying unidimensional latent trait.[17] Items displaying MnSq >1.4 or Zstd >2 or both indicate that the responses are erratic (i.e., misfitting), or perhaps that the item has been inaccurately scored, or belongs to a different construct. In contrast, an item displaying MnSq <0.6 or Zstd <-2 or both indicates that the item has less variation (i.e., overfit), and therefore, might be redundant for a test. However, such an overfit situation must be considered reasonable for developmental tests measuring the developmental sequence of certain abilities.[17] This current analysis, therefore, focused only on the misfitting items presenting high MnSq or Zstd values or both that might present potential departures from the unidimensionality of the PDMS-2-FM subtests or composite scale, rather than overfitting items. The misfitting items were removed from subsequent analyses in a stepwise manner, and successive Rasch analyses were performed until all remaining items showed acceptable fit statistics. Further, the same fit statistics and criteria were used to examine the extent to which the typically developing children's FM patterns in this study were consistent with the Rasch model's expectations.

Principal component analysis of the residuals[28,29] was conducted to examine further the unidimensionality of both the PDMS-2-FM subtests separately and the composite scale by combing the two subtests. It is expected that after removal of the Rasch measure (principal component) from the data, the residuals for item/person interactions should be randomly distributed and uncorrelated (e.g., explaining $<5\%$ of the variance).[28] That is, there should be no further principal components apart from the one identified by the Rasch model.

The third part of the analysis was to evaluate whether there was DIF in the PDMS-2-FM items across FM status (delayed in the clinical sample *vs.* typically developing in the normative sample) and sex (between boys and girls in the normative group). DIF analysis[17,30] might involve using a scatter plot to compare item calibrations that were computed by separate Rasch analyses for each group. If an item was measurably easier or harder when used with a different group of children (i.e., the item location fell outside the 95% confidence intervals), this is prima facie evidence for lack of measurement invariance across groups, suggesting FM status- or sex-specific items.

Finally, we examined how well the PDMS-2-FM items targeted the normative sample of the study by visually inspecting the Rasch item-person map. In this map, the PDMS-2-FM item difficulties and children's FM abilities are displayed along the same linear interval level measurement continuum, and that allows investigation of whether the PDMS-2-FM items adequately encompass (or target) the range of children's FM abilities in the sample of this study. The item-person map also enables identification of ceiling/floor effects and possible gaps (i.e., few or no items exist to differentiate children at a certain ability level[17]) among the PDMS-2-FM items.

RESULTS

Rating Scale Analysis

Rasch analysis showed that the children's average measures in all PDMS-2-FM items increased monotonically across 3 categories, but a large percentage of the items (63%) had <10 children located in the middle category (score = 1), and a few items (20%) exhibited misfit within the categories. Moreover, more than half of the PDMS-2-FM items (56%) displayed disordered category thresholds (i.e., the threshold between categories 1 and 2 was indicated as less difficult than that of the threshold between categories 0 and 1 for the item). Accordingly, the three rating categories (0-1-2) were reorganized as a dichotomy (0-0-1) by collapsing the middle category and combining it with the lower (0) category. This reorganization was implemented for most items presenting with problematic rating scale use, but not in 7 of the grasping items and 20 of the VMI items (see the notes in Tables 2 and 3 for details). The mix of the three-category and dichotomous rating scales among the PDMS-2-FM items was used in the remaining Rasch analyses.

TABLE 2 Misfitting items in the Rasch analyses of the 26-item grasping subtest ($n = 315$)

Misfitting Items[a] (Item Number: Description)	Infit Statistics		Outfit Statistics	
	MnSq	Zstd	MnSq	Zstd
G16: manipulating paper	*1.62*	1.0	*1.50*	0
G8: pulling string	*2.11*	*2.1*	0.67	0
G22: grasping marker	*1.62*	*3.0*	0.93	−0.1
G2: grasping cloth	*1.64*	0.9	0.11	−0.4

The misfitting items above the G22 were eliminated from the grasping subtest, but the items below the G8 were retained.
Values in italics indicate that the MnSq or Zstd values were beyond the fit criteria, i.e., MnSq >1.4 or Zstd >2.0.
Three items (G1, G3, and G4) were dropped from the Rasch item analysis, because most children passed the items.
Seven items (G14, G16, G19, G21, G22, G24, and G25) used a three-category rating scale whereas the remaining items used a dichotomous scale in the Rasch item analysis.
[a] Items ordered according to the extent of misfit to the Rasch model.

TABLE 3 Misfitting items in the Rasch analyses of the 72-item visual-motor integration subtest ($n = 312$)

Misfitting Items[a] (Item Number: Description)	Infit Statistics		Outfit Statistics	
	MnSq	Zstd	MnSq	Zstd
V67: connecting dots	*1.69*	*4.1*	*1.64*	1.5
V58: lacing string	*1.53*	*2.1*	*3.13*	0.8
V15: transferring cube	*1.79*	*2.4*	0.63	0
V34: tapping spoon	*1.52*	*2.4*	0.86	0
V37: scribbling	*2.58*	*2.2*	0.71	0
V41: turning pages	*1.63*	*2.5*	*2.19*	0.3
V26: removing socks	*1.70*	1.3	*1.79*	0.1
V44: imitating vertical strokes	*1.44*	1.6	*4.46*	0.3
V23: manipulating string	*1.60*	1.9	0.91	0
V13 extending arm	*1.45*	0.6	0.38	−0.2
V61: copying cross	1.27	*2.1*	1.28	0.2
V55: copying circle	1.03	0.1	*9.90*	1.3
V45: removing top	1.33	1.6	*9.90*	0.5
V21: clapping hands	1.15	0.5	*6.40*	0.2

The misfitting items above the V23 were eliminated from the visual-motor integration subtest, but the items below the V44 were retained.

Values in italics indicate that the MnSq or Zstd values were beyond the fit criteria, i.e., MnSq >1.4 or Zstd >2.0.

Four items (V1, V2, V4, and V8) were dropped from the Rasch item analysis, because most children passed the items.

Twenty items (V20, V22, V23, V30, V31, V34, V41, V45, V50, V53, V54, V58, V61, V62, V65, V66, V68, and V70-V72) employed a three-category rating scale whereas the remaining items used a dichotomous scale in the Rasch item analysis.

[a] Items ordered according to the extent of misfit to the Rasch model.

Unidimensionality

With the optimized rating scales in the PDMS-2-FM items, the fit of the current normative sample to the Rasch model's expectations was evaluated before the unidimensionality examination of the PDMS-2-FM items. By using fit statistics, the Rasch analyses revealed 27 (12%), 30 (10%), and 25 (8%) children presenting with infit/outfit MnSq >1.4 or Zstd >2 or both in the grasping subtest, VMI subtest, and FM composite scale, respectively. These children's responses to the PDMS-2-FM items were identified as misfitting to the underlying model, indicating the responses as unexpected and somewhat unpredictably different from those of other children. The misfitting children's response patterns might seriously affect fit at the item level, even though they were part of the normative sample. Thus, these responses were excluded temporarily from the corresponding subtests or composite scale to examine the unidimensionality of the PDMS-2-FM items; however, they were included in the other analyses such as item calibration and DIF.

Tables 2 and 3 show that 4 of the 23 grasping items and 14 of the 68 VMI items had fit statistics that exceeded our predefined infit/outfit MnSq or Zstd values or both. Ten misfitting items were removed from the subtests, except for one grasping item and six VMI items that exhibited marginal misfit (i.e., only one fit statistic exceeding the appropriate criterion). The retention of these items was based on recent recommendations in the literature that the items showing misfit in only one fit statistic could be retained and add value when considering their clinical importance.[17,20] These less-misfitting items were all considered to have potential practical significance for the PDMS-2-FM, and therefore they were retained. Moreover, the item G22-"grasping marker," although showing misfit in two fit statistics, was retained in the grasping subtest, because it was the only item in the gap between 16 and 41 mos of age. The principal component analysis of each subtest that included less-misfitting items further revealed <1% of variance in the residual components, and no obvious clustering of item residuals was found in each subtest. Therefore, the remaining items in each reduced subtest (i.e., 21 grasping items and 60 VMI items), even including the less-misfitting items, were sufficiently unidimensional for the usual PDMS-2-FM purposes. The final item sets were used in all of the following analyses performed for each subtest.

Similarly, we investigated the unidimensionality of the FM composite scale by combining the grasping and VMI subtests. A series of Rasch analyses identified four grasping items and nine VMI items showing misfit on >2 fit statistics (Table 4). The 13 items were accordingly eliminated, but we retained 10 additional items (3 from the grasping subtest and 7 from the VMI subtest) each having had only 1 misfitting statistic. The principal com-

TABLE 4 Misfitting items in the Rasch analyses of the 98-item fine motor composite scale (n = 317)

Misfitting Items[a] (Item Number[b]: Description)	Infit Statistics		Outfit Statistics	
	MnSq	Zstd	MnSq	Zstd
G16: manipulating paper	*2.79*	*4.8*	*9.90*	0.6
V67: connecting dots	*1.65*	*4.1*	*1.56*	1.7
V41: turning pages	*1.50*	*2.1*	*1.89*	0.4
G25: grasping marker[c]	*1.50*	*4.1*	*1.58*	*3.3*
V34: tapping spoon	*1.47*	*2.5*	0.84	0
V44: imitating vertical strokes	*1.43*	1.5	*5.33*	0.5
G22: grasping marker[c]	*1.74*	*4.0*	1.09	0.1
V58: lacing string	*1.50*	*2.2*	*2.86*	0.9
V15: transferring cube	*1.65*	*2.3*	0.59	0
V26: removing socks	*1.46*	1.0	*1.66*	0.1
V23: manipulating string[c]	*1.69*	*2.2*	0.96	0
G21: grasping marker[c]	*1.70*	*2.5*	0.66	0
V37: scribbling	*2.74*	*2.5*	0.79	0
G2: grasping cloth	*1.53*	1.0	0.05	−0.6
V5: regarding hands	*1.53*	0.9	0.06	−0.4
V61: copying cross	1.28	*2.2*	1.29	0.3
V33: placing pegs	0.90	−0.3	*9.90*	0.9
V55: copying circle	0.90	−0.4	*9.90*	1.0
V21: clapping hands	1.23	0.9	*7.73*	0.3
V45: removing top	1.35	1.7	*6.56*	0.3
G13: shaking rattle	1.23	0.5	*2.90*	0.1
V54: building bridge	1.19	0.7	*2.19*	0.1
G26: touching fingers	1.20	1.3	*1.55*	0.9

The misfitting items above the G2 were eliminated from the fine motor composite scale, but the items below the V37 were retained.

Values in italics indicate that the MnSq or Zstd values were beyond the fit criteria, i.e., MnSq >1.4 or Zstd >2.0.

Seven items (G1, G3, G4, V1, V2, V4, and V8) were dropped from the Rasch item analysis, because most children passed the items.

[a] Items ordered according to the extent of misfit to the Rasch model.

[b] G denotes the grasping subtest and V denotes the visual-motor integration subtest.

[c] Items that did not fit the fine motor composite scale but did fit the corresponding grasping or visual-motor integration scale.

ponent analysis revealed little common variance (<1%) among the residuals of the remaining items in the FM composite scale. This indicated that these items (including 19 grasping items and 59 VMI items) had acceptable fit to the Rasch model's requirements, substantiating the unidimensionality of the reduced FM composite scale. All the additional results of the FM composite scale were based on this final reduced item set.

Differential Item Functioning

Seven items (i.e., four grasping and three VMI items) were found to fall outside the confidence intervals in the DIF scatter plots displaying the item calibrations of typically developing children, plotted against those of children with FM problems. The four grasping items (G13-"shaking rattle," G20-"grasping cubes," G21-"grasping marker," and G23-"unbuttoning buttons") and three VMI items (V49-"stringing beads," V64-"dropping pellets," and V72-"folding papers") were thereby identified as showing FM status-specific lack of measurement invariance. Likewise, the DIF analysis found seven grasping items (G5-"holding rattle," G18-"grasping pellets," G20-"grasping cubes," G22- and G25-"grasping maker," G23-"unbuttoning buttons," and G24-"buttoning buttons") and three VMI items (V3-"placing hand," V13-"eextending arm," and V18-"poking finger") with DIF by sex, indicating that these items had measurably different levels of difficulty across sex.

Targeting

Figure 1 shows the distribution of the item and person measures plotted along the same axis, with the harder items to complete and children with better FM ability located closer to the top of the latent trait. For convenience, the PDMS-2-FM subtests and the composite scale are shown in the same figure. The mean logit measures (M) of the children in the normative sample and the mean of the items were separated by extremely large ranges of 10–13 logits for each subtest or composite scale. These large differences between the mean sample

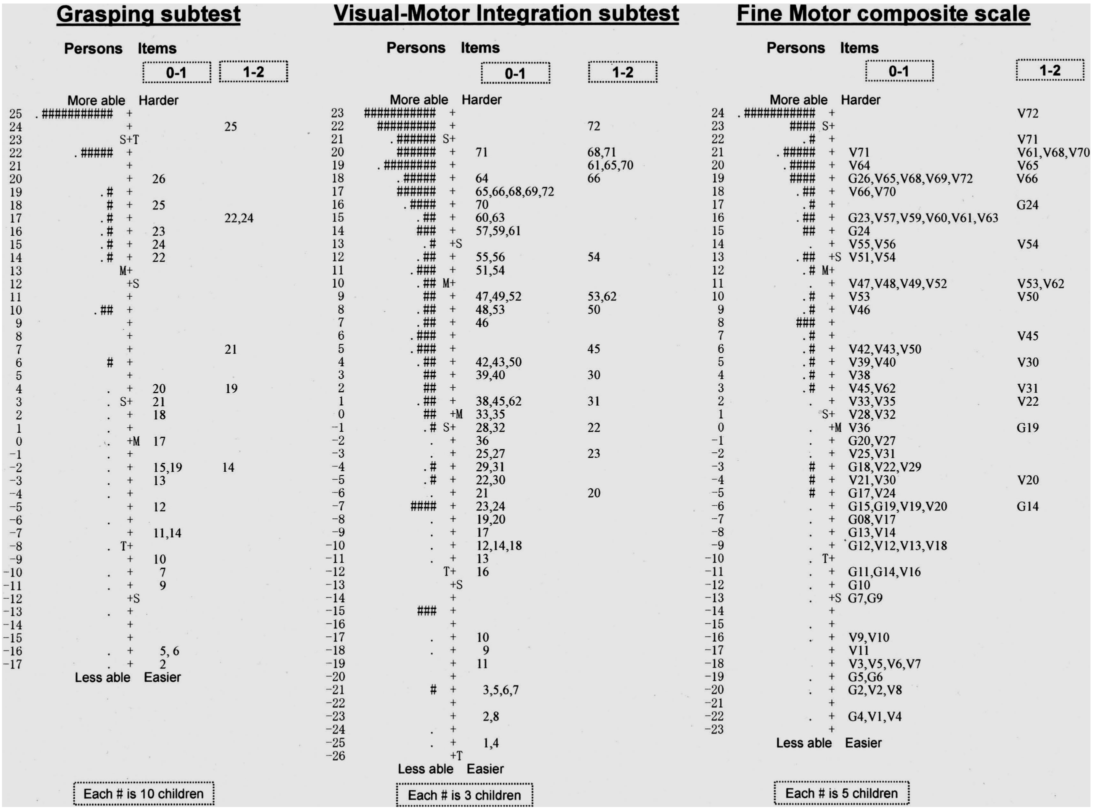

FIGURE 1 *Item-person maps for the grasping and visual-motor integration subtests and the fine motor composite scale: person ability measures in relation to item difficulty calibrations including step calibration (0–1 and 1–2) for the rating scales. Higher measures indicate higher person ability and higher item difficulty. M denotes mean value, S denotes one standard deviation from the mean, and T denotes 2 standard deviations from the mean.*

and item mean measures indicate inadequate item/person targeting. In addition, Figure 1 shows that there were ceiling effects: in the grasping subtest (34.8% of the sample scored at the top of the subtest); in the VMI subtest (9.6%); and in the FM composite scale (9.4%), but no significant floor effect was identified. The item-person maps also show possible gaps between items G21 and G22 and between items G6 and G9 in the grasping subtest and between items V10 and V16 in the VMI subtests, but the FM composite scale presents as gap-free.

In addition, the Rasch item and person reliability indices (which are interpreted similarly to Cronbach's α) for each subtest and the FM composite scale were >0.95. The item and person separation indices for each subtest and the FM composite scale ranged from 4.4 to 19.6 (more than a commonly accepted criterion of 3.0[17,31]). The values indicate that the PDMS-2-FM items cover a wide range of difficulty levels that would be appropriate for measuring children with a wide range of FM abilities. Furthermore, the children's Rasch ability measures in the PDMS-2-FM subtests and composite scale correlated significantly with their chronological age (Pearson's $r > 0.92$), reflecting that the children's test scores increase progressively with age.

DISCUSSION

The PDMS-2-FM is a widely used clinical and research instrument in rehabilitation and early intervention settings, but its measurement properties are yet to be thoroughly explored. This study was the first to use the Rasch model to investigate rating scale performance, unidimensionality, and possible DIF in items of the PDMS-2-FM. Overall, the rating scale performance for most of the PDMS-2-FM items was improved by collapsing the middle categories. Unidimensionality of the grasping and VMI subtests could be achieved by removing a limited number of items and, consequently, most items in the composite FM scale contributed to the measurement of a single construct. The analysis identified a few items that showed DIF across sex or FM status. Ceiling effects and inadequate targeting were found when the PDMS-2-FM subtests and its composite scale were applied to these Taiwanese children. In all, the findings provide information to support further PDMS-2-FM revision or development of a subsequent edition, in terms of the number of response categories and removal/revision of misfitting and DIF items. Furthermore, the addition of more challenging items would help to ameliorate the ceiling effects of the PDMS-2-FM.

There is no convincing evidence for the appropriateness of the three-category response scales of the PDMS-2-FM items. We found that the middle categories of most items provided little information about the children in the combined sample and were therefore redundant. For this reason, the problematic rating categories of these items were simplified to dichotomous categories. Although we retained good rating scale performances for the remaining items thereby supporting the PDMS-2-FM developers' original intention in which the middle categories may capture progressive change of children with FM delays. The combined use of two- and three-point rating scales, however, might be disadvantageous for ease of PDMS-2-FM administration. It is, therefore, suggested that future studies examine the clinical impact of combining the two response formats. Alternatively, the scoring criteria of the items showing problematic three-point rating scales should be refined in a way that can differentiate children's performances more clearly. In addition, future studies that recruit a suitably large group of children with FM delays are recommended to re-examine the appropriateness of the three-point rating scales with such participants. This recommendation is important because the combined sample of this study comprised many more typically developing children who are likely to have proceeded more capably and quickly toward FM mastery than would children with FM delays. It might be that the three-category scoring in certain items is somewhat redundant for only typically developing children.

The results of the Rasch analysis indicated that most of the PDMS-2-FM items in individual subtests and composite scale assessed the intended constructs and were essentially unidimensional. However, there exists some inconsistencies with regard to the item fit results of the individual subtests and the FM composite scale. Four items were removed from the FM composite scale because of misfit, but they were not eliminated in the individual subtests (Table 4). Three of the four items were the series of "grasping marker" items (G21, G22, and G25) in the grasping subtest and one was the item V23-"manipulating string" in the VMI subtest. In contrast, there was one grasping item (i.e., G8-"pulling string") that was retained in the FM composite scale but not the grasping subtest. This finding implied that the FM composite scale, although being a broader construct, may not cover all of the items in the two narrower individual constructs or be automatically constructed from the combination of the narrower individual constructs. Nine items, however, were found to exhibit misfit consistently in both the individual subtests and in the FM composite scale, indicating that these items contributed to neither the individual constructs nor the overall FM construct.

There were some possible reasons why the nine items demonstrated serious misfit to the rest of the items in both the composite scale and corresponding subtests. First, these misfitting items might reflect a different construct from the VMI/grasping and FM abilities, or have influences confounded by other aspects (e.g., movement experience) rather than the targeted ability per se. For example, a child who had less or no experience with lacing a string through a six-hole strip might not perform the item V58-"lacing string" in a manner consistent with that child's overall ability on the PDMS-2-FM. Second, ambiguous scoring criteria might present an additional reason for items showing misfit. In the item V37-"scribbling," for example, a child who makes at least one scribble >1-inch long obtains a full score. However, it was frequently observed that 9- to 12-mo-old children were able to make a 1-inch long scribble by accident, e.g., postural reflex movements. Because of this less challenging criterion, the false, unintentional performance might receive credit when it should not; thereby, contributing to item misfit. Third, items that are difficult to administer/facilitate might be identified as misfitting. In particular, we found that some children younger than 2 yrs did not comply to the examiner's demonstration and verbal instructions on the item V15-"transferring cubes," item V26-"removing socks," and item V34-"tapping spoon" in the VMI subtest. The younger children tended to behave in their own quotidian ways, rather than producing the desired responses. In this case, the performances that did not comply with the instructions for these items were instead scored as no score, according to the test manual. However, these substitute scores might not faithfully indicate children's true performance on the items, thereby contributing to potential item misfit. Accordingly, we suggest that future researchers might need to revise the instructions or demands for some of the misfitting items to match the level appropriate to the development of younger children. Review of scoring criteria is also suggested to minimize children's false positive or unintentional performances. Furthermore, providing up to five practice trials during administration of some of the misfitting items might allow less-experienced children to achieve performances more indicative of their actual underlying abilities.

The DIF results demonstrated only a small number of items (ten items) showing DIF across sex, generally in accordance with the test developers' findings[6] (five items). Given that sex differences are expected in children's motor development,[2] sex DIF items were not eliminated from the PDMS-2-FM in this study. However, the grasping subtest exhibited quite a large percentage (29%) of DIF items. Therefore, the content of these grasping items might need further review and, if no revision is made to the items, splitting sex-DIF items into boys and girls for analysis might be a way of accounting for sex DIF on the grasping subtest. This study also used DIF analysis to identify FM status-specific PDMS-2-FM items. We found seven items with DIF across FM status (delayed *vs.* not delayed), indicating that those items had largely different levels of difficulty between children with/without FM delays. In other words, children with FM delays might perform poorly in specific activities with these DIF items. Because the FM status-specific items are likely to be useful indicators for screening children with FM delays, we suggest their retention in the PDMS-2-FM. However, because of the small sample size and diverse conditions of the clinical sample used, future studies that recruit larger and more carefully specified clinical groups are warranted to examine the current DIF finding.

East Asian children have been found to demonstrate more advanced grasping patterns or manual dexterity than do their Western counterparts.[22,23] The ceiling effects and inadequate targeting of the PDMS-2-FM in this study may have resulted from some cultural influence on children's FM abilities. The American-developed PDMS-2-FM items were found to be quite easy for the Taiwanese children and would be suitable only for younger children or those with FM problems in Taiwan. Some possible item gaps were revealed in each of the individual subtests. For these reasons, future studies are needed to avoid the problems by adding more challenging, suitable items to the PDMS-2-FM.

The current study had three specific limitations. First, seven items (three from the grasping subtest and four from the VMI subtest) were dropped following item fit analyses, because most children in our normative sample managed to pass these items. These discarded items were originally designed for children within an age range of birth to 12 mos. Although this study included a reasonable number of 48 children in this age bracket, future studies that recruit more children aged <12 mos are needed to enable the results of item fit analysis for discarded items to be substantiated. Second, most of the misfitting items were removed from the PDMS-2-FM subtests or composite scale, but some were retained because they possess potential clinical benefits and were judged as not seriously affecting unidimensionality. More replicate studies are warranted to explore if the retained misfitting items contribute substantially to the unidimensionality of the PDMS-2-FM. Third, Rasch measurement analysis is claimed to be more stringent than the item response theory analysis[6] used to examine the PDMS-2-FM by the test developers.[17] However, as their detailed reports and statistics were not presented in the test manual,[8] it is impossible to undertake an in-depth comparison between their results and those obtained in this study.

CONCLUSIONS

This study provides preliminary evidence that simplifying rating scales of most PDMS-2-FM items to dichotomous categories could optimize the measurement qualities of the PDMS-2-FM scaling. Unidimensional grasping and VMI subtests and the FM composite scale with the optimized rating scales might be substantiated after removal of a few items. Although the reduced PDMS-2-FM with reorganized rating scales cannot be used immediately in the place of a suitably norm-referenced test, this revised instrument effectively reflects a unidimensional construct of children's FM ability, and children's test scores increase progressively with age. Moreover, most item calibrations of the reduced PDMS-2-FM are independent of sex and FM status. Therefore, the findings support that the reduced PDMS-2-FM could be used as a criterion-referenced outcome measure to document changes of children's FM ability in clinical and research settings. Further studies could apply the reduced PDMS-2-FM to a group of children with FM delays to determine whether it describes the progression of children's delayed FM abilities appropriately. Continued work is also needed to explore the addition of more challenging items suitable to compensate for the ceiling effects and item gaps.

ACKNOWLEDGMENTS

We thank Ms. Vanessa Carson for her assistance in editing this manuscript before submission. Support and comments from Professor Nola Alloway at James Cook University and Professor Ching-Lin Hsieh at National Taiwan University have been central to the success of this project.

REFERENCES

1. Edwards SL, Sarwark JF: Infant and child motor development. *Clin Orthop Relat Res* 2005;434:33–9
2. Henderson A, Pehoski C: *Hand Function in the Child: Foundations for Remediation*, ed 2. Philadelphia, PA, Mosby Elsevier, 2006
3. Shonkoff JP, Hauser-Cram P: Early intervention for disabled infants and their families: A quantitative analysis. *Pediatrics* 1987;80:650–8
4. Blauw-Hosper CH, Hadders-Algra M: A systematic review of the effects of early intervention on motor development. *Dev Med Child Neurol* 2005;47:421–32

5. Burton AW, Miller DE: *Movement Skill Assessment*. Champaign, IL, Human Kinetics, 1998

6. Folio MR, Fewell RR: *Peabody Developmental Motor Scales: Examiner's Manual*, ed 2. Austin, TX, PRO-ED, 2000

7. Burtner PA, McMain MP, Crowe TK: Survey of occupational therapy practitioners in southwestern schools: Assessments used and preparation of students for school-based practice. *Phys Occup Ther Pediatr* 2002;22:25–39

8. Wang HH, Liao HF, Hsieh CL: Reliability, sensitivity to change, and responsiveness of the Peabody developmental motor scales-second edition for children with cerebral palsy. *Phys Ther* 2006;86:1351–9

9. van Hartingsveldt MJ, Cup EH, Oostendorp RA: Reliability and validity of the fine motor scale of the Peabody developmental motor scales-2. *Occup Ther Int* 2005;12:1–13

10. Darrah J, Magill-Evans J, Volden J, et al: Scores of typically developing children on the Peabody developmental motor scales: Infancy to preschool. *Phys Occup Ther Pediatr* 2007;27:5–19

11. Provost B, Heimerl S, McLain C, et al: Concurrent validity of the Bayley scales of infant development II motor scale and the Peabody developmental motor scales-2 in children with developmental delays. *Pediatr Phys Ther* 2004;16:149–56

12. American Educational Research Association, American Psychological Association, National Council on Measurement in Education: *Standards for Educational and Psychological Testing*. Washington, DC, American Educational Research Association, 1999

13. Zou KH, Tuncali K, Silverman SG: Correlation and simple linear regression. *Radiology* 2003;227: 617–22

14. Waugh RF, Chapman ES: An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *J Appl Meas* 2005;6:80–99

15. Rasch G: *Probabilistic Models for Some Intelligent and Attainment Tests*. Copenhagen, Dermark, Dermarks Paedagogiske Institute, 1960

16. Chen CC, Bode RK, Granger CV, et al: Psychometric properties and developmental differences in children's ADL item hierarchy: A study of the WeeFIM instrument. *Am J Phys Med Rehabil* 2005;84:671–9

17. Bond TG, Fox CM: *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. ed 2. Mahwah, NJ, Lawrence Erlbaum Associates, 2007

18. Gallagher P, Horgan O, Franchignoni F, et al: Body image in people with lower-limb amputation: A Rasch analysis of the amputee body image scale. *Am J Phys Med Rehabil* 2007;86:205–15

19. Fujiwara T, Liu M, Tsuji T, et al: Development of a new measure to assess trunk impairment after stroke (trunk impairment scale): Its psychometric properties. *Am J Phys Med Rehabil* 2004;83:681–8

20. Krumlinde-Sundholm L, Holmefur M, Kottorp A, et al: The assisting hand assessment: Current evidence of validity, reliability, and responsiveness to change. *Dev Med Child Neurol* 2007;49:259–64

21. Arnould C, Penta M, Renders A, et al: ABILHAND-kids: A measure of manual ability in children with cerebral palsy. *Neurology* 2004;63:1045–52

22. Chow SM, Henderson SE, Barnett AL: The movement assessment battery for children: A comparison of 4-year-old to 6-year-old children from Hong Kong and the United States. *Am J Occup Ther* 2001;55: 55–61

23. Tseng MH: Development of pencil grip position in preschool children. *Occup Ther J Res* 1998;18: 207–24

24. Hanna SE, Law MC, Rosenbaum PL, et al: Development of hand function among children with cerebral palsy: Growth curve analysis for ages 16 to 70 months. *Dev Med Child Neurol* 2003;45:448–55

25. Linacre JM: *A Users Guide to Winsteps and Ministeps Rasch Model Computer Programs*. Chicago, IL, Winsteps, 2006

26. Wright BD, Masters GN: *Rating Scale Analysis: Rasch Measurement*. Chicago, IL, MESA Press, 1982

27. Linacre JM: Optimizing rating scale category effectiveness. *J Appl Meas* 2002;3:85–106

28. Smith EV Jr: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002;3:205–31

29. Linacre JM: Detecting multidimensionality: Which residual data-type works best? *J Outcome Meas* 1998;2:266–83

30. Holland PW, Wainer H: *Differential Item Functioning*. Hillsdale, NJ, Erlbaum, 1993

31. Duncan PW, Bode RK, Lai SM, et al: Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. *Arch Phys Med Rehabil* 2003;84: 950–63