JOURNAL OF APPLIED MEASUREMENT, *3*(2), 190–204
Copyright© 2002

Development of a Functional Movement Scale for Infants

Suzann K. Campbell
University of Illinois at Chicago

Benjamin D. Wright

J. Michael Linacre
University of Chicago

The increasing survival rate of infants with a complicated birth and perinatal history generated the need for a test of functional motor performance with the capability of identifying children under four months of age with delayed development which could be addressed with physical therapy. This paper describes a Rasch analysis of the psychometric qualities of the Test of Infant Motor Performance (TIMP) for the purpose of reducing the length of the test while maintaining its precision as a measurement device. Following analysis of fit statistics, item-to-total correlations, redundancy of item difficulty measures, and consideration of clinically-relevant features of test items from analysis of 1732 tests, the TIMP was reduced from 59 to 42 items forming a functional motor scale for prematurely born infants. The resulting person separation index was 4.85 and the item separation index was 23.79.

Requests for reprints should be sent to Suzann K. Campbell, Department of Physical Therapy, University of Illinois at Chicago, 1919 W. Taylor Street M/C 898, Chicago, IL 60612-7251, email: skc@uic.edu.

An increasing number of infants with perinatal complications now survive as a result of advanced technology and caregiving interventions provided in special care nurseries. The developmental morbidity associated with this trend, however, appears to be unchanging (Fanaroff, Wright, Stevenson, Shankaran, Donovan, Ehrenkrans, Younes, Korones, Stoll, Tyson, Bauer, Oh, Lemons, Papile, and Verter, 1995). For example, infants with brain insults (Pinto-Martin, Whitaker, Feldman, Van Rossem, and Paneth, 2000) or chronic lung disease (Majnemer, Riley, Shevell, Birnbaum, Greenstone, and Coates, 2000), and those born extremely early tend to have delayed motor development (McCormick, McCarton, Tonascia, and Brooks-Gunn, 1993). Improving the accuracy with which infants at risk for disability are identified remains a challenge, in part because infants can recover from some neurodevelopmental abnormalities during the first year of life (Wildin, Smith, Anderson, Swank, Denson, and Landry, 1997). The purpose of this paper is to describe the use of Rasch analysis in the development of a test of functional motor performance intended to identify infants whose movement capabilities demonstrate delayed development. The specific goal of this analysis was to reduce the length of the test.

The Test of Infant Motor Performance

The Test of Infant Motor Performance (TIMP) is a 25-35 minute assessment of the posture and selective control of movement needed by infants under four months of age for functional performance in daily life. The TIMP was developed to 1) identify infants with delayed motor development, 2) discriminate among infants with varying degrees of risk for poor motor outcome, and 3) measure change resulting from intervention. The TIMP has gone through three research versions to date.

Version 1 of the TIMP was developed by Girolami for use in assessing the effects on posture and movement of physical therapy provided to premature infants at risk for movement dysfunction in a special care nursery (Girolami and Campbell, 1994). The test included fifteen dichotomous items that were scored on the basis of observing infants' spontaneous movements. An example of these Observed Items is one used to assess the infant's ability to hold the head centered along the midline of the body, i.e., keep the head from falling to the side while backlying. Twenty-eight additional Elicited Items were scored on Likert-type scales reflecting the infant's responses to being placed in various positions or stimulated with interesting sights and sounds, such as a rattle or the examiner's voice. Results of a controlled clinical trial demonstrated that the TIMP was a valid tool for capturing the effects of intervention (Girolami and Campbell, 1994).

Based on the sensitivity of TIMP scores for detection of intervention effects, the decision was made to expand the test in order to increase the age range to include performance of premature infants as young as 32 weeks postconceptional age, i.e., 2 months prior to expected date of birth, as well as infants up to 3-4 months of age (chronologic or corrected for premature delivery). The resulting Version 2 of the TIMP had 27 dichotomous Observed Items and 30 Elicited Items scored with rating scales. The rating scales for some TIMP items allow one to assess whether or how much of an activity an infant can do while others are used to assess changes in how an infant solves a movement problem, i.e., changing patterns of movement synergies used in response to stimulation.

With funding from the Foundation for Physical Therapy, TIMP Version 2 was studied in a cross-sectional sample of 137 infants from three race/ethnicity groups in the Chicago metropolitan area: non-Latino/a white, black (African or African-American), and Latino/a (Mexican or Puerto Rican) (Campbell, Osten, Kolobe, and Fisher, 1993; Campbell, Kolobe, Osten, Lenke, and Girolami, 1995). Data from 174 tests on these subjects was analyzed with BIGSTEPS V.2.2. Clarity of the measure as reflected in the item separation index was 7.38 (root mean square error=0.19) with a separation reliability of .98. The items separated the infants into about 6 levels of ability (person separate index of 6.02 with root mean square error=0.21) which was thought to be excellent in view of the 5-6-month age range of the test. TIMP performance correlated with age at r = .83 (Campbell, Kolobe, Osten, Lenke, and Girolami, 1995).

After study of Rasch analysis results on TIMP Version 2 and a review of new literature on possible predictors of abnormal outcome, the test developers honed their theoretical understanding of what the TIMP should accomplish and made several new changes to the test. As a result, Version 3 of the TIMP was developed which included 28 Observed Items and 31 Elicited Items, 6 of which were the same items but used to test different sides of the body so that asymmetry of movement could be assessed.

Various aspects of the TIMP's validity were explored in research by graduate students or with funding from the National Center for Medical Rehabilitation Research of the National Institutes of Health. V.3 of the TIMP was shown to 1) relate to movement demands on infants provided by caretakers in naturalistic interactions such as bathing, dressing, and play (Murney and Campbell, 1998), 2) be responsive to changes produced by maturation and to differentiate among groups of infants with differing degrees of risk for poor motor outcome based on medical conditions (Campbell and Hedeker, 2001), and 3) predict 12-month outcome on the Alberta Infant Motor Scale with high sensitivity and specificity (Campbell, Kolobe, Wright, and Linacre, in press). Prediction to 5-year outcome is currently under study (Kolobe and Bulanda, personal communication).

Despite strong support for the proposed uses of the TIMP resulting from our research, a shorter tool was desirable for practical clinical use. The purpose of this paper is to present evidence from Rasch analysis of TIMP V.3 data and to describe how Rasch item fit statistics and other considerations were used to shorten the test to create Version 4.

Methods

Subjects

The subjects in this study were a sample of convenience born during the years 1996-98 and recruited from the special care nurseries of three hospitals or from the community within the Chicago metropolitan area. Subject recruitment methods were approved by the Institutional Review Board for the protection of the rights of human subjects at the University of Illinois at Chicago (#H-99-1158) and at each field testing site. Subjects were 159 infants with a range of medical complications who participated either in a study of test-retest reliability over the space of three days (n=56) or in both the test-retest and a longitudinal study (n=103) of performance on the TIMP. Fifty-five percent of the infants were male, 42% female (3% had missing data on sex). The distribution of race/ethnicity was 38% white, non-Latino/a; 31% black (African or African-American); 26% Latino/a; and the rest Asian (1%), mixed (1%), or missing (3%).

Procedures

With the informed consent of a parent to test an eligible subject and medical clearance from the infant's physician, the infants in the longitudinal study were scheduled for testing with the TIMP every week until approximately 4 months corrected age. The number of weekly tests each infant received ranged from 2-23. Test numbers varied because of 1) age and health status of the infant at recruitment, 2) illness, and 3) family scheduling conflicts. In the test-retest study, infants were tested twice within the space of 3 days.

At initial testing, infants were required to be off mechanical ventilation and cleared for testing by their physician, but could be receiving oxygen by nasal canula. Thus infants began testing at different ages based on health. The infant was tested in its current environment: isolette with vital signs monitors in place, open crib, home, or occasionally during an outpatient clinic visit. Testers were not told the age or medical history of the infant before testing (unless information was needed to guarantee safe handling of an infant during assessment). Testing occurred about one hour prior to expected feeding time for preterm infants or about mid-way between feedings for older infants.

Twelve testers participated in this study. Each tester had experienced a period of training in use of the TIMP which consisted of a 4-hour workshop on development and validation of the test and how to score it, independent reading of research on the test, practice in testing at least 10 infants of varying ages, and rating of item performances on 14 videotapes of infants of different ages with or without a variety of medical complications. Rater consistency was evaluated with the Facets computer program for Rasch psychometric analysis (Linacre, 1988); raters needed to have fewer than 5% misfitting ratings, i.e., unexpected ratings given the infant's level of ability on an item and the item difficulty, in order to qualify for being a tester in this study.

Data Analysis

Scores on the TIMP were subjected to Rasch analysis using the BIGSTEPS computer program Ver. 2.65 in order to transform the raw ordinal scores into interval-level logit measures (Wright and Linacre, 1996; Wright and Masters, 1982). According to the Rasch model, the probability of passing an item is based only on the ability of the subject and the difficulty of the item and its various scale levels. Analysis yields both population-independent estimates of item parameters, and individual ability estimates for the latent trait being measured (Hambleton, 2000), in this case functional motor performance in early infancy.

The Rasch analytical rating scale model was the Andrich model, but with each Elicited Item conceptualized as having its own rating scale structure as in the Master's partial credit model. The partial credit model was selected because the rating scale for each Elicited Item is unique and the number of levels varies from 5 to 7. Given their dichotomous nature, Observed Items were assumed to form a group as a whole. BIGSTEPS begins with a central estimate for each person measure, item calibration and rating scale step calibration. An iterative version of the normal approximation algorithm is used to reach a rough convergence to the observed data pattern. The unconditional maximum likelihood method, using proportional curve fitting, is then iterated to obtain more exact estimates, standard errors and fit statistics. The scale mean for item difficulty was set to 50 with one logit (log-odds unit) equal to 10 points.

Item fit to the Rasch model was investigated using infit (information-weighted fit statistic which is sensitive to unexpected behavior affecting responses to items near the person's ability measure) and outfit (an outlier-sensitive fit statistic sensitive to unexpected behavior by persons on items far from the person's ability level) measures. Fit statistics are reported as mean square residuals which have approximate chi-square distributions. In BIGSTEPS values of standardized fit statistics are obtained from the squared residuals by means of an asymptotic normal distribution (Windmeijer, 1990). Mean-square fit statistics of 1.2 or greater were used as the cutoff for identifying misfitting items for further evaluation. Values below one were not addressed in the analysis.

Items with unsatisfactory fit to the Rasch model were reviewed for the possibility of removal from the test or revision to improve item fit. Before consideration of removing any misfitting item from the test, several factors other than Rasch results were also considered. Elicited Items were compared with data on frequency of occurrence as a naturalistic demand for movement during caregiver-infant interactions so that no item believed to be of functional significance in daily life would be removed without considering whether the item could instead be revised to improve item response characteristics. Second, how well therapists and parents liked these items was considered; item deletion was supported by finding that the items were hard to administer, seemed particularly demanding for fragile infants, or otherwise were difficult for parents to understand based on therapists' impressions of parent and child responses. A final consideration was whether removal of an item would be likely to decrease test precision because its item difficulty was not duplicated by other items at the same level of scale performance. A finding of redundancy in difficulty with another item was also grounds for consideration of deleting an item. A goal of the analysis was to reduce the number of items, if possible, in order to minimize infant stress and fatigue and to increase the practicality of using the test by reducing the time required to administer it.

Results

Testing in the longitudinal study and the test-retest reliability study resulted in data from 1723 tests crossing the entire range of age from 32 weeks postconceptional age through 4 months post-term. Only the first test of a test-retest pair was used in the data analysis. Tests from 1719 non-extreme scoring individuals were assessed. The mean Rasch measure was 55.43 with a SD of 10.49 and mean SE of 1.98 with a SD of .45. Overall infit mean square for persons was .98 and outfit mean square was 1.10. The person separation index was 4.57 with a reliability of .95.

A wide range of item difficulty was achieved with an item separation index of 21.37 and reliability of 1.00. The mean item measure (as set for the analysis) was 50.00 with a SD of 11.92, model error of .48 with a SD of .23. The third column from the left in Figure 1 shows that the average item difficulties ranged from 20 to 83 (D represents dichotomously scored Observed Items and X's represent scaled Elicited Items). Calibrations for the easiest step on Elicited Items extend the range of the scale down to 10 (Fig. 1, second column) while calibrations for the hardest step extend the scale up to 85 (Fig. 1, fourth column). Point biserial correlations for items ranged from -.10 to .78 with 29 of 59 (49%) greater than .50. Infit mean square for items was 1.01 with an outfit of 1.21.

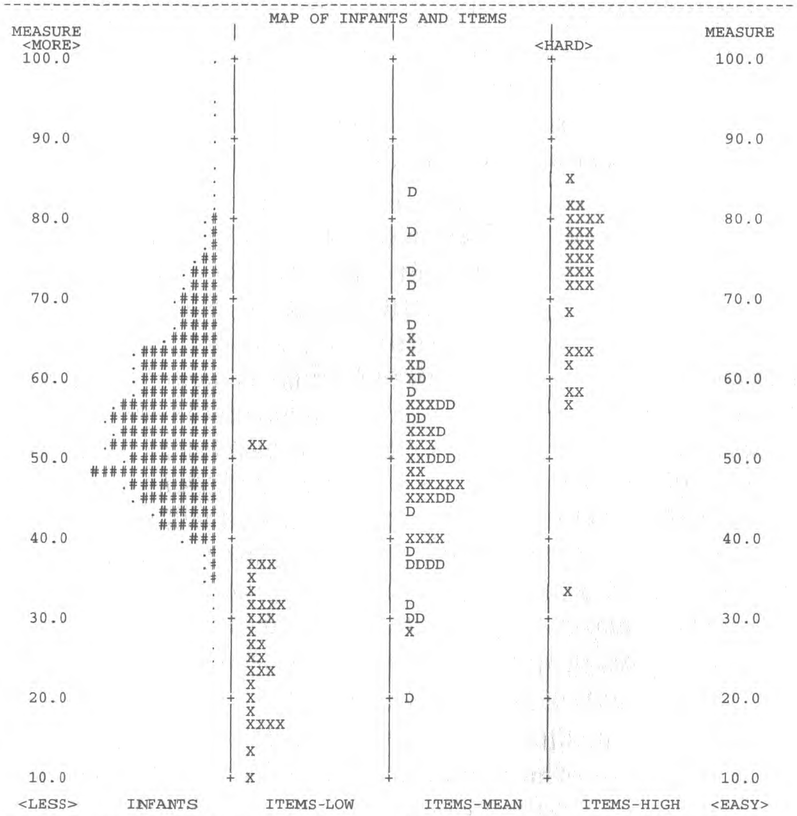
Table 1

*Difficulty Calibrations, Infit and Outfit Mean Square Values, and Point Biserial Correlations for Items Misfitting the Rasch Model**

ITEM	DIFFICULTY	INFIT MNSQ	OUTFIT MNSQ	PT. BISERIAL
O2 Head turn L	55	1.13	**1.23**	.30
O3 Head turn R	55	1.15	**1.24**	.29
O4 Hands together	58	1.19	**1.27**	.24
O5 R Hand to mouth	46	**1.45**	**2.23**	-.10
O6 L Hand to mouth	50	**1.44**	**1.89**	-.03
O7 Mouths R hand	72	1.19	**1.38**	.19
O8 Mouths L hand	73	1.17	**1.36**	.20
O9 Individual R finger movements	37	**1.23**	**1.71**	.05
O10 Individual L finger movements	39	**1.26**	**1.78**	.02
O11 R wrist movements	37	1.11	**1.29**	.18
O12 L wrist movements	37	1.11	**1.30**	.18
O15 Pelvic lift in supine	43	**1.23**	**1.55**	.12
O16 Hip/knee flexion	20	1.04	**1.38**	.10
O17 R ankle movements	29	1.06	**1.28**	.17
O18 L ankle movements	30	1.06	**1.31**	.17
O23 Arm movements with forearm off surface	53	**1.33**	**1.47**	.10
O24 Arm movements with upper arm off surface	45	**1.26**	**1.46**	.12
E3 Straighten spine with head supported	47	**1.55**	**2.26**	.40
E11 R neck rotation	47	1.06	**1.35**	.59
E13 Defensive head movements	28	**1.46**	**7.90**	.10
E14 Defensive arm movements	61	**1.31**	**1.35**	.56
E28 Arm release in prone	57	**1.58**	**1.53**	.49
E29 Standing	62	**1.21**	1.17	.58

* Bold lettering indicates value exceeding the targeted range for item fit statistic of 1.2.

Information on misfitting items is presented in Table 1. The identified items had difficulty values ranging from 20 to 73 and point biserial correlations ranging from -.10 to .59. Review of the infit mean square data for each item revealed 12 items with infit greater than 1.2: 5 Elicited Items (E3, E13, E14, E28, and E29) and 7 Observed Items (O5, O6, O9,



EACH '#' IN THE INFANT COLUMN IS 10 INFANTS; EACH '.' IS 1 TO 9 INFANTS

Figure 1. TIMP BIGSTEPS—Version 3 Item Scaling Properties

O10, O15, O23, and O24). Of the 12, 11 items also had unacceptable outfit statistics. In addition, 11 items—O2, O3, O4, O7, O8, O11, O12, O16, O17, O18, and E11—had outfit values greater than 1.2.

Based on these results, the following Observed Items were deleted in forming V.4 of the TIMP: O2 through O8, O11, O12, O15, O23, and O24. With the exception of O7 and O8, other items with similar average difficulty levels as the items deleted were already available in the scale. With respect to O7 and O8 with difficulty measures of 72 and 73, respectively, although other items of similar average difficulty are lacking in this ability region, the rightmost column of Figure 1 shows that several Elicited Items have their highest step level in this region. Thus precision at the higher ability levels should be maintained without O7 and O8. Finally, point biserial correlations for all removed Observed Items were low (<.31).

Items O20, O21, and O22 were also deleted, although their fit statistics were satisfactory, because during testing they engender a position change from supine to prone that was an undesirable disturbance during spontaneous movement observation, and because they were redundant in difficulty with reliably obtained Elicited Items E24, E26, and E27 which assess the same movements under conditions of stimulation.

Of the misfitting Observed Items, only O9, O10, O16, O17, and O18 were retained. O9 and O10 are used to assess the ability to perform individual finger movements. Their misfit values were just above 1.2 (1.23 and 1.26, respectively). They were relatively easy movements for infants to display (difficulty 37 and 39, respectively) in an area of the scale with few items of similar average difficulty so deleting them might reduce the precision of the scale for the lowest functioning infants. In addition these delicate finger movements were items that fascinated parents when pointed out by testers. Observed Item O16 is used to measure kicking and was the easiest item in the test with no other items of similar difficulty. O16's infit was satisfactory and kicking is a fundamental movement skill that the authors believed should be retained in the test. O17 and O18 are used to assess isolated movements of the right and left ankle, respectively, again easy items that were not redundant in difficulty with other items. Infit values were satisfactory so the items were retained.

Of the five Elicited Items selected for possible revision or deletion from the TIMP because of high infit values, E3 and E28 were removed. In addition to support for deleting these items based on misfit statistics, both were items deemed by therapists to be ones families liked less well than others and both often resulted in crying. Although the point biserial value for E3 was moderate at .40, neither item was identified as being similar to caretaker demands for movement in Murney and Campbell's study (1998) of the ecologic relevance of TIMP Elicited Items. Other items were available with similar difficulties so removal did not seem likely to affect the precision of the test.

Items E13, E14, and E29 were not removed but were instead revised, in part because each of them was found to occur moderately or frequently often in naturalistic interactions (Murney and Campbell, 1998). Problems with E13, one of the easiest items in the test, were believed to come from infrequent use of several response categories and little variation in infant performance, leading to an exceptionally high outfit value (7.90) and poor point biserial value (.10). Rather than removing the item from the test, it was revised by collapsing two levels, making the item a 4-level rather than 5-level item, and by changing the time allowed for the response to occur to make it a more difficult item.

E14 had a point biserial correlation of .56 and was not redundant with other items in average difficulty. The item, therefore, was revised to combine the old steps 2 and 3 into one step, leaving a 5- rather than 6-level item.

E29 is a relatively difficult item (difficulty measure 62) that is used to assess supported standing, an activity therapists believed could not be removed from the test because virtually all infants are placed in supported standing by their parents (91% in the study of ecologic relevance of TIMP Elicited Items), making it important in daily life interactions. E29's point biserial correlation of .58 also favored keeping the item but revising it to better reflect what infants actually did during item administration. To accomplish this task, videoclips of the standing item from 55 infant observations were reviewed by an experienced tester. She proposed changes to the item that reflected the infant performances caught on videotape and the item descriptions were revised.

Finally, item E11 with a poor outfit was not removed because another item used to test the same activity on the opposite side of the body (E12) had excellent fit statistics and E11's point biserial correlation of .59 favored keeping the item. Items that would be capable of reflecting asymmetry of performance on the two sides of the body were believed to be important items to retain because of their possible diagnostic value.

The resulting V.4 of the TIMP has 42 items—13 Observed Items and 29 Elicited Items—6 of which are tests of the same activity on either side of the body. After deleting the 17 selected items from the test, a new BIGSTEPS analysis was run using the 42 items of V.4 (before revision of rating scale descriptions of items E13, E14, and E29). The resulting summary statistics demonstrate that deletion of these items shortens the length of the test but does not significantly change the quality of the measuring device. The mean Rasch measure increased to 57.58 with a SD of 13.05 and mean SE of 2.34 with a SD of .64. Overall infit mean square for persons increased slightly to 1.01 and outfit mean square to 1.14. The person separation index improved to 4.85 with a reliability of .96.

Again, a wide range of item difficulty was demonstrated with an item separation index of 23.79 and reliability of 1.00. The mean item measure (as set for the analysis) was 50.00 with a SD of 13.49, model error of .47 with a SD of .25. Item difficulties ranged from 20 to 87. Point biserial correlations for items ranged from .01 to .81 with 28 of 42 (67%) greater than .50. Infit mean square for items was 1.02 with an outfit of 1.35.

Review of the infit mean square data for each item revealed that 5 items remained with infits greater than 1.2: O9 (1.35), O10 (1.39), E13 (1.63), E14 (1.57), and E29 (1.37), all items previously identified and purposely not removed from the test. As previously mentioned, E13, E14, and E29 have been revised with the expectation that they will perform better in data analyses with a new sample of infants. O9 and O10 will be retained for the reasons previously given and because further reanalysis removing these items did not produce substantial improvement in Rasch fit statistics for the test as a whole. Thirteen items remain with high outfit statistics. Although these could perhaps be deleted without affecting test reliability or precision, they are items which assess meaningful activities that therapists wish to have included for the purposes of obtaining a fuller understanding of infant development for use in treatment planning and anticipatory guidance of families regarding their infant's development.

Discussion

Rasch analysis was used to reduce the length of a new assessment of infant posture and movement. The 42-item TIMP has high precision, good fit to the Rasch psychometric model, and good fit to the ability levels of infants tested in the age range for which the test is intended, especially those with low to moderately high levels of ability (measure ability range from 10-85). The lack of a floor effect indicates sufficient room for assessment of low functioning infants, a major purpose of the examination. Harder items would need to be added if precision was desired for 4-month-old infants of the highest ability. This is not the test developers' intent, however, because other tests are available for assessing infants from 3-4 months of age onward while no other quantitative assessment is currently available with precision for assessing infants' functional motor performance under the age of 3 months.

With the elimination of 15 items from the set of Observed Items, many of which involved arm and hand functions, it becomes more obvious that the TIMP is an assessment primarily of gross (large muscle), rather than fine, motor function. It is likely that the misfit of these items reflected the fact that they are part of a different construct than that of posture and movement needed for early functional activities or that it is too early in life to reliably assess hand and arm functions. Unfortunately, removal of the large number of items that were Observed Items did little to reduce the physical demands of the TIMP nor the time required for testing because Observed Items do not involve handling. Reducing the number of Observed Items from 28 to 13, however, does reduce the attentional demand of the tester because these items must be observed for throughout the examination. In addition, one position change was deleted and removal of two Elicited Items contributes slightly to reducing the time required for testing. Future research will assess the feasibility of using a smaller set of the best items as a screening test and for assessment of the most fragile infants.

Because we are aware of no other test for young infants that was developed using Rasch analytic methods, we have no other comparable data to evaluate relative to our results. In general, however, other tests for newborns and premature infants have a small range of age, are not interval-scaled, and have not been studied with longitudinal assessment of infants to document a linear relationship between age or ability and test scores. Neither do they provide the capability to scale individual items to identify difficulty level of steps involved in learning various tasks and how they should appear in the sequence of early development of prematurely born infants.

Scales developed with Rasch methods are available for older children with disabilities (Coster, Ludlow, and Mancini, 1999), including the Gross Motor Function Measure (GMFM). A recent report of the use of Rasch analysis to reduce the number of items in the GMFM showed that a 66-item test was as sensitive to change with age in children with cerebral palsy as the earlier 88-item test, but the criteria used to delete items were not reported (Russell, Avery, Rosenbaum, Raina, Walter, and Palisano, 2000).

Along with evidence that the TIMP discriminates among infants based on early medical complications (Campbell and Hedeker, 2001), predicts development at 12 months from TIMP assessment at 3 months (Campbell, Kolobe, Wright, and Linacre, in press), and has ecologic validity for assessing activities that are important in daily life (Campbell and Murney, 1998), the evidence provided by Rasch analysis provides support for the validity of the TIMP for use in clinical practice and research to measure the development of motor skills in early infancy. The next stage in development of the TIMP will use V.4 for a re-assessment of the test's scaling properties in a population-based sample of 1200 infants selected to match the race/ethnicity distribution of the population of low-birth-weight infants in the U.S. Use of data from this cross-sectional sample in a new Rasch analysis will alleviate the limitations of the current data obtained from repeated assessment of about two-thirds of the infants. Data from the national sample will also be used to establish age standards for performance of infants in 14 different age groups in support of its use as a diagnostic measure of infant motor development.

Acknowledgments

Research was conducted at the University of Illinois at Chicago Medical Center, the University of Chicago Hospitals, and Lutheran General Hospital, Park Ridge, IL, with funding (R01 HD 32567) from the National Center for Medical Rehabilitation Research of the National Institute of Child Health and Human Development, USPHS. We are grateful to the physicians at each institution—Nagamani Beligere, Jaideep Singh, and David Sheftel—for providing access to their patients. The authors would like to thank Dolores Schorr, Pat Byrne-Bowens, Dawn Kuerschner, Carrie Ryan, and Kathy Tolzien for assistance in recruiting subjects; Beth Osten, Kathy Kamm, Pai-jun Mao Liao, Vinod Subramonian, and Sriram Balachandran for assistance with data management; and Betty Brannen, Mary Carter, Judy Flegel, LouAnn Gouker, Thubi Kolobe, Maureen Lenke, Pai-jun Mao Liao, Gail Liberg, Mary Murney, Beth Osten, Celina Wise Martin, and Laura Zawacki for testing of infants. Without the participation of infants and their parents who believed that our research would help other families, this work would not have been possible.

References

Campbell, S. K., and Hedeker, D. (2001) Validity of the Test of Infant Motor Performance for discriminating among infants with varying risk for poor motor outcome. *Journal of Pediatrics, 139,* 546-551.

Campbell, S. K., Kolobe, T. H. A., Osten, E.T., Lenke, M., and Girolami, G.L. (1995). Construct validity of the Test of Infant Motor Performance. *Physical Therapy, 75,* 585-596.

Campbell, S. K., Kolobe, T. H. A., Wright, B. D., and Linacre, J. M. (In press). Predictive validity of the Test of Infant Motor Performance with the Alberta Infant Motor Scale. *Developmental Medicine and Child Neurology.*

Campbell, S. K., Osten, E. T., Kolobe, T. H. A., and Fisher, A. G. (1993). Development of the Test of Infant Motor Performance. *Physical Medicine and Rehabilitation Clinics of North America, 4(3),* 541-550.

Coster, W., Ludlow L., and Mancini, M. (1999). Using IRT variable maps to enrich understanding of rehabilitation data. *Journal of Outcome Measurement, 3,* 123-133.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38 (9, Suppl. II),* 60-65.

Fanaroff, A. A., Wright, L. L., Stevenson, D. K., Shankaran, S., Donovan, E. F., Ehrenkrans, R. A., Younes, N., Korones, S. B., Stoll, B. J., Tyson, J. E., Bauer, C. R., Oh, W., Lemons, J. A., Papile, L. A., and Verter, J. (1995). Very-low-birth-weight outcomes of the National Institute of Child Health and Human Development Neonatal Research Network, May 1991 through December 1992. *American Journal of Obstetrics and Gynecology, 173,* 1423-1431.

Girolami, G., and Campbell, S. K. (1994). Efficacy of a Neuro-Developmental Treatment program to improve motor control of preterm infants. *Pediatric Physical Therapy, 6,* 175-184.

Linacre, J. M. (1988). *FACETS. Computer Program for Many-faceted Rasch Measurement.* Chicago: MESA Press.

Majnemer, A., Riley, P., Shevell, M., Birnbaum, R., Greenstone, H., and Coates, A. L. (2000). Severe bronchopulmonary dysplasia increases risk for later neurological and motor sequelae in preterm survivors. *Developmental Medicine and Child Neurology, 42,* 53-60.

McCormick, M. C., McCarton, C., Tonascia, J., and Brooks-Gunn, J. (1993). Early educational intervention for very low birth weight infants: Results from the Infant Health and Development Program. *Journal of Pediatrics, 123,* 527-533.

Murney, M. E., and Campbell, S. K. (1998). The ecological relevance of the Test of Infant Motor Performance Elicited Scale items. *Physical Therapy, 78,* 479-489.

Pinto-Martin, J. A., Whitaker, A. H., Feldman, J. F., Van Rossem, R., and Paneth, N. (2000). Relation of cranial ultrasound abnormalities in low-birthweight infants to motor or cognitive performance at ages 2, 6, and 9 years. *Developmental Medicine and Child Neurology, 41,* 826-833.

Russell, D. J., Avery, L. M., Rosenbaum, P. L., Parminder, S. R., Walter, S. D., and Palisano, R. J. (2000). Improved scaling of the Gross Motor Function Measures for children with cerebral palsy: evidence of reliability and validity. *Physical Therapy, 80,* 873-885.

Wildin, S. R., Smith, K. E., Anderson, A. E., Swank, P. R., Denson, S. E., and Landry, S. H. (1997). Prediction of developmental patterns through 40 months from 6- and 12-month neurologic examinations in very low birth weight infants. *Developmental and Behavioral Pediatrics, 18,* 215-221.

Windmeijer, F. A. G. (1990). The asymptotic distribution of the sum of weighted squared residuals in binary choice models. *Statistica Neerlandica, 44(2),* 69-78.

Wright, B. D., and Linacre, J. M. (1996). *A User's Guide to BIGSTEPS.* Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

JOURNAL OF APPLIED MEASUREMENT, *3*(2), 205–231

Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals

Everett V. Smith, Jr.
The University of Illinois at Chicago

The purpose of this research is twofold. First is to extend the work of Smith (1992, 1996) and Smith and Miao (1991, 1994) in comparing item fit statistics and principal component analysis as tools for assessing the unidimensionality requirement of Rasch models. Second is to demonstrate methods to explore how violations of the unidimensionality requirement influence person measurement. For the first study, rating scale data were simulated to represent varying degrees of multidimensionality and the proportion of items contributing to each component. The second study used responses to a 24 item Attention Deficit Hyperactivity Disorder scale obtained from 317 college undergraduates. The simulation study reveals both an iterative item fit approach and principal component analysis of standardized residuals are effective in detecting items simulated to contribute to multidimensionality. The methods presented in Study 2 demonstrate the potential impact of multidimensionality on norm and criterion-reference person measure interpretations. The results provide researchers with quantitative information to help assist with the qualitative judgment as to whether the impact of multidimensionality is severe enough to warrant removing items from the analysis.

Requests for reprints should be sent to Everett Smith, University of Illinois at Chicago, 1040 West Harrison Street, M/C 147, Chicago, IL, 60607, e-mail: evsmith@uic.edu.