The WeeFIM Instrument: Its Utility in Detecting Change in Children With Developmental Disabilities

Kenneth J. Ottenbacher, PhD, Michael E. Msall, MD, Nancy Lyon, RN, PNP, Linda C. Duffy, PhD, Jenny Ziviani, PhD, Carl V. Granger, MD, Susan Braun, MLS, OTR, Roger C. Feidler, PhD

ABSTRACT. Ottenbacher KJ, Msall ME, Lyon N, Duffy LC, Ziviani J, Granger CV, Braun S, Feidler RC. The WeeFIM instrument: its utility in detecting change in children with developmental disabilities. Arch Phys Med Rehabil 2000;81: 1317-26.

Objective: To examine the utility of the WeeFIM® instrument (''WeeFIM®'') in detecting changes in the functional status of children with disability.

Design: Prospective longitudinal design with correlation and responsiveness analysis.

Setting: Three facilities providing services to children with developmental disabilities in western New York State.

Participants: Two hundred five children (72 girls, 133 boys) with identified medical disabilities receiving special services were administered the WeeFIM. Subjects ranged in age from 11 to 87 months and came from diverse socioeconomic and ethnic backgrounds. Scores for 174 children were available for 3 administrations performed over a 1-year period.

Main Outcome Measures: The responsiveness of the WeeFIM instrument was examined using 5 statistical procedures: Reliability Change Index, Proportional Change Index, effect size, standardized response means, and paired *t* tests.

Results: All 5 indexes of responsiveness indicated statistically significant ($p < .05$) or reliable changes over time. The transfer subscale of the WeeFIM showed a skewed distribution that affected the results for some responsiveness indexes. The advantage, limitations, and assumptions of the responsiveness indexes are described and graphic examples of change over time are presented to validate the responsiveness of the WeeFIM instrument.

Conclusion: The WeeFIM instrument showed the ability to document change in functional abilities over a 1-year period in children with chronic disabilities.

Key Words: Disability evaluation; Disabled children; Rehabilitation.

IN DISABILITIES among children, Wegner et al[1] stated that an estimated 4 million children and adolescents, or 6.1% of the US population aged younger than 18 years, have disabilities. *Disability* is defined broadly to include any limitation in activity caused by a chronic health condition or impairment. Ing and Tewey[2] estimated that the percentage of children with disability ranges from 2.6% in children 0 to 3 years of age up to 12.4% in children 15 to 17 years of age. The most common chronic conditions resulting in disability in children include learning disabilities, mental retardation, and orthopedic impairments.[3] The evaluation and assessment of children with disability is an important and growing area of pediatric practice. A recent supplement to *Pediatrics* reviewed the use of health status, functional outcome, and quality-of-life instruments in evaluating children with disabilities.[4] In the supplement, Hack[4] states that there is a need to reexamine the use of growth and neurodevelopmental status instruments and to supplement these with measures of health status, functional outcomes, and quality of life for pediatric patients and their families. As the scope of pediatric assessment instruments expands to include functional outcomes and health-related quality of life, clinicians need information on the reliability, validity, and responsiveness of new evaluation and screening tools. Aylward[5] described the process of selecting assessment and evaluation instruments for use in clinical and educational settings. He proposed an evaluation matrix that includes the results of the assessment, the child's environment, the area of function assessed, age of the child, and medical history. Aylward[5] emphasized the need to establish the sensitivity and responsiveness of assessment instruments in addition to the traditional measurement attributes of reliability and validity.

Responsiveness is defined as the ability of an instrument to detect clinically important differences over time.[6] Guyatt et al[6] noted that the usefulness of a test to measure change over time is dependent not only on reliability and validity, but also on the measurement property of responsiveness. They stated that ''it is possible for instruments to be reliable, but unresponsive to change.''[6]

The purpose of this investigation is to examine systematically the responsiveness of the WeeFIM® instrument (''WeeFIM''®).[7] The WeeFIM instrument provides an indication of functional outcomes in children and is modeled on the FIM™ instrument (''FIM''). The FIM instrument is widely used in adult rehabilitation settings.[8] The WeeFIM instrument includes 18 measurement items and can be administered in 20 minutes or less. The goal of the WeeFIM instrument is to ''measure changes in function over time to weigh the burden of care in terms of physical, technologic, and financial resources.''[9] Recent reports on the reliability and validity of the WeeFIM instrument indicate that the assessment has excellent consistency across raters and provides scores that are stable.[10-14] Good equivalence reliability has also been shown between WeeFIM ratings obtained from direct observation and from reports by parents and teachers.[12,13] The responsiveness of the WeeFIM instrument has not been previously examined.

From the University of Texas Medical Branch, Galveston, TX (Ottenbacher); Child Development Center, Providence, RI (Msall); Robert Warner Rehabilitation Center (Lyon); Children's Hospital of Buffalo (Duffy); State University of New York (Granger, Braun, Feidler); Buffalo, NY; and University of Queensland, Brisbane, Australia (Ziviani).

Accepted in revised form February 8, 2000.

Supported by the Department of Health and Human Services, Health Resources and Services Administration, Maternal and Child Health Bureau (grant no. MCJ-360646).

The authors have chosen not to select a disclosure statement.

Reprint requests to Kenneth J. Ottenbacher, PhD, University of Texas Medical Branch, SAHS 4.202, 301 University Blvd, Galveston, TX 77555-1028, e-mail: *kottenba@utmb.edu*.

0003-9993/00/8110-5915$3.00/0

doi:10.1053/apmr.2000.9387

METHODS

Participants

Two hundred five children with developmental disabilities, ranging in age from 11 to 87 months, participated in the investigation. All children had a confirmed medical diagnosis and were receiving treatment and/or developmental child support services in early intervention or school-based programs. The sample was recruited from 3 early childhood education programs and developmental disabilities/rehabilitation facilities in western New York State. All 3 facilities included educational day programs designed specifically for children with disabilities. Many of the children were in day programs that included children without disabilities. All children included in the sample received occupational, speech, or physical therapy as part of their individualized educational programs or individualized family service plans. A proportional sampling plan was used to ensure that children were evenly distributed based on severity, type of disability, and age. The proportional sampling plan involved creating cells based on gender, age, medical diagnosis, and severity of disability and then selecting the predetermined number of children to fill each cell. This approach helped ensure that such potential moderator variables as age, gender, and severity of disability were evenly distributed across the 205 children included in the initial sample. Severity of disability was estimated based on scores from standardized developmental assessments, including the Bayley Scales of Infant Development,[15] Clinical Adaptive Test/Clinical Linguistic Auditory Milestone Scale (CAT/CLAMS),[16,17] and McCarthy Scales.[18] The children were tested by licensed professionals with the appropriate instrument based on medical condition at their initial entry into the health care system. Scores on these assessments were available in the children's health care and/or educational records. Information on socioeconomic status (SES) was collected using demographic factors, including family income, educational level and occupation of parents, availability of transportation, use of paid outside help in the home, and presence of a telephone. The most common medical impairments were cerebral palsy, prematurity, Down syndrome, spina bifida, epilepsy, and genetic disorders.

The study protocol was reviewed and approved by the appropriate institutional review boards. All parents and teachers participating in the investigation granted written informed consent before providing information about the children in their care.

Assessment Instrument

The children participating in the investigation were evaluated using the WeeFIM.[7] It is a pediatric functional assessment developed by health and rehabilitation professionals.[7,14,19,20] The instrument is derived from the items of the FIM, which was originally designed to assess severity of disability in adults.[8] Key characteristics of the WeeFIM instrument are its use of a minimal data set, emphasis on consistent actual performance, and the ability to be used by multiple disciplines.[21]

The WeeFIM instrument (version 4.0)[7] contains 18 measurement items that are divided into 6 areas: self-care (6 items), sphincter control (2 items), transfers (3 items), locomotion (2 items), communication (2 items), and social cognition (3 items). The motor subscale includes the areas of self-care, sphincter control, transfer, and locomotion; it contains 13 items. The remaining 2 areas (communication, social cognition) comprise the cognitive subscale. A 7-level ordinal rating system ranging from 7 (complete independence) to 1 (total assistance) is used to rate performance. A rating from 1 to 4 indicates that the child requires some level of assistance from another person to complete the activity. A rating of 5 means the child requires supervision or adult cues to set up the task. A rating of 6 means that the child can complete the activity independently but may require an assistive device, more than a reasonable amount of time, or safety is a concern. Table 1 lists the WeeFIM items and rating protocol. The polar graph in figure 1 shows how the WeeFIM instrument demonstrates changes in performance for each of the individual items.

The WeeFIM instrument can be administered through direct observation, interview, or a combination of observation and interview.[7] Each item must be rated. No zeros or nonapplicable ratings can be given. The minimum possible total rating is 18 (total dependence in all skills); the maximum possible rating is 126 (complete independence in all skills). The WeeFIM instrument is designed for use by a variety of professionals. Training is recommended to ensure appropriate administration and rating (discussed later). Validity and interrater reliability have been examined in various studies and found to be excellent (interclass correlation coefficients $> .95$).[11-14]

Procedure

Parents were initially interviewed by a trained rater in the facility in which the child received intervention or follow-along services. The purpose of the study was explained, and each

Table 1: Sample WeeFIM Instrument Rating Form*

Table Unavailable Online.
Please See Print Journal.

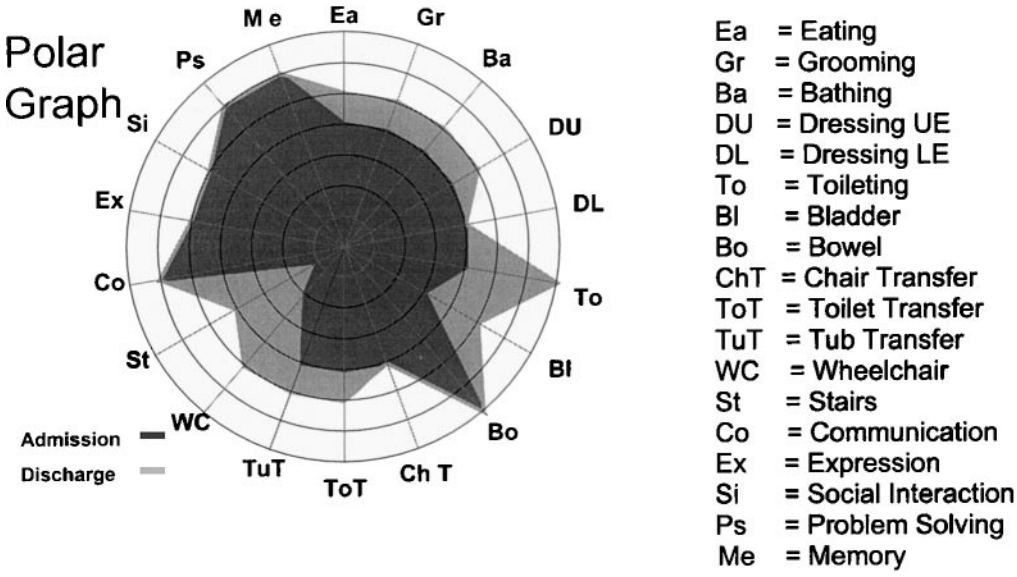

Fig 1. Sample polar graph showing changes in WeeFIM ratings over time for each individual item. All items scored on a 1- to 7-point scale. Inner ring indicates score of 1; outer ring indicates score of 7. The expanding rings indicate ratings of the WeeFIM. The penultimate outer ring represents a score of 6. UE, upper extremity; LE, lower extremity.

parent received an information sheet. In cases in which neither parent was available, the assessment was administered to a caregiver designated by the parent as familiar with the child's functional abilities. In all cases in which the parent was not available, the interview was completed with the child's teacher ($n = 87$).

Interviewers were blind to the child's health status and previous scores on developmental tests before the initial assessment. The date and time of all assessments were recorded, and every effort was made to schedule the second assessment on the same day of the week and at approximately the same time as the initial assessment. The same person (parent or teacher) interviewed initially was also interviewed during the follow-up assessments. The second assessment was administered to all 205 caregivers of the children within a period of 5 to 20 days after the first testing. These 2 administrations were part of a larger investigation of the validity and reliability of the WeeFIM. The results of the reliability and validity analysis are available in other sources.[10-12] The final (third) administration of the WeeFIM instrument took place 1 year after the initial administration. Of the original 205 children, 174 were available for all 3 assessments. One-year follow-up interviews were conducted with the same caregiver when possible. In all, 136 (78%) of the 1-year follow-up interviews were conducted with the original parent or teacher. The time of day of the original interview was recorded and duplicated if possible in the 1-year follow-up interview. There was no statistically significant difference in demographic characteristics (age, sex, SES, etc) between the original sample ($n = 205$) and the sample available at 1 year ($n = 174$). To maintain a homogenous sample, all analyses reported here were conducted using the 174 children with complete test data for all 3 administrations.

Interviewers. The primary interviewer, who collected 63% of the initial WeeFIM data and all the 1-year follow-up data, was a pediatric nurse practitioner with more than 20 years experience in developmental disabilities and rehabilitation. Other interviewers collecting WeeFIM information were health, developmental, or rehabilitation professionals with at least 3 years of experience working with children with disabilities and their families. Each interviewer completed training in the administration of the WeeFIM instrument, which included review of the administration and scoring protocol and viewing a 25-minute videotape.[22] Successful completion of the training required 90% agreement with case study material. If the 90% criterion on the first assessment was not achieved after training, the protocol was repeated. Each interviewer recorded a minimum of 2 pilot WeeFIM assessments to establish the interview format.

Statistical Analysis

Five measures of responsiveness were computed. The 5 measures were selected based on their previous use in the research literature and their relevance to measuring clinically important change in children with developmental disabilities. The measures are described next.

Reliability Change Index. The Reliability Change Index (RCI) was originally proposed by Jacobson et al[23] as a general purpose measure of clinical change. They argued that for a change in performance to be clinically significant, it must be statistically reliable (ie, there must be some way to determine that the change is not caused by chance variation or measurement error). To make this determination, Jacobson[23] proposed the use of the RCI. To compute the RCI, the following information is required: (1) the child's initial or preintervention score; (2) the follow-up or posttest score; and (3) the standard error of measurement (SEM) for the test. The SEM represents the spread or the distribution of repeated performances for a given individual. It is influenced by the reliability of the test and may be computed from the following formula:

$$\text{SEM} = S_E \sqrt{1 - r}$$

in which S_E equals the standard deviation for the test and r is the reliability coefficient (test-retest) for the instrument. The RCI is computed as follows:

$$\text{RCI} = (X_2 - X_1)/\text{SEM}$$

in which X_2 is the follow-up or posttest score, X_1 is the initial or pretest score, and SEM is the standard error of measurement, as previously defined. Jacobson and Truax[24] modified the RCI denominator to include the standard error of the difference (SE_{Diff}) for the SEM. Jacobson et al[23] argued that an RCI more than ± 1.96 would be unlikely to occur ($p < .05$) without actual change.

Proportional Change Index. The Proportional Change Index (PCI) was introduced by Wolery[25] as an alternative to previously proposed efficacy indexes.[26] The PCI is designed to

measure developmental improvement in any domain (motor, cognitive, language, etc). To compute the PCI, the following information is required: (1) the child's chronologic age; (2) the time between first and second measurement; (3) the child's initial (preintervention) score; and (4) the child's follow-up (postintervention) score. The PCI is computed using the following formula:

PCI = [(developmental gain/time between assessments) ÷ (initial developmental age/follow-up chronologic age)].

Wolery argued that the PCI compares the child's rate of development before initial assessment to his/her rate of development during the period between first and second measurement.[21] The amount of developmental gain (follow-up assessment minus developmental age at initial testing) is divided by the duration between assessments. The rate of development before initial testing is computed by dividing the child's pretest developmental age by his/her pretest chronologic age. For example, assume a developmental assessment such as the Bayley Scales of Infant Development is administered to a 12-month-old infant and the infant received a score that translated into a developmental age of 8 months. After the evaluation, the child begins a program of intervention that lasts 10 months. At the end of this period, the child is retested and receives a score on the Bayley Scales that translates to a developmental age of 22 months. The developmental gain over the 10-month period is: 22 months (follow-up assessment) − 8 months (initial test) = 14 months. The PCI score would be: PCI = [(14mo/10mo) ÷ (8mo/12mo)] = 2.1.

Children who continue to develop during the period between testing at the same rate as they did before initial testing will receive a PCI score of 1. Children whose rate of development is greater during the period between first and second testing than their (estimated) rate of development before testing will receive a PCI score greater than 1. Children whose rate of development during the period between initial and posttesting is less than their (estimated) rate of development before initial testing will receive a PCI score less than 1.

The PCI is not just a gain score. Two children who show the same number of months of improvement between first (pre) and second (post) testing may have different PCI scores because of different rates of estimated development before initial testing. Consider the following 2 PCI scores:

Child 1: PCI = [(14mo/10mo) ÷ (12mo/12mo)] = 1.4.

Child 2: PCI = [(14mo/10mo) ÷ (8mo/12mo)] = 2.1.

The time between initial assessment and follow-up testing was the same for both children (ie, 10mo), and both made similar developmental gains (ie, 14mo). The PCI score for child 2 is larger because her rate of gain between the first and second assessment was greater relative to her estimated rate of development before initial testing.

Effect Size Index. The Effect Size (ES) relates the magnitude of the change score to the variability in scores.[27] In general, the ES is calculated by taking the mean change found in a particular variable and dividing it by the standard deviation of that variable:

$$ES = (X_2 - X_1)/SD$$

where X_2 is the follow-up (posttest score), X_1 is the initial (pretest) score, and *SD* is the average standard deviation. There is some controversy regarding which standard deviation to use.[28] The options for computing the standard effect size (d-index) for 2 group comparisons are using the standard deviation of the initial (pretest) score or using the pooled standard deviation for the initial and follow-up scores. We used the standard effect size d-index, computed by subtracting the initial score from the follow-up score and dividing by the pooled standard deviation following the procedure outlined by Cohen.[27] Laing et al[29] used a variation of the ES they called the standardized response mean (SRM). SRMs are calculated by dividing mean change scores for before and after comparisons by the standard deviation of the difference scores. Higher values indicate better responsiveness. We computed SRMs using the method described by Laing.[29]

p values. Guyatt et al[6] recommended the *p* value as an index of responsiveness. They propose using *p* values generated by paired *t* tests. We used the paired *t* test to compare initial and follow-up WeeFIM ratings and reported the exact *p* value.[6]

RESULTS

Demographic Characteristics

Seventy percent of the children in the sample were white, with 21% black, 6% Hispanic, and 3% other. One hundred eleven were boys and 63 were girls. As noted, severity of disability was determined based on scores obtained by the children on standardized instruments when they initially entered the health care delivery system. Fifty children (29%) had original standardized scores on the Bayley Scales, CAT/CLAMS, or McCarthy Scales between −1 and −2 standard deviations less than the mean (mild disability); 94 children (54%) had standardized developmental scores between −2.1 and −3 standard deviations (moderate disability); and 30 children (17%) had standardized scores greater than −3 standard deviations less than the mean (severe disability). Additional demographic information for the sample is listed in table 2.

Pearson's product moment correlations were computed between age and WeeFIM ratings. The correlation between age (in months) and total WeeFIM rating was statistically significant ($r = .67$, $p < .01$), with younger children showing lower scores than older children. The correlation between SES and initial total WeeFIM rating was not statistically significant ($r = -.01$, p = NS). As expected, there was a statistically significant difference ($p < .01$) in initial WeeFIM instrument ratings across the 3 levels of severity of disability. The mean initial total WeeFIM rating for children categorized as having mild disability was 30.54 ± 14.95; those with moderate disability had a mean total WeeFIM rating of 64.32 ± 21.55; and those children identified as severely disabled had a mean total WeeFIM rating of 73.03 ± 23.25. The relationship of severity of disability and total WeeFIM rating across the 3 test administrations is shown in figure 2.

A chi-square analysis was conducted to examine the relation between severity and gender. As noted, 3 levels of severity were defined (mild, moderate, severe) based on the child's initial assessment at entrance into the health care system. The chi-square analysis of 3.91 ($df = 2$, $p = .186$) was not statistically significant. We also examined if there were differences in age across the 3 levels of severity using a 1-way analysis of variance (ANOVA). The ANOVA showed F = 0.342 ($df = 2$, $p = .711$), suggesting no statistically significant differences in age across levels of severity.

Several of the responsiveness indices described in the previous section might be influenced by distributions that are not normal. For example, the standard deviation used in computing effect size measures, SRMs, and paired *t* tests is influenced by departures from normality. Skewness and kurtosis coefficients were computed for the 6 WeeFIM areas (self-care, sphincter control, transfers, locomotion, communication, social cognition), subscales (motor, cognitive), and total

Table 2: Demographic Information for Sample of Children With Developmental Disabilities (n = 174)

Gender	
Boys	111 (64%)
Girls	63 (36%)
Age	
Mean	59.98 ± 20.37mo
Median	56.00mo (range, 23.0–108.0)
SES	
Mean	13.19 ± 4.82
Median	14.0 (range, 1–20)
Ethnic background	
White	128 (74%)
Black	31 (18%)
Hispanic	11 (6%)
Other	4 (2%)
Level of severity*	
Between −1 and −2 sd below mean	50 (29%)
Between −2 and −3 sd below mean	94 (54%)
Greater −3 sd below mean	30 (17%)
Medical condition**	
Congenital impairment	n = 9
Down syndrome	n = 12
Developmental disorder	n = 32
Mental retardation	n = 54
Motor control	n = 36
Cerebral palsy	n = 41
Communication disorder	n = 72

* Level of severity determined by standardized scores on initial assessment into health care system.
** Medical conditions total >174 because some children had more than 1 condition.

ratings. The skewness values ranged from .59 to 1.16. Skewness coefficients greater than 1 were found for 2 WeeFIM areas (transfers, locomotion). Normality plots were computed for motor and cognitive subscales and for total WeeFIM instrument ratings. The Kolmogorov-Smirnov statistic with Lilliefors significance level for testing normality[30] was computed and indicated normal distributions for the 2 combined subscales (motor, cognitive) and for total WeeFIM instrument ratings.

Figure 3 presents the mean total WeeFIM ratings for first, second, and follow-up assessments for the children. A statistically significant difference in initial total WeeFIM ratings was found for boys and girls (t = 2.49, p = .14, df = 172). This finding contradicts previous research that has generally not reported statistically significant differences in WeeFIM ratings based on gender.[21] Further analysis of this finding showed a difference in age for the 174 boys and girls in this subsample of the original 205 children. The average age for boys was 42.03 ± 19.78 months; the average age for girls was 46.70 ± 20.15 months. This difference was statistically significant (t = 2.49, $p < .05$, df = 172). When age was controlled as a covariate, there was no statistically significant difference between boys and girls in initial total WeeFIM instrument ratings.

Table 3 includes the means and standard deviations for the WeeFIM instrument area ratings (self-care, sphincter control, transfers, locomotion, communication, social cognition), the WeeFIM subscale ratings (motor, cognitive), and the total WeeFIM ratings for all 3 test administrations. The relation between the ratings over time is graphically presented in figures 4 and 5. Inspection of figure 4 shows little difference between the first 2 ratings (obtained within a period of 5–20 days). In contrast, the third rating taken at 1-year follow-up shows a substantial increase across all WeeFIM areas assessed. The responsiveness of this change as measured by the RCI, PCI, effect sizes, SRMs, and p values is listed in table 4.

Reliability Change Index. The RCI was computed using the formula described previously (RCI = $(X_2 - X_1)/SE_{Diff}$). An RCI was computed for each child for the 6 WeeFIM areas, the 2 subscales, and the total rating. In computing the RCI, the average of the first 2 ratings was used for the initial score (X_1). The test-retest reliability for the WeeFIM instrument subscales and total ratings reported by Ottenbacher et al[10] were used in computing the SEM used in the RCI computations. Descriptive statistics for the RCI for all WeeFIM instrument areas, subscales, and total ratings are included in table 4. All the RCI values except the transfers subscale are greater than 2, suggesting that the changes in WeeFIM instrument ratings from initial testing to follow-up assessment were larger than what could be expected because of measurement error or chance variation. The results imply the WeeFIM instrument was responsive to changes in functional outcomes caused by nonchance factors, ie, maturation, intervention programs, or some interaction between maturation and intervention.

Proportional Change Index. To compute the PCI, the WeeFIM instrument ratings were converted to developmental ages using age norm tables provided in the WeeFIM instrument manual (version 4.0).[7] Using developmental ages, the PCI was computed for each child. The initial WeeFIM instrument rating, initial developmental age, follow-up WeeFIM rating, and follow-up developmental age were used to compute the PCI values. All the mean PCI scores are greater than 1 (see table 4), indicating that during the period between initial assessment and follow-up (1yr), the children in this sample progressed at a rate equal to or greater than that experienced before the initial assessment.

Effect size and SRMs. Table 4 also includes the effect size (d-index) and SRM values for the WeeFIM areas, subscales, and total ratings. The d-index values ranged from .31 to .81 for WeeFIM instrument ratings. Cohen[27] has proposed a classification system for the d-index in which a value of .20 to .50 is considered a small effect size, .51 to .80 is considered a medium effect size, and a d-index greater than .80 is labeled as a large effect size.

p values. The p values generated by the paired t test analysis are also included in table 4. Exact p values are reported following the recommendation of Gyuatt et al.[6] All the paired t test comparisons were made using the initial WeeFIM instrument ratings and the follow-up ratings obtained at 1 year. All tests were statistically significant at the p less than .001 alpha level.

Miscellaneous analyses. We examined floor and ceiling effects and the impact of measurement error by calculating correlation coefficients (ρ) between the individual difference scores (follow-up rating minus initial rating) and the initial WeeFIM instrument score. This analysis was designed to determine if the difference ratings were dependent on the initial WeeFIM instrument score. The correlation for total WeeFIM rating was −.24, the correlation for the motor subscale was −.29, and the correlation for the cognitive subscale was −.17. Difference correlations were also computed for the 6 WeeFIM instrument areas and ranged from −.12 (self-care) to −.51 (locomotion). The correlation of −.51 is statistically significant and suggests that those children with initial low ratings in locomotion items showed the largest difference scores at 1-year follow-up.

Floor and ceiling effects were further examined by plotting initial ratings against WeeFIM instrument ratings at 1-year follow-

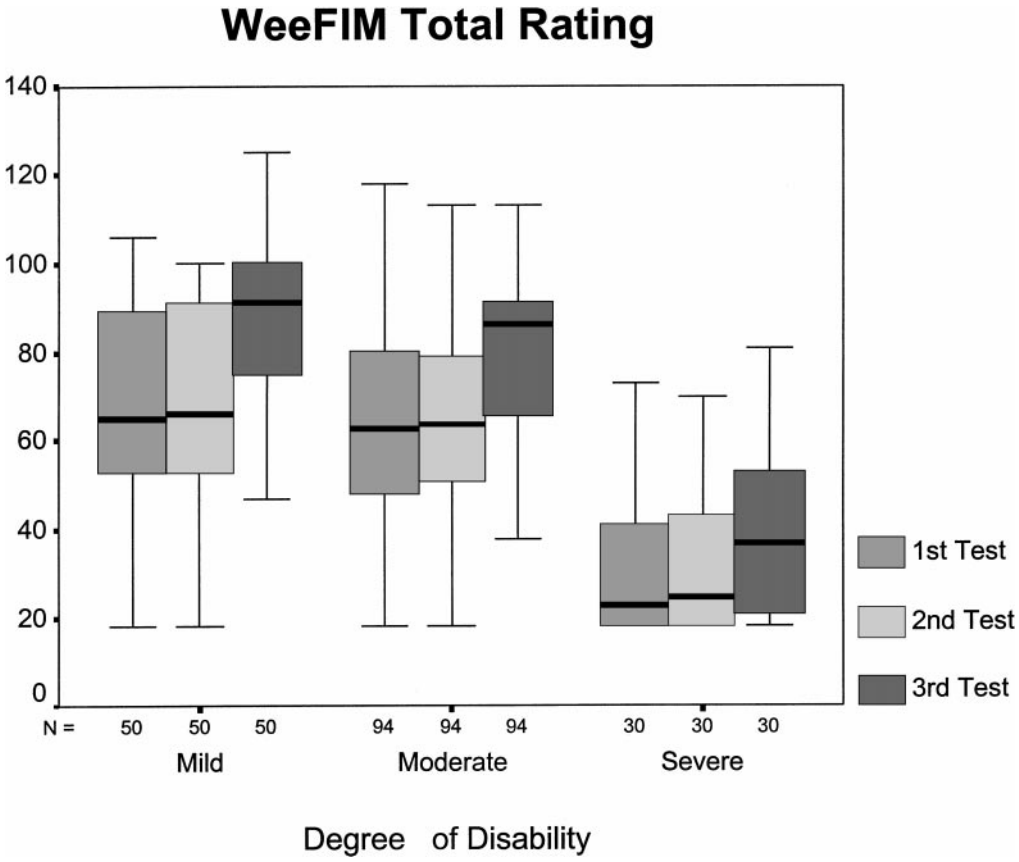

Fig 2. Comparison of total WeeFIM rating for 3 administrations across the 3 levels of disability: mild, moderate, and severe.

up. These plots for total WeeFIM ratings and the motor and cognitive subscales appear in figure 6. Points falling above the diagonal represent improvement, points falling directly on the diagonal represent no change, and points falling below the diagonal indicate deterioration from initial rating to follow-up at 1 year.

Table 4 includes discrete values reflecting responsiveness of the WeeFIM instrument over time. Depending on the reliability of the instrument, all follow-up scores will be somewhat imprecise. The RCI was used to derive confidence intervals that define the range in which an individual score is likely to fluctuate because of measurement error. Figure 6 illustrates the use of confidence intervals as a band (dashed lines) around the diagonal line. Points falling outside the band around the diagonal represent changes that are ''statistically'' reliable based on the confidence interval of the RCI. Those children represented by points falling within the band showed changes that were not ''statistically'' reliable as determined by the RCI.

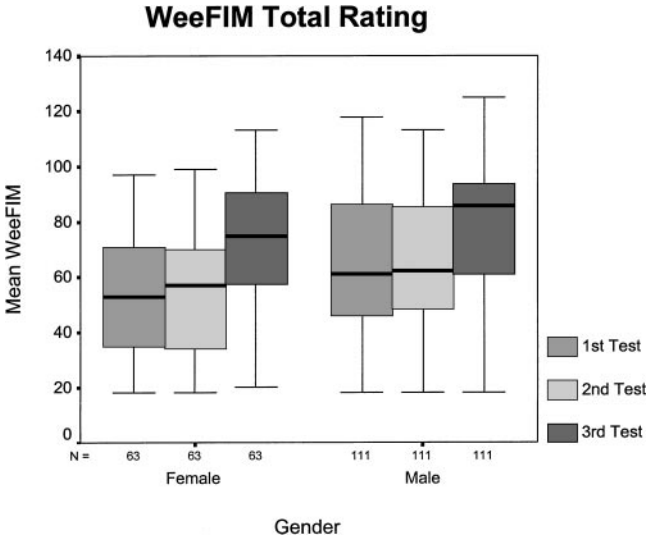

Fig 3. Comparison of total WeeFIM rating for boys and girls across the 3 test administrations.

DISCUSSION

The purpose of this investigation was to examine the responsiveness of the WeeFIM instrument—its utility in detecting clinically important change. Several indices of responsiveness were examined. Different measures were used because there is no empirical consensus regarding the single best procedure for determining responsiveness using clinical rating scales.[30] The results suggest that the WeeFIM instrument was able to detect change over time in the 174 children with disabilities in this sample. Indications of change were obtained for all 5 responsiveness measures included in the analyses. The results for any single responsiveness index are constrained by method and statistical limitations. For example, the PCI requires that performance be measured or transformed into a developmental age equivalent, and it assumes that development progress is linear.

Four of the 5 responsiveness indices used relied on such parametric statistics as means and standard deviations that assume normally distributed scores. For at least 2 of the WeeFIM areas (transfers, locomotion), the distributions for the initial ratings were positively skewed, and the responsiveness

Table 3: Means and Standard Deviations for WeeFIM Instrument Across 3 Assessments

WeeFIM Variables	1st Assessment	2nd Assessment	1-Year Follow-Up
Area			
Self-Care	14.0 Median	16.0 Median	21.0 Median
Potential range (1-42)	16.34 Mean	16.90 Mean	20.98 Mean
	(7.97 sd)	(7.61 sd)	(8.16 sd)
Sphincter Control	4.0	4.0	9.5
Potential range (1-14)	6.02	6.27	8.63
	(4.71)	(4.65)	(4.85)
Transfers	13.0	13.0	18.0
Potential range (1-21)	12.37	12.50	14.94
	(6.40)	(6.32)	(5.89)
Locomotion	12.5	13.0	13.0
Potential range (1-14)	10.18	10.23	11.40
	(4.12)	(3.95)	(3.33)
Communication	6.0	6.0	8.0
Potential range (1-14)	6.24	6.03	7.95
	(2.79)	(2.60)	(3.03)
Social Cognition	8.0	8.0	11.0
Potential range (1-21)	8.25	8.26	11.11
	(3.58)	(3.36)	(3.92)
Subscale			
Motor	43.5	45.0	59.0
Potential range (1-91)	44.89	45.91	55.97
	(20.65)	(20.25)	(19.35)
Cognitive	14.0	14.0	19.0
Potential range (1-35)	14.49	14.28	19.06
	(6.01)	(5.75)	(6.07)
Total	59.0	59.0	80.0
Potential range (1-126)	59.38	60.21	74.82
	(24.89)	(24.56)	(24.35)

index for these 2 WeeFIM areas must be interpreted cautiously. The only WeeFIM area that did not produce a statistically reliable change using the RCI was transfers; it also had the smallest effect size and SRM. Transfers was 1 of the WeeFIM areas with a skewed distribution.

Floor effects may have also influenced the ability of the responsiveness indexes to detect change. Figure 6 suggests that a number of children showed floor effects. There were also significant correlations between initial rating and difference scores for 2 of the WeeFIM areas (transfers, locomotion). These correlations suggest that children with the lowest initial scores showed the largest difference scores at 1-year follow-up. This raises the possibility of regression toward the mean, where extreme scores are more likely to move toward the average on follow-up or retest. Some measurement error may have been introduced because at 1-year follow-up, not all information was collected from the original caregivers. In approximately 12% of the cases, the person who provided information on the child at 1-year follow-up was not the same person interviewed at the initial assessment. However, the WeeFIM instrument has shown good interrater reliability in past studies.[10-12]

One way to deal with measurement error and such related problems as nonnormal distributions would be to use a proportional difference as the criterion for change, or to conduct a logarithmic data transformation. Hebert et al[31] argued that this type of transformation has not been entirely successful when used with adult disability rating scales. They found that a square root transformation addressed some of the problems associated with skewed distributions and heterogeneity of variance in

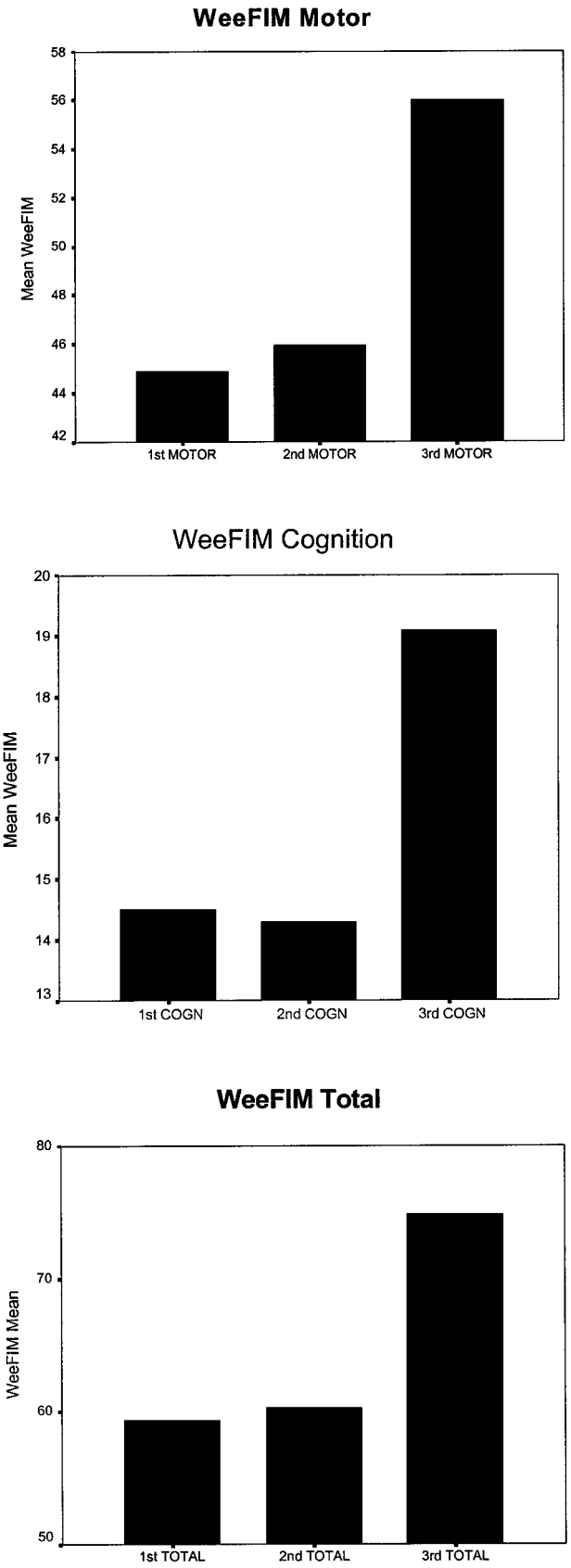

Fig 4. Comparison of total WeeFIM ratings across 3 test administrations for motor, cognitive, and total WeeFIM ratings.

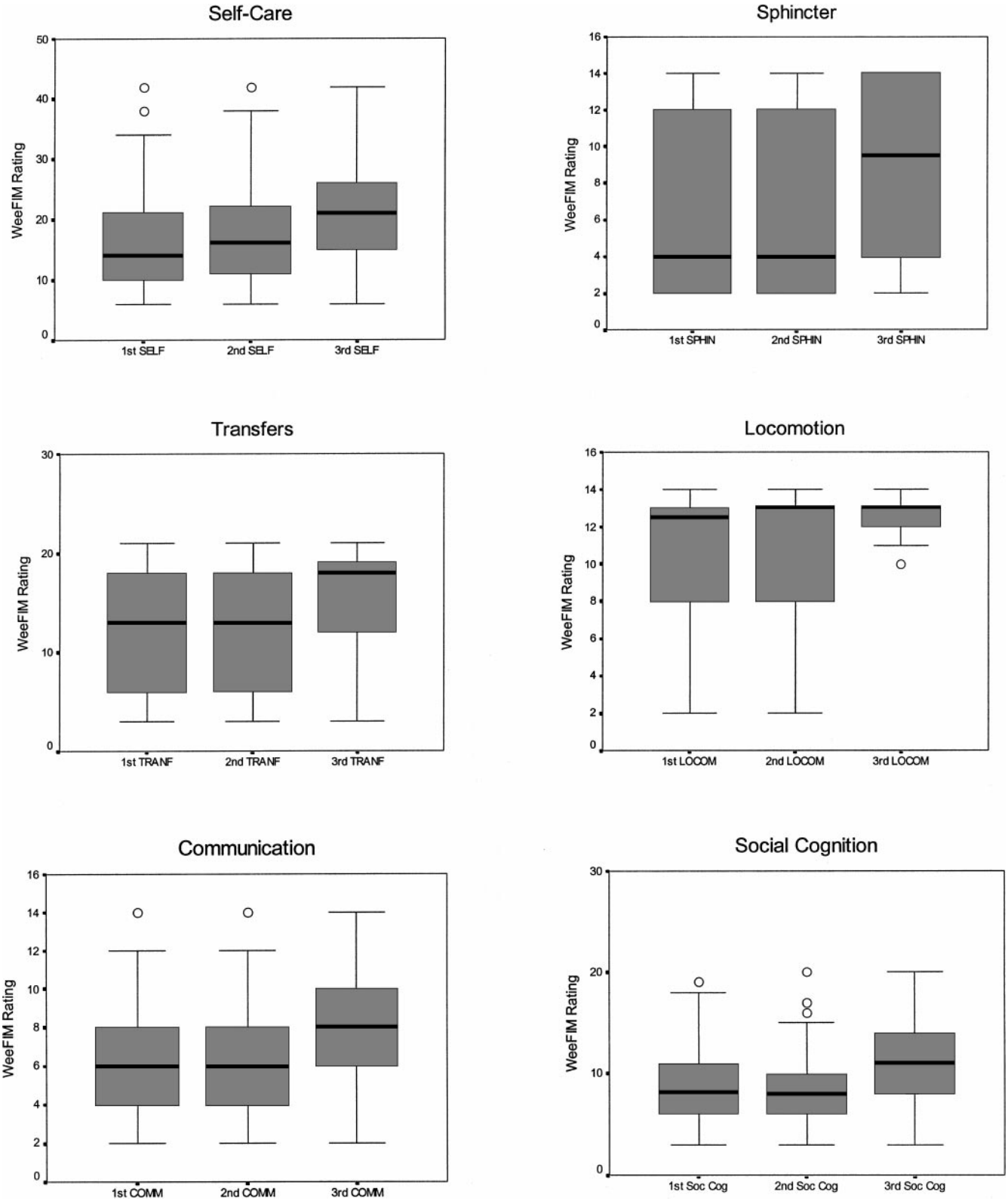

Fig 5. Comparison of WeeFIM ratings across 6 WeeFIM areas over a 1-year period.

Table 4: Mean Responsiveness Measures for WeeFIM Instrument Ratings

	Responsiveness Index*				
WeeFIM Variables	RCI	PCI	ES	SRM	*p* Value
Area					
Self-Care	2.64	1.81	.58	1.02	<.00
Sphincter control	2.13	1.21	.55	.70	<.00
Transfers	1.83	1.02	.40	.75	<.00
Locomotion	2.09	1.11	.31	.52	<.00
Communication	2.29	1.71	.65	.80	<.00
Social cognition	2.84	1.68	.81	1.04	<.00
Subscale					
Motor	2.46	1.51	.54	1.21	<.00
Cognitive	2.96	1.31	.76	1.12	<.00
Total	2.82	1.41	.62	1.31	<.00

* RCI, Reliability Change Index (values >1.96 statistically significant $p < .05$); PCI, Proportional Change Index; ES, Effect Size (d-index); SRM, Standardized Response Means; *p* value, exact *p* value for paired *t* test.

difference scores, but noted ''such a transformation is barely understandable and quite artificial.''[31]

Another approach to dealing with floor effects and heterogeneity of difference scores is to weight items differently according to their relative importance. Some items or areas in the WeeFIM instrument are more important and imply greater consequences in terms of disability and resource costs than others. In the WeeFIM instrument, as in other disability rating scales, all individual items have the same weight in the calculation of the total scores. The argument could be made that there is a hierarchy of functional outcomes with some basic activities of daily living (ADL) being more ''important'' than other instrumental ADL.[32] For example, the bowel and bladder items in the area of sphincter control may affect a number of other daily living skills and have greater impact on total disability than other ADL items, such as dressing upper body or climbing stairs. Weighting items is a complex process that must be done at multiple levels; theoretical, clinical, and empirical. Empirical weighting requires comparison with a criterion common to all abilities. This criterion could be amount of nursing care, time to complete specific ADL tasks, or costs for aides and assistive devices. This type of data has been collected in the study of functional outcomes in adults with disabilities.[33-36] Similar research is needed as functional status instruments are developed and introduced into pediatric practice and research.

A related issue relevant to the use of the WeeFIM is the distinction between changes that are clinically important versus statistically significant. The use of traditional statistical methods, such as the paired *t* test used in this study, is limited in 2 important respects. First, the test provides no information on the nature of responses from an individual. The statistical results are derived from average (pooled) information and provide no indication of the importance or degree of change recorded for an individual patient. Second, whether change exists in the statistical sense has little to do with the clinical or practical importance of the effect. It is widely known that statistical significance can overestimate or underestimate clinical importance based on sample size and other factors.[30,31] Questions regarding efficacy and responsiveness refer to benefits derived from changes in ability and the impact those changes have on the person's daily life. Conventional statistical comparisons tell us little about the clinical and personal value of changes in functional status. In contrast to statistical significance, judg-

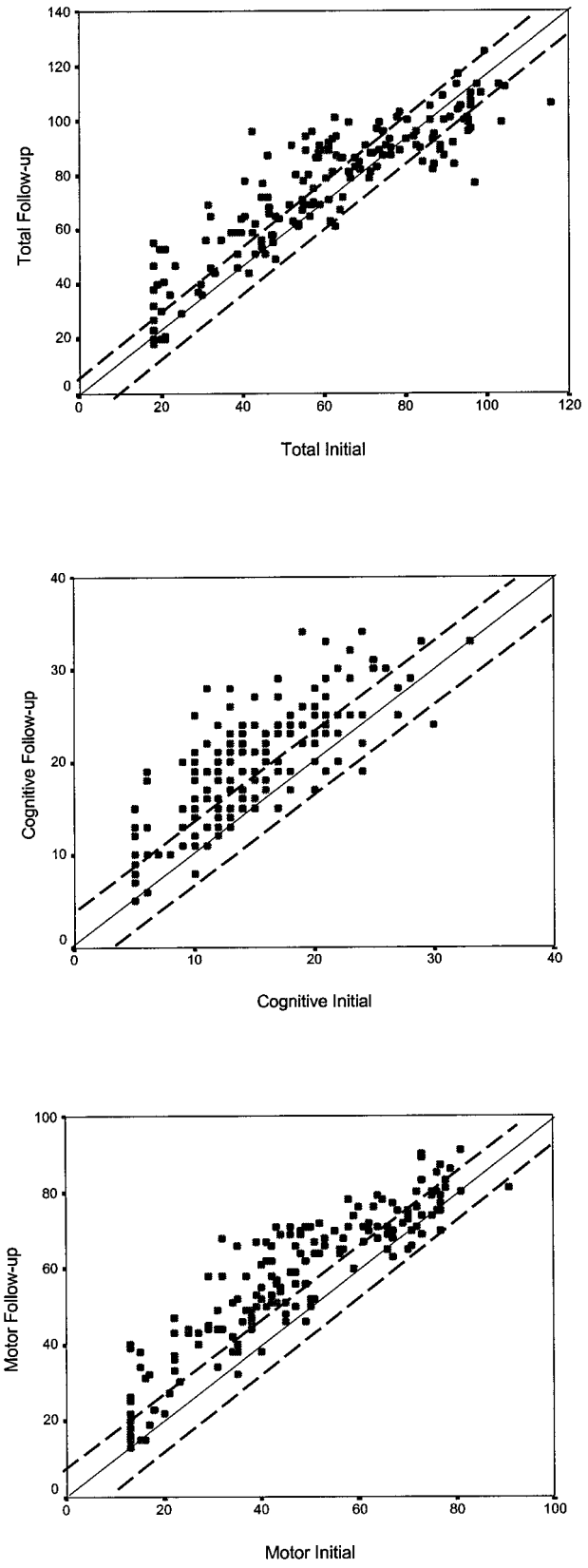

Fig 6. Plot of WeeFIM initial ratings against follow-up ratings.

ments regarding clinical importance are based on external standards provided by interested parties, including the patient, families, teachers, and other consumers. Researchers are aware of the need to integrate clinical and statistical significance, and procedures have been proposed to encourage this process.[27,28]

CONCLUSION

The information provided here is valuable because it provides clinicians with basic information for making distinctions between clinical and statistical significance in using the WeeFIM. Some of the responsiveness measures directly reflect statistical significance, whereas others, such as the RCI and PCI, are related to detecting change that is not caused by measurement error or is proportional to assumed previous development. This information should be useful in providing individualized interpretations of change over time in children with developmental delays and disabilities. The results also provide evidence that the WeeFIM instrument is sensitive enough to document changes in basic ADL functions over time in children with disabilities. Future research on the responsiveness of the WeeFIM instrument and other pediatric disability rating scales must continue to focus on functional outcomes that are not only statistically significant, but also important in the daily lives of children with disabilities and their families.

References

1. Wegner BL, Kaye S, LaPlante MP. Disabilities among children. Disability Statistics Abstracts, No. 15. Washington (DC): US Department of Education, National Institute on Disability and Rehabilitation Research; March 1996.
2. Ing CD, Tewey BP. Summary of data on children and youth with disabilities. Washington (DC): US Department of Education, National Institute on Disability and Rehabilitation Research; 1994. Contract No. HN93027011.
3. Newacheck PW, Taylor WR. Childhood chronic illness: prevalence, severity, and impact. Am J Public Health 1992;82:364-70.
4. Hack M. Consideration of the use of health status, functional outcomes, and quality-of-life to monitor neonatal intensive care practice. Pediatrics 1999;103(Suppl E):319-28.
5. Aylward GP. Conceptual issues in developmental screening and assessment. J Dev Behav Pediatr 1997;18:340-9.
6. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.
7. Guide for the Functional Independence Measure for Children (WeeFIM) of the Uniform Data System for Medical Rehabilitation, Version 4.0—Community/Outpatient. Buffalo (NY): State University of New York at Buffalo; 1993.
8. Guide for the Uniform Data Set for Medical Rehabilitation (Adult FIM), Version 4.0. Buffalo (NY): State University of New York at Buffalo; 1993.
9. Braun S, Granger CV. A practical approach to functional assessment in pediatrics. Occup Ther Pract 1991;2:46-51.
10. Ottenbacher KJ, Msall ME, Lyon NR, Duffy LC, Granger CV, Braun S. Interrater agreement and stability of the Functional Independence Measure for Children (WeeFIM): use in children with developmental disabilities. Arch Phys Med Rehabil 1997;78:1309-25.
11. Ottenbacher KJ, Taylor ET, Msall ME, Braun S, Lane S, Granger CV, et al. The stability and equivalence reliability of the Functional Independence Measure for Children (WeeFIM). Dev Med Child Neurol 1996;38:907-16.
12. Sperle PA, Ottenbacher KJ, Braun S, Lane SJ, Nochajski S. Equivalency reliability of the Functional Independence Measure for Children (WeeFIM) administration methods. Am J Occup Ther 1997;51:35-41.
13. DiScala C, Grant CC, Brooke MM, Gans B. Functional outcome in children with traumatic brain injury: agreement between clinical judgement and the functional independence measure. Am J Phys Med Rehabil 1992;71:145-8.
14. Msall ME. Functional assessment in neurodevelopmental disabilities. In: Capute AJ, Accardo PJ, editors. Developmental disabilities in infancy and children. Vol 2, 2nd ed. Baltimore (MD): Paul Brookes; 1996. p. 311-8.
15. Bayley N. Bayley Scales of Infant Development. 2nd ed. San Antonio (TX): Psychological Corp; 1994.
16. Hoon AH, Pulsifier MB, Goplan R, Palmer FB, Capute AJ. Clinical Adaptive Test/Clinical Linguistic Auditory Milestone Scales in early cognitive assessment. J Pediatr 1993;113:S1-8.
17. Rossman MJ, Hyman SL, Rorabaugh ML, Berlin LE, Allen MC, Modlin JF. The CAT/CLAMS assessment for early intervention services. Clin Pediatr 1994;33:404-9.
18. McCarthy DA. Manual for the McCarthy Scales of children abilities. New York: Psychological Corp; 1972.
19. Msall ME, DiGaudio K, Duffy LC, LaForest S, Braun S, Granger CV. WeeFIM—normative sample of an instrument for tracking functional independence in children. Clin Pediatr 1994;65:431-8.
20. Msall ME, DiGaudio K, Rogers BT, LaForest S, Catanzaro NL, Campbell J, et al. The Functional Independence Measure for Children (WeeFIM): conceptual basis and pilot use in children with developmental disabilities. Clin Pediatr 1994;33;421-30.
21. Msall ME, DiGaudio KM, Duffy LC. Use of functional assessment in children with developmental disabilities. Phys Med Rehabil Clin North Am 1993;4:517-27.
22. Msall ME, Braun S, Granger CV. Use of the Functional Independence Measure for Children (WeeFIM): an interdisciplinary training tape [abstract]. Dev Med Child Neurol 1990;32(Suppl):90.
23. Jacobson N, Follette W, Revenstorf D. Psychotherapy research: methods for reporting variability and evaluating clinical significance. Behav Ther 1984;15:336-52.
24. Jacobson N, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy. J Consult Clin Psychol 1991;59:12-9.
25. Wolery M. Proportional change index: an alternative for comparing child change data. Except Child 1983;50:167-71.
26. Simeonsson RJ, Wiegernik R. Accountability: a dilemma in infant intervention. Except Child 1975;45:474-81.
27. Cohen J. Applied statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Earlbaum; 1988.
28. Stucki G, Daltroy JN, Katx N, Johannesson M, Laing MH. Interpretation of change scores in ordinal scales and health status measures. Why the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711-7.
29. Laing MH, Larson MG, Cullen KE, Schwartx JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.
30. Hays WL. Statistics for the social sciences. 2nd ed. New York: Holt, Rinehart & Winston; 1973.
31. Hebert R, Spiegelhalter DJ, Brayne C. Setting the minimally detectable change on disability rating scales. Arch Phys Med Rehabil 1997;78:1305-8.
32. Coster WJ, Haley SM. Conceptualization and measurement of disablement in infants and young children. Infant Young Child 1992;4:11-22.
33. Spector WD, Katz S, Murphy JB, Fulton JP. The hierarchical relationship between activities of daily living and instrumental activities of daily living. J Chronic Dis 1987;40:481-9.
34. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the Functional Independence Measure. Am J Phys Med Rehabil 1993;72:84-9.
35. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: a study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990;71:870-5.
36. Granger CV, Cotter AC, Hamilton BB, Fiedler RC. Functional assessment scales: a study of persons after stroke. Arch Phys Med Rehabil 1993;74:133-8.