The Reliability of the Motor Development Checklist

Marcia Gevelinger,
Kenneth J. Ottenbacher,
Trina Tiffany

Key Words: developmental disabilities • motor skills evaluation • test reliability • tests, by title, Motor Development Checklist

A study was conducted to determine the interrater and test-retest reliability of the Motor Development Checklist (Doudlah, 1976). Thirty-six subjects with diagnosed multiple disabilities and severe and profound mental retardation were evaluated by two experienced occupational therapists. Reliability was established by means of the intraclass correlation approach, and results indicated that coefficients ranged from .93 to .99. The study results suggest that the Motor Development Checklist, when administered by experienced raters, can provide consistent information regarding the spontaneous motor performance of persons with multiple disabilities. Implications for practice and future research are briefly discussed.

Persons with severe and profound mental retardation often have a combination of physical, cognitive, and sensory disabilities (Batshaw & Perret, 1981). Because of the interactive effects of the varied impairments, it is difficult to accurately isolate and assess motor development levels in persons with multiple disabilities. Many standardized motor assessments require the use of visual or auditory cues or the application of a specific physical stimulus to elicit a response in a testing situation. For example, the Motor subscale of the Bayley Scales of Infant Development contains items that require visual ability for the performance of the motor tasks (Bayley, 1969).

The Motor Development Checklist (MDC) developed at the Central Wisconsin Center for the Developmentally Disabled in Madison, Wisconsin, was designed specifically to record spontaneous motor behaviors in persons with severe developmental disability (Doudlah, 1976). The approach employed in the MDC reflects the belief that the spontaneous action patterns displayed by individuals are the best descriptors of their developmental status (Slaton, 1981). All items reflect only motor performance and do not require the use of visual or auditory cues. This feature of the MDC makes it an appropriate assessment device for use with the severely and profoundly retarded population.

The MDC also permits the longitudinal collection of data. The format of the instrument allows for a 2-year period of monthly recording covering the sequence of motor development from birth to ambulation. Thus, the rate of development of motor milestones can be monitored.

In developing the MDC, Doudlah (1981) addressed the concern that the criteria for the normal motor development sequence used in most assessments were not based on the observation of spontaneous movement. Doudlah constructed the MDC using visual records of the spontaneous motor behavior of 20 normal children who were filmed monthly in familiar surroundings over an 18-month period. Only the spontaneous behavior of the children was filmed. Children were never prompted or made to perform activities they were not capable of doing independently. Hence, the criteria for success on the MDC are based solely on spontaneously observed motor performance. No physical prompts or visual and/or auditory cues are necessary. (See Figure 1 for a copy of the MDC.)

The MDC has several strengths that make it an ideal tool for motor assessment in persons with severe and profound mental retardation. These strengths are as follows: (a) the criteria for success on the MDC are based solely on motor performance, (b) the MDC permits sequential measurements that can be used to establish the rate of motor development, and (c) the
assessments are derived from a normal developmental sequence.

To establish the clinical usefulness of an instrument that is based on observation, such as the MDC, it must be demonstrated that the assessment tool can provide consistent information across different occasions as well as different observers. In other words, the reliability of the MDC must be established. Deter-
mining reliability is a logical first step in establishing the psychometric adequacy of any assessment that is based on observation. It is widely understood that the validity of an instrument cannot exceed the square root of its reliability (Neufeld, 1977). Thus, instruments used in the clinic must produce stable results if they are to provide therapeutically meaningful and valid information.

The purpose of the present study was to determine the interrater and test-retest reliability for the MDC with a sample of severely and profoundly retarded, multiply disabled persons. A natural setting correlational research design was used.

Method

Subjects

Thirty-six subjects who were residents of the Central Wisconsin Center (CWC) for the Developmentally Disabled were observed on their living unit. This particular living unit at CWC was chosen because of the variety of motoric levels demonstrated by the residents living there and the potential to incorporate completed assessments into routine therapy evaluations. Within this unit, all available residents were observed. All residents were functioning at 0.8-month to 20-month levels of development, as determined on the Wisconsin Behavior Rating Scale (WBRS) (Song et al., 1984). The chronological age range of subjects was from 9 to 26 years, with a mean of 18.25 years. (See Table 1 for diagnoses of the subjects’ disabilities and related descriptive information.)

Measurements

The MDC was used to collect data on the observed level of motor behavior. MDC scores were obtained by two raters who were occupational therapists working at the CWC; each rater had at least 3 years’ experience in the area of developmental disabilities. The two raters were asked to score the MDC using the following rating scale as described by Doudlah (1976): 0—does not perform task; 1—beginning to attempt task; 2—performs task occasionally; and 3—performs task skillfully. The following definitions were developed specifically for this study to help standardize scoring between raters:

**Beginning to attempt task:** The subject is observed completing some components of the task, but a completed performance is not observed. Performance is never of the quality observed in children developing normally.

**Performs task occasionally:** Performance at this level may be seen only with facilitation. Tasks in this category are often performed at a level observed in nondisabled infants and children. However, the task is not performed consistently.

**Performs task skillfully:** The task is effectively and consistently performed in appropriate situations or as necessary in interacting with the environment. Quality is equal to that observed in nondisabled infants and children.

Training Program

Each rater was trained in the use of the MDC. An 18-minute videotape produced by Doudlah, “Motor Development—Birth to Walking,” which was available at the CWC, was shown to the therapists to familiarize them with the terminology and developmental sequences described in the MDC. Both examiners rated the items on the MDC while viewing the training videotape and other tapes of developmentally disabled individuals; they reached an 80% agreement rate in scoring selected videotapes before the actual study was begun. When the raters disagreed, they discussed the item until they reached a consensus. This procedure was used to ensure that the raters had a similar understanding of rating procedures prior to beginning the study. The entire training program was completed in 6 hours.

Design of Study

Two therapists simultaneously observed each of the 36 subjects individually for a 10-minute period and independently scored each subject’s motor behavior on the MDC, using the rating system previously described. A stopwatch was used to determine the beginning and the end of the observation periods. Only spontaneous motor behaviors exhibited during that time period were recorded. During the observations,

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description of Subjects</strong></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td><strong>Race</strong></td>
</tr>
<tr>
<td>Caucasian</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Oriental</td>
</tr>
<tr>
<td>Indian</td>
</tr>
<tr>
<td><strong>Diagnosis</strong></td>
</tr>
<tr>
<td>Post trauma/illness</td>
</tr>
<tr>
<td>Cerebral palsy</td>
</tr>
<tr>
<td>Genetic disorder</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td><strong>Age (in months)</strong></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Standard deviation</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td><strong>WBRS scores (in months)</strong></td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Standard deviation</td>
</tr>
<tr>
<td>Range</td>
</tr>
</tbody>
</table>

Note: WBRS = Wisconsin Behavior Rating Scale (Song et al., 1984).
no direct interaction occurred between the raters and the subjects. However, the observers were visible to the subjects; hence the possibility exists that the subjects were aware of being observed.

No modifications of the environment were made. If subjects were routinely placed in positioning devices to facilitate motor development or minimize orthopedic deformity, observations were made with the subject in these positions. A record of the subject’s position was maintained to ensure consistency in reevaluation. Subjects were allowed to interact with staff, peers, and objects in their living units as if no observations were occurring.

After the observations were completed, the completed checklists from each rater were placed in separate envelopes. To reduce any bias which could have affected the retest values, scores were not reviewed or totalled until the end of the study.

One therapist from the original pair of raters observed each subject a second time, 2 to 13 days after the first observation. The variation in the retest period occurred as a result of the need to coordinate with school schedules, home visits, and periods of illness. Environmental conditions as similar as possible to those of the initial observation were maintained on reevaluation. A list of any environmental conditions not consistent with the usual unit routine was compiled to be used by the therapist rating the subjects during second observations. If a significant difference in the environment was perceived at the time of the attempted reevaluation, the observation was rescheduled. Subjects who were positioned in external support devices for initial evaluations were observed a second time only when positioned in the same way.

Data Analysis

We computed an intraclass correlation coefficient using a method based on generalizability theory to determine the consistency of ratings across two examiners and over time. According to Berk (1979), only the intraclass correlation generalizability theory approach offers the comprehensiveness, precision, and flexibility required to deal with reliability. Berk (1979) noted that although most observational studies assess reliability in terms of observer agreement or interclass correlation, these measures are inflexible and lacking in precision. Traditional bivariate correlations cannot account for sources of variance that could confound the interpretation of reliability (cf. Cronbach, Ikeda, & Auner, 1964). For example, a high bivariate correlation coefficient, such as a Pearson product moment correlation, may merely reflect parallelism between scores. Unless the means and standard deviations for the two tests or raters are similar, a high correlation indicates only that the scores covary in a similar way. One can resolve or reduce many problems associated with more traditional methods of assessing reliability by using the intraclass correlation generalizability theory approach. The intraclass correlation is generally more informative because it takes into account mean differences among scores and parallelism between test results (Jones, 1977). The intraclass correlation expresses the classical theory of measurement error relationship between true and observed variance. It is computed from analysis of variance components. Reliability corresponds to the proportion of the total variance in test scores that is due to true differences among individuals in the quality evaluated by the test (cf. Berk, 1979).

Results

The initial analysis of the data included computations of the means and standard deviations of the sets of scores recorded on the MDC. The mean of the scores for Rater 1 was 27.78 (SD = 31.34), and the mean for Rater 2 was 31.28 (SD = 35.05). The mean for retest by Rater 1 was 25.69 (SD = 29.19). Table 2 summarizes descriptive information for the scores obtained by each rater.

To provide additional information on the variability of the scores, we determined quartile values. The middle quartile is the median value of the scores, and the interquartile range (Q3–Q1) is the middle 50% of a set of scores. Table 3, which contains this information for ratings of the MDC, indicates that the scores obtained on the MDC ranged from 0 to 102.
with subjects generally scoring at the lower end of the scale. The maximum obtainable score on the MDC is 105.

A one-way analysis of variance (ANOVA) was conducted to determine if the total MDC scores varied significantly across the four diagnostic categories or by sex. The ANOVA computed with sex as a factor resulted in an $F$ of 0.07 ($p = .79$, $df = 1,35$). The ANOVA for diagnostic categories produced an $F$ of 1.53 ($p = .23$, $df = 3,35$). The results indicate that no statistically significant differences in performance existed across these two factors. All further analyses were conducted with the total data set collapsed across sex and diagnostic categories.

To graphically illustrate the nature of the relationship between the ratings on the MDC, we developed scatter plots for the scores obtained for Rater 1 and Rater 2, and for the initial test and the retest administrations (see Figures 2 and 3). In both figures, the data points approximate a straight line and represent a positive linear relationship.

The intraclass correlation coefficient (ICC) measuring the covariation of scores of Evaluator 1 and Evaluator 2 was .98. The ICC for association between initial and reevaluation was .98.

The MDC is composed of five sections. The five sections include three separate phases of items and two transitions between these phases (see Figure 1). In addition to determining the relationship of total scores, intraclass correlation coefficients were computed for scores obtained in each of these phases and transitions. ICCs obtained ranged from .93 to .99. These intraclass correlation coefficients are reported in Table 4.

Finally, we computed Pearson product moment correlation coefficients to determine the existence of any significant relationship between age, time between test and retest, and the total scores obtained by each evaluator. The coefficients ranged from $-0.06$ to $0.10$ and revealed no significant relationship among any of the variables.

Discussion and Conclusions

The MDC meets several criteria that are important in evaluating the motor development of persons with mental retardation and severe physical disabilities. Items are based on a normal developmental sequence, criteria for success require only motor performance, and the checklist can easily be used for se-

![Figure 2](image_url)

**Figure 2**
Scatter Diagram of Scores Recorded by Rater 1 and Rater 2 (Contains Four Hidden Observations)

![Figure 3](image_url)

**Figure 3**
Scatter Diagram of Scores Recorded on Initial Evaluation and on Retest (Contains Seven Hidden Observations)

<table>
<thead>
<tr>
<th>Phase</th>
<th>Between Raters</th>
<th>Test-Retest</th>
<th>Number of Scale Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>.94</td>
<td>.93</td>
<td>8</td>
</tr>
<tr>
<td>Transition 1</td>
<td>.96</td>
<td>.97</td>
<td>4</td>
</tr>
<tr>
<td>Phase 2</td>
<td>.97</td>
<td>.98</td>
<td>7</td>
</tr>
<tr>
<td>Transition 2</td>
<td>.97</td>
<td>.97</td>
<td>8</td>
</tr>
<tr>
<td>Phase 3</td>
<td>.97</td>
<td>.99</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>.98</td>
<td>.98</td>
<td>35</td>
</tr>
</tbody>
</table>
quential recording to document longitudinal changes.

The results of this study indicate that the interrater and test–retest reliability coefficients for the MDC are very high when the test is administered by trained, experienced therapists. In interpreting the reliability measures, delimiting factors of the study must be considered. All subjects in this investigation were residents of a specific living unit at Central Wisconsin Center. Although this unit offered diversity in motor development, it is possible that different measures would have been obtained for observations of patients in other institutional or community settings.

The two raters in this study were experienced in working with a developmentally disabled population and were trained in the use of the MDC. Because differences in experience level and training can influence the outcome of a study, it is possible that different reliability estimates would have been obtained with changes in any one of the above factors.

The raters involved in this study were aware that reliability was being checked—a factor that may have contributed to a higher level of agreement between raters. However, steps were taken to minimize the threat of this potential bias. Raters did not discuss the observed motor behaviors, and the completed evaluations were separated and scores not totalled until the end of the study.

Finally, test–retest reliability coefficients are typically influenced by the interval of time between evaluations, with shorter time intervals resulting in higher reliability coefficients (Wentling, 1980); however, scores on the reevaluations in this study were not significantly related to the number of days between tests.

Although observations of the described population were highly reliable, the degree of validity of the MDC is not known. Since it is possible for a test to measure performance consistently, yet not accurately measure what it was designed to measure, validity must also be assessed. Validity is an essential property of any clinic instrument (Mitchell, 1979). The clinical usefulness of the MDC cannot be confirmed until additional studies investigating the validity of the assessment device have been conducted and reported.

The determination of interrater and test–retest reliability for other raters is another important area for future investigation. Scores on the MDC may vary as a function of the experience or backgrounds of the raters. Future research should address the impact of the training period on the reliability of the MDC. Can therapists who do not receive any instruction in the administration of the MDC use it to reliably record the spontaneous motor behavior of patients with severe and profound disabilities? What is the reliability of the MDC when used with a nondisabled sample of infants and children? Answers to these and other questions will provide the information necessary to establish the clinical usefulness of the MDC.

Acknowledgments

We thank the staff and the residents of the Central Wisconsin Center for their cooperation in this study.

The first author completed her contribution to this study in partial fulfillment of the requirements for the Master of Science in Therapeutic Science degree at the University of Wisconsin–Madison. The advice of the thesis committee members Rita Hohlsstein, Alice Punwar, and G. Alan Stull is gratefully acknowledged.

Copies of the Motor Development Checklist can be obtained by writing to the Library and Information Center, Central Wisconsin Center for the Developmentally Disabled, 317 Krauson Drive, Madison, Wisconsin 53704.

References

Children with handicaps: A Medical primer. 
Baltimore, MD: Paul H. Brooks.

New York: Psychological Corporation.

American Journal of Mental Deficiency, 83, 460–472.

Psychological Reports, 15, 727–736.

Madison, WI: Central Wisconsin Center.


Mitchell, S. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. 

New York: Grane & Stratton.


American Journal of Mental Deficiency, 88, 401–410.