The present study attempted to construct a standardized Development Battery for Indian children aged 0 to 7 years and 11 months in five key areas of development. The Development Battery utilizes input from parents/caregivers, observation, and direct testing of the child to provide standard scores in five areas of development, namely motor, adaptive, social-emotional, cognition, and communication, and a total composite standard score.

**General Description and Scoring of the Development Battery**

The Development Battery is a criterion referenced, individually administered, standardized assessment battery to measure the development skills of Indian children aged birth through 7 years and 11 months in five domains of development. This is the first multi-domain development battery for use with Indian children with a wide age range and one which provides standard scores at the domain and at the test composite score level. In addition, the battery provides percentile ranks for each of the standard scores across domain and total score level, and errors of measurement.

Many researchers have argued that in India the lack of availability of indigenously developed intelligence and developmental tests for very young children have constrained researchers to use instruments developed in the Western countries (Anadalakshmy, 1982; Bevli et al., 1989; Malhi & Singhi, 1999; Murlidharan, 1992). The dangers of importing Western tests for use in other socio-cultural settings have been well documented and include the use of culturally inappropriate and irrelevant test items, inappropriate testing methods, and use of tests not standardized on the target population (Bevli, 1990; Bhave et al., 2009; Gupta & Patel, 1991; World Health Organization, 1992). Several Indian authors have emphasized the need to ensure that developmental assessment tests should be culturally sensitive, developed and standardized on population which is representative of the population to be tested (Chaudhari, 1996; Kumar et al., 1995; Malhi & Singhi, 1999; Singhania & Sonksen, 2004; Singhi, 1992). Hence, the currently developed test provides child health professionals a culturally sensitive development battery which can assess development across all domains.
The Development Battery has five domains and each of these domains further has two sub-domains each, with items ranging from 21 to 73 items, designed to assess the development and functioning of children from birth to eight years. For example, the Motor Domain has two sub-domains; the gross motor and the fine motor with 45 and 40 items, respectively. The Adaptive domain has two sub-domains; the self-care and personal-responsibility with 38 and 28 items, respectively. The Social-Emotional domain also has two sub-domains, namely social role and interpersonal interaction with 21 and 27 items, respectively. The Cognitive domain has two sub-domains, namely the attention and perception and concepts and academics with 73 and 67 items, respectively. The Communication domain has two sub-domains; receptive and expressive with 35 and 38 items, respectively. The entire battery takes about 60 to 75 minutes to complete and this is similar to the time taken to administer other comprehensive development batteries. For example, both the Vineland Adaptive Behaviour Scales (VABS, Sparrow et al., 1984) and the Battelle Developmental Inventory II (BDI II, Newborg, 2005) take about an hour or more to complete.

It is noteworthy that none of the developmental assessment instruments constructed in India provide standard scores at the domain or at the total composite level for children across the entire development range. For example, the recently developed Lucknow Development Screen for Indian Children (Bhave et al., 2009) is a 27 item screen which assesses children aged 6 months to 2 years of age and assigns only a pass/fail score. The only Indian development assessment tool which provides standard scores is the Developmental Assessment Scales for Indian Infants (DASII; Misra & Phatak, 1996), however, it assesses only two developmental domains, i.e., the motor and mental, for children aged up to two and half years only. The most popular test for assessment of children is the Indian adaptation of the Vineland Social Maturity Scale (Malin, 1971) and this assesses only the adaptive behaviour functioning of the child and provides a ratio quotient score. Hence, child health professionals working in India have to assess children's development by taking recourse to using developmental assessment batteries developed in the West such as the Developmental Profile II (Malhi et al., 2006, Malhi & Singh, 2003; Sidhu et al., 2010a) and the Battelle Developmental Inventory II (Kaur et al., 2009, 2010).

The scoring process of the Development Battery is somewhat complicated, particularly the process of establishing basal and ceiling levels and
hence requires careful administration and some amount of training. In this context, one may note that batteries which involve establishing basal and ceiling levels have been found to be prone to several scoring errors (Bailey, Burchinal, & McWilliam, 1993). For example, in scoring the Battelle Developmental Inventory II (BDI, Newborg, 2005) which involves a similar process of establishing a basal and ceiling level, Bailey et al. (1993) found that teachers took longer than 30 minutes to score and made errors like adding (45%), errors in establishing basal level (43%), and scoring items below the basal level (23%). Only 14.5% of the teachers were found to make no errors. Clearly then, the administrators of the present battery would require some training for being able to administer, score, and interpret it in the correct way.

The Development Battery is administered by first finding a basal age, which is the age level below which the child gets full credit for all items in a sub-domain. This has been placed as scoring 2 on four consecutive items. The ceiling age level is determined as an age beyond which a child would fail all items. This has been placed as scoring 0 on four consecutive items. A score of 2 indicates that the child manifests the skill in a consistent manner or the response is correct; 1 means that the skill is emerging and the child demonstrates the skill sometimes; and 0 is scored when the response is incorrect or the child does not demonstrate the skill. It is important to note that the basal age may not be established in the youngest age groups and the ceiling may not be established for the oldest age groups because of limited number of items which are available at the extreme age ranges. This scoring system is useful as it provides guidelines for the examiner to build on skills which are emerging while designing remediation programs. Similar scoring systems have been utilized by other comprehensive development batteries such as the Battelle Developmental Inventory II (BDI II, Newborg, 2005).

Standardization

The standardization of the Development Battery was a multi-phase exercise which extended over four years. In the initial phase new items and items from other development batteries were collected and field tested. Then a total of 412 items were included in a pilot phase to obtain a feedback about the items, administration difficulties, problems with understanding, and problems with scoring. Items that were determined to have problematic psychometric properties, difficult to administer, score
or understand by parents and/or child, or age inappropriate were either removed or modified.

The standardization sample consisted of 626 children, aged birth through 7 years and 11 months. A total of 18 age groups formed the sample; with more children being recruited at younger ages as the development is most rapid at the youngest ages. Data were collected over 24 months. Care was taken to enroll equal number of boys and girls, children from both rural and urban areas, and belonging to different socio-economic status in order to have a standardization sample which would most closely resemble school going children in North India. The size of the sample used for standardization was somewhat small as compared to other development batteries. In this context it may be noted that other assessment batteries have been standardized on larger samples. For example, the Developmental Profile 3 (DP 3, Alpern, 2007) was standardized on 2216 children, the Battelle Developmental Inventory II (BDI II, Newborg, 2005) on 2500 children, and the Vineland Adaptive Behaviour Scales (VABS, Sparrow et al., 1984) on 3000 children. These batteries were multi-centric studies and utilized a nationally representative sample, and multiple examiners. In contrast, the present battery was developed only in the Union Territory of Chandigarh and utilized a single examiner.

The present study utilized a classical method of test standardization which is particularly useful for small samples. This procedure uses the fitted progression of data across age levels to develop a normalized aggregate distribution of all available cases, with appropriate corrections for the moments of individual distributions. This aggregate distribution is used to represent the norms at every age level, after adjustment for differences in level and dispersion (Angoff & Robertson, 1987). However, one may note that there are several other newer and more efficient methods which can be used for calculating standard scores like Mantel-Haenszel statistics (Holland & Thayer, 1988), WINSTEPS (Linacre, 2003), and Rasch based methods (Rasch, 1960). These computer based techniques have been used in the test construction of some recent batteries such as the Battelle Development Battery II (Newborg, 2005) and the Developmental Profile 3 (Alpern, 2007).

Standard Scores at Domain and Total Test Composite Level

The battery provides standard scores at the domain and at the total composite
level. It may be recalled that standard scores express in standard deviation units the extent to which a person's score is higher or lower than the mean score of all individuals in the same age range and are the most easily understood and valid type of derived scores (Anastasi & Urbina, 2008).

Standard scores have many advantages as they are easy to interpret and have exceptional psychometric properties. The main advantage is that they consist of equal-interval units; therefore the difference between the standard score of 90 to 100 is the same as the difference between 50 and 60. This advantage makes it easy to understand and interpret the standard scores. In addition, standard scores are the most statistically valid measurement to use when assessing the child's development over a period of time. For example, a child who at age 4 obtained a standard score of 118 and when retested at age 6 obtains a standard score of 124 can be assumed to be scoring in the above average range at both points of time. On the other hand, a child who obtains a standard score of 100 at age 4 years and at retest obtains a score of 80 can be assumed to be developing at a slower pace.

Standard scores are also the best measure to compare performance across tests. However, one must recognize that parents and teachers and other individuals not trained in psychology may have considerable difficulty in understanding standard scores. For persons, other than professionals, it is recommended that percentile rank scores and descriptive categories be used to communicate results of evaluation for easy comprehensibility.

One must keep in mind several problems when calculating domain total scores, especially at the youngest ages. Since the range of possible behaviours is small during the first few months of life, there are only a few items at the lowest age range available for each scale. It is, therefore, theoretically possible that a child obtains a zero score on both the sub-domains of a domain, and has a total raw domain score of zero. However, a look at the tables in appendix B shows that technically a total domain raw score of zero translates into some DQ, albeit a low DQ. In such a case, the accurate estimation of DQ may be questionable and extreme caution may be used in interpreting such an obtained score. Such a scenario is possible at younger ages and in the markedly delayed children. Despite the best efforts to construct a battery which spans the entire development period and the full ability range from the
very low to high, and inclusion of items which test the ability of all the age ranges for
all the sub-domains and domains, there may still be situations when the very young
and very profoundly delayed children would obtain a score of zero. The examiner
should then exercise caution and sensitivity in interpreting the scores and conveying
the results of the testing to the parents of such a child.

On the other hand, at the other end of the scale, many typically developing
children obtain perfect scores. This is because of ceiling effects as major
developmental milestones level out at the older ages. Therefore, there is a restricted
range of scores at the higher age groups for all the domains, and for children above 7
years, the battery can primarily be used as a means of identifying below-average skills
or delayed skills and for providing assurance that the skills of the typically developing
children are within the normal range. The examiner needs to exercise caution while
interpreting extreme scores for the oldest and the youngest ages or for the most
delayed and the advanced child.

**Percentile Ranks**

The present Development Battery also provides percentile ranks for each of
the standard score for each domain at each level. In addition, percentile ranks are also
provided for the total composite test score. The percentile score represents the
percentage of individuals in the normative sample who scored lower than the child
being evaluated. In the school settings, percentile ranks can provide useful
information to the school teacher. However, percentile ranks have some limitations as
well. The main problem with the percentile score is that the differences between
scores are not the same and the difference depends on where the percentile falls. The
difference between scores is higher at the extreme ends than in the middle. In other
words, the difference between the 50th and 60th percentile score is much smaller than a
performance difference between 80th and 90th percentile score. This is mainly because
the percentile scores are grouped more closely at the middle of the scale than at either
end of the scale. This property can lead to errors in interpreting differences between
two percentile ranks. For this reason, the user must not attach much importance to
large differences in percentile ranks near the centre of distribution. On the other hand,
even small differences in percentile ranks at the two extreme ends of the distribution
are important and should not be disregarded as they translate into bigger differences
in the ability being assessed.
Classification Categories

In order to facilitate understanding of the standard scores at the domain and at the total test level, it is sometimes more useful to categorize the score into a classification category. The present Development Battery provides classifications of the scores and a descriptive label to the level of developmental at which the child is functioning. The classification presented in the present battery is similar to the one provided for other tests and batteries including the Vineland Adaptive Behaviour Scales (VABS, Sparrow et al., 1984) and Wechsler Intelligence Scale for Children (Wechsler, 1991). The standard scores are grouped into ranges and the following classification of scores are provided, in addition, to standard scores, such as very superior, superior, above average, average, low average, and borderline. Children scoring in the mentally retarded range are further classified into mild, moderate, severe, and profound. Provision of classification categories along with standard scores and percentile ranks help the parents and teachers to comprehend the meaning of the standard scores.

Item Gradients

In each of the domains, item gradients from raw scores to standard scores were found to be acceptable as a raw score increase of 1 resulted in no more than 1 to 3 points increase of standard score. The increase in standard scores with an increase in raw score meets the generally accepted criteria (Bliss, 2007). According to Bracken (1987), for adequate item gradients each raw score point should be equal to or less than one-third of a standard deviation. For example, for the present battery, in each of the domains, a raw score increase of 1 point resulted in no more than a 2-point increase in standard score. It seems then that the present battery met the current standards of item gradient and its results are in consonance with well established standardized batteries.

Another point of concern is that because of cutoffs between adjacent age categories in the norm tables, very different scores may result from one day to next when a child’s chronological age falls at a cutoff point. For example, a child 2 years 5 months and 29 days who obtains a raw score of 80 on the Cognitive domain would receive a Cognitive DQ of 92. However, when tested the next day when the child is 2 years and 6 months, the same raw score of 80 on the Cognitive domain would
translate into a Cognitive DQ of 81. In other words, depending on the day of testing this child’s Cognitive DQ can increase or decline by 9 standard scores. This holds true for all domains and at all ages, although the shift in DQs is more at younger ages and much less at the older ages. However, this problem is not unique to the current battery, and previous developmental batteries, such as the BDI-II (Newborg, 2005), also have similar difficulties (Bradley-Johnson, 2001).

It is, therefore, recommended that while testing children at the extreme end of any age group, the examiner needs to also check the standard score for the same raw score for the adjacent age group and take cognizance of the extent of change, and advice and guide parents accordingly. The administrators, therefore, should be aware of this lacuna while interpreting scores at the extreme age groups. In any case, a multidisciplinary assessment approach is necessary while testing and no single test is sufficient for determining eligibility of a child into special program and education. Moreover, the results of assessment must always be interpreted with caution and understanding of the instrument being used (American Academy of Pediatrics, 2001, 2006; Rydz et al., 2005; Sutton, 2006).

**Reliability**

The Development Battery reports the split-half, test-retest, and inter-rater reliabilities; and also the standard error of measurement.

Internal consistency was determined by calculating the split-half reliability coefficients obtained for the five domains and for the total score separately for all the 18 age groups. Previous assessment batteries, such as the Battelle Developmental Inventory II (Newborg, 2005) and the Vineland Adaptive Behaviour Scale (Sparrow et al., 1984), also use the split-half reliability coefficients for assessing internal consistency. It is recommended that the reliability coefficients should be above 0.80 for the sub-domain scores and 0.90 for the domain and diagnostic decisions (Bracken, 1987). Examination of the Development Battery reliability coefficients revealed that total score reliability coefficients ranged from 0.90 to 0.97 across the 18 age groups with a median of 0.94. These coefficients were highly acceptable. However, the domain coefficients were less acceptable and the median coefficients were 0.63 for the Social-Emotional Domain, 0.78 for the Adaptive domain, 0.79 for the Motor domain, 0.79 for the Communication Domain, and 0.89 for the Cognitive Domain.
The median correlations were highest for the Cognitive domain and lowest for the Social-Emotional domain. In general, the correlation coefficients were moderate for the domains but high and satisfactory for the total composite score. It is noteworthy that the split-half reliability coefficients for the total composite score were excellent and more satisfactory than the domain split half estimates.

Although calculating the split-half reliability is the norm for assessment batteries there are some limitations of the split half measure which needs to be acknowledged. The results indicated that the split half reliability could not be calculated at the older ages for the two sub-domains of the motor domain. This is mainly due to the fact that at older ages, the correlation could not be computed because of lack of variability, as all children in the age group obtained perfect scores for all items in one of the split halves within a sub-domain. Moreover, at the youngest and at the oldest age ranges the correlations tend to be affected by the restriction in range of obtained scores. In such cases it is recommended that the median split-half reliability coefficient for that sub-domain or domain be used for ages where the value could not be calculated because of the above mentioned limitation. It may be mentioned, despite this limitation, most assessment batteries have used the split half reliability coefficient as a measure of internal consistency.

Test-retest reliability represents the stability of the Development Battery scores over time and involves administering the test to the same child and same parent on two different occasions and then correlating the scores from each administration. The results indicate that the test-retest reliability coefficients for the domains and the total composite score were very high. For example, the coefficients ranged from 0.93 for the Communication domain to a high of 0.99 for the Social-Emotional domain. The test-retest reliability coefficient for the total test composite score was 0.95 which is excellent. In general, the average differences between scores from the first testing and the second testing were small. The means reported for domains and total test score were generally slightly higher for the second administration. Except for the Cognitive domain, where the retest score increased by one-half of a standard deviation, the other domain scores increased negligibly.

Since objective scoring procedures were defined for each item, the items administered to 30 children aged 1 year to 7 years in a single session, were re-scored
by a trained examiner to create a second set of item scores to compare with the
original examiner’s item scores. Both examiners calculated domain standard scores and
total test composite scores based on their interpretation of the child’s responses to the
test battery. The inter-rater reliability was very high and ranged from 0.94
(Communication domain) to 0.97 (Cognitive domain). The mean difference in the
standard scores was low and ranged from 0.67 (Cognitive domain) to 1.77
(Communication domain). It is interesting to note that the inter rater reliability is as
high as the test-retest reliability.

In sum, an examination of the reliability of the Development Battery reveals
that the internal consistency reliability coefficients for the total score were very high
across all the age groups and were highly acceptable. However, the domain
coefficients were less acceptable and lower than the accepted standard of 0.80 to 0.90.
The test-retest reliability coefficient for the total test composite score was also high
and excellent and the mean differences between scores from the first testing and the
second testing were small. The inter rater reliability was also above 0.90.

Standard Error of Measurement (SEM) and Confidence Intervals

All test scores are subject to measurement error including errors related to
informant’s accurate knowledge about the child’s functioning, skills and
development; over-reporting about child’s skills; misunderstanding about the question
asked and the skill being tested; child’s shyness and fatigue; incorrect scoring of the
item by the examiner; and wrong addition. All these factors contribute to
measurement error. In order to acknowledge the measurement errors, confidence
intervals are provided for each domain at each level as well at the composite score
level. The standard errors of measurement are provided to compute the bands of error
for standard scores for different age levels at the 68%, 85%, 90%, 95% and 99%
confidence level. The bands of error are smaller as the confidence level declines and
are larger at the 99% confidence level. It is recommended that the 95% confidence
interval be used as the common practice for most intelligence tests and development
batteries. It is further recommended that while reporting a standard score it must
always be banded with the amount of error. The range of standard score, plus and
minus the band of error, ideally at 95% confidence interval, should receive more
emphasis than the specific obtained standard score. This implies that the DQ score of
any domain should not be considered as fixed and static, but rather as dynamic. It should be noted that the confidence intervals are higher for the domain DQs and lower for the total DQ.

The standard error of measurement (SEM) values for the standardization sample by age for domain and total test composite score are provided for the battery. The results indicated that the SEM values ranged from 2.10 to 4.65 standard score units (Median = 3.60) for the total composite score, with most of the SEMs ranging from 3 to 4 standard score points. The domain specific SEMs were, however, much higher. Smaller the standard error of measurement, the more sure the examiner is of the score a child obtains on the test is close to the child’s true score. Clearly then, the total DQ is a more reliable measure of child’s true score than the domain DQ.

Validity

Validity refers to the degree to which a meaningful interpretation can be inferred from a test. The content, construct, and criterion related validity were examined for the battery.

Content validity of the battery was assessed by examining the three dimensions: professional judgment of content, coverage of important constructs, and empirical item analysis conducted at the time of the tryout phase. Professional judgment of the content was determined by the opinions of the experts during the test development try out phase. Coverage of important constructs was also determined by experts who examined whether each of the five domains was addressed by the items selected in the initial pool. Item analysis was conducted only in the tryout phase and items with a poor fit were deleted.

Construct validity reflects the ability of an instrument to measure an abstract concept or construct (Gregory, 2005). The construction of the Development Battery was based on several general developmental principles. First, a child who performs well in one domain should perform well on other domains as well. The correlations generally were the highest between the domain and the Total DQ score for all the domains. It is noteworthy that the correlations between the domains are in the low to moderate range. This is not unexpected given the fact that each scale represents one unique aspect of the child’s development. The low to moderate correlations between
domains also indicate that development is inter-related and all the domains are to some extent related to one another. However, each domain has a higher correlation with the Total DQ than with any other domain. This provides support for the separate scoring and interpretation of the five scales.

Second, developmental abilities should increase with age from birth to about eight years with the maturing of the brain (Santrock, 1999). Since the Development Battery was designed to measure developmental growth, differences in the scores across age groups provide evidence of construct validity. This was established by examining the progression of raw scores with age. The results indicated that the raw scores of the various sub-domains and domains consistently increased with the age of the child. The domain curves across age ranges showed that before 3 years the Motor, Adaptive and Social-Emotional curves showed the steepest slope or more growth and then showed a flattening out at older ages. The Cognitive and Communication domains, on the other hand, showed relatively a more flat slope in the initial years and then a steady steep slope till 8 years with no signs of flattening. This demonstrated that the motor, adaptive, and social-emotional development is the most rapid in the earlier years while the cognitive and communication development coincided with the age when the child begins attending pre-school and continued developing till 8 years. The generally positive growth trends for all the development domains, with more rapid development at younger ages, provides evidence for construct validity of the Development Battery.

Factor analyses were employed to gain information for the validity of the structure of the Development Battery. The principal component analyses were conducted with the domain standard scores to confirm the underlying structure of the test and to determine the percentage of variance accounted for by the first principal component. For each age group, the analysis produced one significant factor, using the "eigenvalue equal to or greater than 1.0" criterion, which for the eight age groups accounted for 47.3 to 64.7% of the variance in domain standard scores. This indicated that the total test composite score is an appropriate index of development. Further, principal factor analyses were conducted for the sub-domain raw scores for each of the eight age groups. Intercorrelation matrices of the sub-domain raw scores, with the effects of chronological age removed, were analysed with the resulting factors rotated orthogonally. It is noteworthy that at the higher age groups, the factor
structure becomes clearer, i.e., the sub-domains fall more clearly in the five domains and have a more defined structure. Clearly then, the domain-wise development of the child is more apparent at older than at younger ages, wherein there is some overlap between the developmental domains. The factor structure obtained in the present study, thus, lends support for the rationale underlying the battery and all the sub-domains were found to share primarily their functions as measures of the general development index.

Criterion validity was determined by examining the correlations of the Development Battery with several existing developmental batteries, intelligence tests, and test which measure specific domains. These included Developmental Profile II (DP-II; Alpern et al., 1986), Developmental Profile 3 (DP 3; Alpern, 2007), Malin’s Intelligence Scale for Indian Children (MISIC; Malin, 1969), Draw a Man test (Phatak, 1993), Bender Visual Motor Gestalt Test (Bender, 1946), the Coloured Progressive Matrices (CPM; Raven, 1965), the Vineland Social Maturity Scale (VSMS; Malin, 1971), and the Clinical Linguistic Auditory Milestone Scale (CLAMS; Capute et al., 1986).

Correlations between the domain scores and the DP II (Alpern et al., 1986) were found to be moderate to high. The correlations were the highest between the domains assessing similar content. For example, the Motor DQ had the highest correlation with the score on the Physical domain of the DP II; the Communication DQ had the highest correlation with the score on the Communication domain of the DP II; and the Cognitive DQ had the highest correlation with the score on the Academic domain of the DP II. This illustrated the utility of the present Development Battery in providing meaningful and valid information about the child’s functioning in each of the five areas of development. Similarly, correlations between the Development Battery and the DP 3 (Alpern, 2007) were found to be moderate to high. The correlations were the highest between the domains assessing similar content. The correlation was the highest between the General Development Score on the DP-3 and the total test composite score of the Development Battery.

The correlations between the domain standard scores of the Development Battery and the three IQ scores of the Malin’s Intelligence Scale for Indian Children (MISIC, Malin, 1969) ranged from a low of 0.10 (between Social-Emotional DQ and
PIQ) to a high of 0.75 (between Cognitive DQ and FSIQ). Since, the MISIC is a measure of the child's intelligence one would expect that it would be most closely related to the Cognitive DQ of the battery and in this context, it is noteworthy, the Cognitive DQ was highly correlated with the Full Scale IQ (FSIQ). In addition, the inter-correlations between the total test composite standard scores and PIQ, VIQ and FSIQ were high and were 0.72, 0.79, and 0.83, respectively. Moreover, the total test composite was also very close to the FSIQ. The results therefore indicate that both the cognitive DQ and the composite test score could be used as measures of a child's intelligence. The inter-correlations between the Malin's Intelligence Scale for Indian Children (MISIC, Malin, 1969) and the Development Battery reflected both convergent validity, by the high correlation of 0.83 between the composite test score of the Development Battery and the FSIQ, and divergent validity, by the relatively low correlations of 0.16 between the Social-Emotional DQ of the Development Battery and the FSIQ.

The Development Battery was also correlated with the Draw-a-Man test (Phatak, 1993). Majority of the correlations were found to be in the moderate range. The highest correlation was with the Cognitive DQ and the lowest with the Communication DQ. The mean standard score for the Draw-a-Man test was higher than the Mean test composite score by 9 points, thereby indicating that the human figure drawing test may perhaps over estimate the child's ability to some extent (Malhi et al., 2003).

The validity of the Development Battery was also determined by intercorrelating it to the Bender Visual Motor Gestalt test (Bender, 1946). The inter-correlations between domain DQs and the Bender Visual Motor Gestalt DQ ranged from low to moderate with the lowest correlation found with the Social-Emotional DQ and the highest with the Cognitive DQ. The highest correlation was with the total test composite score. The large range in the correlations was to be expected because the Development Battery assesses development in five different domains whereas the Bender Gestalt test assesses only perceptual motor skill. Since the perceptual motor skill is assessed in the Cognitive battery in tasks involving copying of shapes, the best correlation of the Bender Gestalt test was found with the Cognitive DQ. These results provide further support for the convergent and divergent validity of the Development Battery.
The inter-correlation of the Development Battery with the Coloured Progressive Matrices (CPM, Raven, 1965) test ranged from low to moderate with the domain DQs, lowest being with Social-Emotional DQ and the highest being with the Cognitive DQ. The highest correlation of the CPM was with the total test composite score. The range in the correlations was not unexpected because the Development Battery assesses development in five different domains whereas the CPM assesses only non-verbal abstract reasoning.

In addition to comparing the Development Battery with other tests of intelligence and general development, the battery was also compared to other tests that examined areas which are specific to only one of the domains of the Development Battery such as the Clinical Linguistic and Auditory Milestone Scale (Capute et al., 1986) and the Vineland Social Maturity Scale (VSMS, Malin, 1971). For example, the Vineland Social Maturity Scale (VSMS) measures the adaptive behaviour functioning of children aged 0 to 15 years and it was expected that the Social Quotient (SQ) generated by the VSMS would be most closely related to the Adaptive DQ of the test battery. The results indicated that the inter-correlations between the domain standard scores and VSMS ranged from 0.68 (Motor DQ and SQ) to a high of 0.81 (total test DQ and SQ). Interestingly, the correlations were high and similar for all the domains, and not only to the Adaptive DQ as one would have expected, thereby pointing to the construct of general child development that appears to underlie the different areas of functioning.

The Clinical Linguistic and Auditory Milestone Scale (CLAMS, Capute et al., 1986) assesses the language skills of children aged less than 3 years. The correlations between the Development Battery and the CLAMS score were in the low to moderate range, the lowest correlation of the CLAMS was with the Motor DQ of the Development Battery ($r = 0.26$) and the highest was with the Communication DQ of the Development Battery ($r = 0.61$) as one would expect. The low correlations with the Motor DQ, Social-Emotional DQ, and Adaptive DQ are not unexpected as these domains did not directly assess the communication skills of the child. These results provide strong evidence for the convergent and divergent validity of the present battery in very young children.
Validity of the Battery for Children Belonging to Different Groups

Several moderator variables in the standardization sample were evaluated to determine whether the constructed Development Battery could be effectively used across groups without a bias towards any group. Comparisons of the groups on each of the five domains and the total development score for the entire standardization sample for five moderator variables including sex, residence, education of the father, education of the mother, and socio-economic status was done. The analyses was conducted to compare the standard scores for each of the five domains standard scores and the total test composite score to the average score for the entire standardization sample, which was 100 by definition. Effect sizes were reported to determine whether the statistically significant difference held any clinical meaning. Effect sizes of 0.10 to 0.30 deviation units (approximately 1 to 4 standard score points) were considered small and not considered clinically meaningful, effect sizes of 0.30 and 0.50 deviation units (approximately 5 to 8 standard score points) were considered moderate, and effect sizes greater than 0.50 deviation units (or greater than approximately 8 standard score points) were considered large (Cohen, 1992). Effect sizes that are not small suggest that clinically meaningful differences do exist.

Overall, the standard scores based on the entire standardization sample, stratified by age, applied to a number of different demographic groups, stratified by sex, parental education, place of residence, and socio-economic groups. Differences found between groups were small and did not suggest that they should be interpreted in different ways for different groups or that different norms were needed for children belonging to different groups. This finding is notable in view of the criticisms against many intelligence tests that they are biased against disadvantaged children (Sattler & Hoge, 2005) and the emphasis on non-biased assessment (Kaufman & Kaufman, 1983). Therefore, the present battery seems appropriate for use with all children, including children from marginalized groups.

Principal Uses of the Battery

The Developmental Battery can be used effectively in a variety of settings and for a range of purposes as it provides norm-based scores and information on child’s areas of strengths and weaknesses. The Development Battery provides a wide range of scores from 20 to 160 and can be used to assess typically developing children as
well as children with marked delays and superior functioning. Several reviewers of assessment tests for young children have reported that tests for preschoolers had poor floors and item gradients making them inadequate for testing of delayed children (Bracken, 1987; Flanagan & Alfonso, 1995). Therefore, it is important that assessment instruments for young children have adequate technical and psychometric properties which are clearly stated so that practitioners working with young children can make informed decisions regarding test selection and appropriate interpretation (Emmons & Alfonso, 2005). Research suggests that early identification of and intervention of children with developmental disabilities is essential (Dworkin, 2006; Rosenberg, Zhang, & Robinson, 2008) and therefore this battery is an important contribution towards this end.

At higher ages, its primary use is in identifying skills that are below average and in providing the assurance that the skills of the older children are in normal and above normal range. In this sense, the effective use of the Development Battery is supported till the age of 7. Since the battery tests full range of skills, this is particularly useful when one is working with children who have severe and profound disabilities. Many of the currently available batteries, such as the BDI II (Newborg, 2005) and the DP 3 (Alpem, 2007), provide a limited range of scores and therefore are not very useful for testing children with severe disabilities. For example, the BDI II scores range from 55 to 145 and the scores for DP 3 range from 50 to 130; at the domain level. The problem with tests which have limited floors, may spuriously inflate test scores of children who are severely delayed. Moreover, diagnosis of mental retardation and severity of retardation is not possible in batteries and tests when the floors are inadequate (Bradley-Johnson, 2001).

It is noteworthy, that the present battery is the first assessment battery constructed in India, which provides standardized scores from ranging from 20 to 160, for children across the critical period of development from birth to 8 years of age. Most of the existing developmental tests in India do not provide standardized scores. For example, the most commonly used test in the Indian setting is the Vineland Social Maturity Scale which can be used for children birth to 15 years. However, the scale only yields an age equivalent score which is then converted to a social quotient. In a recent evaluation of the Vineland Social Maturity Scale (VSMs), Malhi et al. (2006) reported that the VSMs has a low sensitivity of only 58% in a clinic based sample,
suggesting that the VSMS because of its low sensitivity would miss nearly 42% of the young children with cognitive deficits. Hence, despite the availability of an Indian adaptation of VSMS, its ability to pick up children with developmental delays falls far short of expected standards. The only development battery in India which yields standardized scores is the Developmental Assessment Scales for Indian Infants (DASII; Misra & Phatak, 1996), however, its use is limited to age of 30 months and it yields standardized scores for only two developmental domains, i.e., Mental and Motor.

The Development Battery is easy to administer and score by a person who is familiar and competent with psychological or educational testing. Interpretation of the scores and application of the results would, however, require a person with understanding and experience in child development, psychology, and psychometrics. The battery can be used by schools, clinics, hospitals, and in any other setting where the evaluation of a child’s developmental status, strengths and weaknesses could be useful.

Examiners can administer the items for each domain separately, or they could test all the five developmental domains. For example, a physical therapist might find only the motor domain useful and could administer only the items of the Motor domain; an occupational therapist might administer items only from the Adaptive domain; a speech-language therapist might have interest in the results from the Communication domain alone and may use items only from the Communication domain; and the child psychologist or a teacher might administer the cognitive items.

Moreover, the battery does not consist of any timed test which makes it useful for children with motor, language, and cognitive disabilities. The inclusion of toys and the use of engaging tasks make testing for the child interesting and counteract the disadvantage of the long administration time of the battery to some extent, particularly for the younger children.

Many authors have argued that the emphasis of testing should not be merely on disability classification but also to provide an appropriate intervention plan in order to remediate early delays (Bracken & Walker, 1997; Bradley-Johnson, 2001; Peterson, Wall, Raikes, Kisker, Swanson, & Jerald, 2004). The present battery can assist with planning individual educational program by identifying the specific skills
in several sub-domains and domains which particularly need to be taught or enhanced. Since there are adequate numbers of items testing various skills, for the entire age, the battery provides adequate sample of skills to be taught in the individual remediation program for each child based on the results of testing.

The number of norm referenced cognitive and development measures available for Indian children are very limited. For example, the most well known test of intelligence, the Malin’s Intelligence Scale for Indian Children (MISIC, Malin, 1969) was developed nearly four decades back. Evidence indicates that IQ keeps increasing every decade and norms become almost 3 points lower per decade, for some tests, (Flynn, 1984; 1999). Although the Flynn effect was originally identified as relating to IQ scores, a similar effect has been shown with standardized scores of development tests illustrating that the child’s development score may be over-estimated if older norms are used. This change is evident in the transition from the original Bayley Scales of Infant Development (BSID, Bayley, 1993) to the BSID-II (Bayley, 2006). For example, the Mental Developmental Index on the BSID-II was 12 points lower than the BSID Mental Developmental Index, and the Psychomotor Developmental Index was 7 points lower. This decrease was also found in other developmental tests, and supported the argument that tests needed to be periodically updated. Several other developmental and motor tests have been found to show similar increases and this is not easily explained (Aylward, 2009). Therefore, most tests developed more than two decades back need revision and re-standardization to avoid the pitfall of inflated scores. Clearly then, dated tests used by professionals for classification purposes can lead to seriously flawed classification results. The more recently a test is developed, the more likely is a test to reflect test items that are based on current empirical studies (Johnson & Marlow, 2006).

Selection of a test warrants careful thought so that the results yielded can be interpreted accurately. Moreover, one must recognize that assessment is a difficult and complex task, and no single test is sufficient for determining eligibility into special program and education. The results of testing must also incorporate information from other sources such as observation, reports from teachers, parental interview, developmental and school history, medical records reviews, results of previous testing (Bradley-Johnson, 2001; Sattler & Hoge, 2005). A multidisciplinary
assessment approach is necessary as conclusions based on a single testing session may not be accurate.

To conclude, the battery can be used for various purposes and can accomplish several assessment and educational objectives. Firstly, it can serve both as an assessment tool for assessing global developmental delay in young children and mental retardation in children above five years. Secondly, it can identify both the areas of strengths and weaknesses of the child which can aid in designing an individualized remediation plan for the child. Thirdly, the wide age range facilitates the use of the battery in longitudinal comparisons of the same child over the entire developmental period. Finally, it can also be used as a plan for intervention by sensitizing parents and school teachers by identifying the developmental tasks in the various domains which the child is deficient in and at which the child needs to work at achieving. It is important to bear in mind that the battery should not be used in isolation while drawing out a remediation plan; it must always be used in conjunction with other data, such as information derived from history, detailed interviews, observation, and the socio-economic milieu to which the child belongs.

**Further Extensions of the Battery**

Establishing validity is an ongoing process, and there is, therefore, a need that future research should test the battery with different populations, especially children who are mentally retarded and developmentally delayed. There is, thus, a need to establish discriminant validity and test whether the battery can successfully distinguish typically developing children from children with clinical conditions like autism, cerebral palsy, and language disorders. Such studies can lend validity support to the ability of the Development Battery to effectively distinguish between relevant groups of children. There is also a need to assess the use of the battery over a period of time in longitudinal research, both before and after interventions. There is, therefore, a need to conduct intervention studies with the battery. One purpose of developing the battery was to provide educational and mental health professionals with information for instruction and remediation. For example, to provide guidelines regarding how to teach children the items that were failed and then to re-evaluate the same children after instruction. Such studies can provide multidimensional
information on a child’s functioning level and the utility of the battery in planning and evaluating instructional programs in a variety of contexts.

Longitudinal research needs to be conducted with the battery in order to establish the predictive validity of the battery. Prediction helps in informing the clinician if early alarm or reassurance, based on early assessment, has any basis. However, while establishing predictive validity, especially with young children; one needs to recognize that prediction is difficult because of numerous factors which include: rapid developmental change, biologic or environmental variables medical, social or developmental interventions, and the fact that testing itself has an impact on the developmental trajectory (Hix-Small, Marks, Squires, & Nickel, 2007). Moreover, lack of attainment of a developmental skill could be due to the skill being emergent, latent, delayed, deficient, or disordered. Prediction may also vary depending on the domain or area of development that is tested. More stable performance is found in those children at high risk who manifest developmental problems. The most unstable performance is found in those children with mild to moderate problems.

The battery provides standardized and percentile scores at the domain and the total test level. The battery, however, provides no scores at the sub-domain level. Further extensions of the battery should provide some score at the sub-domain level as well. In addition, since standard scores are at times difficult to understand by parents and teachers, age equivalent scores may be generated both at the sub-domain and domain level. Although, age equivalents are problematic scores (Anastasi & Urbina, 2008) and do not compare as well to standard scores in psychometric properties (Newborg, 2005), there addition could help in making scores more interpretable for the lay person.

Since the primary use of developmental tests is with disabled children, future research and extensions with the battery need to suggest adaptations for children with disabilities. Another welcome addition to the battery would be to construct a short development screening measure from the battery for use with professionals, such as pediatricians and teachers, who do not have the time and training to administer a long battery. The relatively long administration time and the need for training in
administration, scoring, and interpretation, may not make it easy for use with para-
professionals.

In sum, the results support the reliability of the data obtained by the Development Battery and its content, construct, and concurrent validity. The strengths of the Development Battery and the comprehensiveness of the domains it measures all make a strong case for the use of the present battery with children as a tool for multi-dimensional assessment, particularly in longitudinal studies, determining developmental trajectories and outcomes, and classifying children.

Limitations of the Study

Although the battery has several strengths, there are some limitations of the study which need to be kept in mind while using the battery. The main limitation is the small sample size of 626 children used in the standardization sample. The size of the sample used for standardization was relatively small as compared to other development batteries developed in the West which have used sample sizes of several thousands. However, one may keep in mind that test construction is a massive task and usually requires inputs of several centres and multiple examiners.

The second limitation is that the battery did not utilize a random stratified sampling technique to recruit the participants for the study. Utmost care was taken to enroll equal number of boys and girls, children from both rural and urban areas, and from different socio-economic status in the standardization sample so that it was representative of school going children in North India, nevertheless, there was an over sampling of boys, children from the upper and middle socio-economic status, and from urban areas.

Finally, although considerable effort was made to select items for various sub-domains of development, the battery provides no standard scores at the sub-domain level. Moreover, the reliability coefficients at the sub-domain level and domain level are also lower than the recommended value of 0.80, thereby indicating that scores at the sub-domain and domain level be used with caution. Given this, it is recommended that the total composite score be used as it is the most reliable measure of the child’s performance.