PEDIATRICS®
OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test

William K. Frankenburg, Josiah Dodds, Philip Archer, Howard Shapiro and Beverly Bresnick

Pediatrics 1992;89;91-97

The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test

William K. Frankenburg, MD, MSPH*‡; Josiah Dodds, PhD§;
Philip Archer, ScD¶; Howard Shapiro, PhD; and Beverly Bresnick, MA‖

ABSTRACT. Since the Denver Developmental Screening Test was first published 23 years ago, it has been utilized worldwide and restandardized in more than a dozen countries. Concerns raised through the years by test users about specific items and features of the Denver Developmental Screening Test, coupled with a need for more current norms, have prompted a major revision and restandardization of the test. For the revision, 336 potential items were administered to more than 2000 children. The average number of times each item was administered was 540. Using regression analysis, composite norms for the total sample and norms for subgroups (based on gender, ethnicity, maternal education, and place of residence), were used to determine new age norms. The final selection of the 125 Denver II items was based on the following criteria: ease of administration and scoring, item appeal to child and examiner, item test-retest and inter-rater reliability, minimal "refusal" scores, minimal "no opportunity" scores, minimal subgroup differences, and a smooth step-like progression of ages at which 90% of children could perform the tasks. The major differences between the Denver II and the Denver Developmental Screening Test are: 1) an 86% increase in language items; 2) two articulation items; 3) a new age scale; 4) a new category of item interpretation to identify milder delays; 6) a behavior rating scale; and 7) new training materials. *Pediatrics* 1992; 89:91–97; *Denver Developmental Screening Test; Denver II.*

The Denver Developmental Screening Test (DDST), standardized and published 23 years ago, has enjoyed worldwide utilization since 1967.[1] During that time, resources to evaluate, treat, and educate children who have developmental delays have been vastly expanded. As the test has been used, a number of concerns have been raised: the need for additional language items[2]; the appropriateness of 1967 norms in the 1990s; the difficulty in administering and/or scoring some DDST items; the appropriateness of the test for various subgroups (such as ethnic groups, groups with different genders, groups with various levels of maternal education, and groups with varying places of residence). An additional concern of the test's developers has been the well-intentioned but inaccurate way in which it sometimes has been administered and/or interpreted.[3]

These concerns led to the decision to revise the test, restandardize it, modify its interpretation, develop a new video training program, and emphasize training and periodic evaluation of proficiency. This paper was designed to give the reader an overview of the revision, standardization, and reliability evaluation.

From the Departments of *Pediatrics, ‡Preventive Medicine, ¶Biometrics, and ‖Community Child Development, University of Colorado Health Sciences Center School of Medicine, and the §School of Professional Psychology, University of Denver.

Received for publication Oct 15 1990; accepted May 28, 1991

Reprint requests to (W.K.F.) University of Colorado Health Sciences Center Box C-223, 4200 E Ninth Ave, Denver, CO 80262.

METHODS

Development of Potential Denver II Items

The authors, the Denver Developmental Screening Test trainer from the Colorado Department of Health, and a consultant speech pathologist reviewed the DDST items. Eighty-two items were left unchanged, 21 were revised, and 43 new items were added. Some of the items had multiple parts (such as "stacks blocks": one, two, three, four, etc); each of the subparts was treated as a separate item. In addition, some items could be scored as two items, either by report of the parent or by observation. The result was a pool of 336 potential items. Each item was assigned to one of four domains of development: personal-social, fine motor-adaptive, language, and gross motor. Directions for the administration and interpretation of each item also were developed to address each of the concerns raised above.

Sample Design

Because of financial and time constraints and the need to assure adequate numbers to facilitate valid comparisons between various subgroups, a quota sample was used. Because developmental change in young children proceeds at a more rapid rate than in older children, a quota sample containing larger numbers of children at the younger ages was devised to assure equal precision in determining the ages at which 25%, 50%, 75%, and 90% of children could perform each of the potential items.

Children were drawn from the following demographic subgroups: maternal education (less than 12 grades completed, 12 grades completed, and 13 or more grades completed); ethnicity (black, Hispanic, and white); gender (male and female); place of residence (urban defined as an incorporated place or Census Designated Place and surrounding densely populated territory with a population of 50 000 or more; semirural defined as a population of 2 500 to 50 000; and rural defined as areas neither urban or semirural). Because social class, ethnicity, gender, and place of residence of children are all related to different areas of their development,[4–6] the authors wanted the sample for this study to parallel the distribution of children vis à vis these variables in the Colorado population. Maternal education was used as a rough index of social class, ethnicity was self-indicated by the mother, and the residence categories were based on those utilized by the US Census Bureau and were used to facilitate later projections of the sample to all Colorado.[7]

The final sample used for this standardization actually consisted of two samples: one from Denver County and one from twenty other Colorado counties. The sampling frame for Denver County was divided into the three ethnic groups and further subdivided on the basis of maternal education. Each of these strata was then subdivided into age groups. The sampling frame for Colorado counties other than Denver County was divided into the three residential groups. Each of these was divided further into the three

maternal education groups and then divided into age groups as with the Denver County sample.

Data Collection

Seventeen examiners were trained to administer the potential items until they obtained at least 90% inter-rater reliability on five consecutive tests. At the conclusion of the study the seven examiners who collected more than 80% of the sample were rechecked. Inter-rater reliability prior to data collection varied from 92.4% to 98.2% and, at the conclusion of the data collection, it varied from 93.8% to 97.8%.

For the two samples, 95% of the quota sample was obtained and all cells were filled at least 50%. Many of the quota cells were overfilled, but all the data collected were analyzed. The Denver sample included 1039 children, and the out-of-Denver County sample contained 1057 making the overall total sample of 2096 children.

In addition to identifying which items each child passed and failed, the examiner evaluated and rated the child's test behavior and speech intelligibility. The child's test behavior was rated as follows: compliance (complies, usually complies, rarely complies), interest in surroundings (alert, somewhat disinterested, seriously disinterested), fearfulness (none, somewhat fearful, very fearful), and attention span (attentive, somewhat distractable, very distractable). These were subjective ratings based on the examiner's internal clinical norms of children's behavior. The children's speech was rated as understandable, understands about half, hardly understands any, did not speak. The purpose of adding these ratings, which later were placed on the face of the final test form, is to help the examiner more systematically attend to and note important aspects of a child's style of interacting with his or her environment than was the case in the DDST.

Statistical Analysis

To determine the ages corresponding to 25%, 50%, 75%, and 90% passing each of the 336 items, logistic regression analysis was utilized. Items passable by observation and/or report were analyzed separately for each. After the development of a fitted curve, each curve was subject to a "goodness of fit" test to determine if the fitted curve was a sufficiently good representation of the data.[8] If the goodness of fit statistic was nonsignificant ($P > .05$), the subgroup variables were examined. If the goodness of fit statistic was significant at the 5% level, indicating a poor fit, refitting was done using a modified logistic regression. This consisted of three logistic splines, fitted to the 0 to 35th percentile, the 35th to 65th percentile, and the 65th to 100th percentile, and constrained to join together to create one continuous curve.

To determine if significant differences existed between subgroups (for example, did children whose mothers had completed less than 12 grades of formal education differ from those whose mothers had 13 or more years of education in the age at which they could "name one color"), the statistically significant subgroup variables were identified by running a backward stepwise logistic regression. To determine the "composite" or average percentiles, all subgroups showing statistically significant difference ($P \leq .10$) were weighted to correspond with the prevalence of the subgroups in the Colorado population (based on 1980 Census data).

A "clinically significant difference" between percentiles of a given subgroup and the composite percentiles was defined as a difference between the age at which 90% of the subgroup passed the item and the age at which 90% of the composite group (weighted total group) passed that item being equal to or greater than one tenth of that composite age. For example, for the item "Name One Color," the composite age at which 90% of the standardization sample passed is 3 years and 8½ months, whereas the age at which 90% of the children whose mothers had less than 12 years of education passed is 4 years and 2 months. The difference is 5½ months, which is more than one tenth of the composite age. (One tenth of 3 years and 8½ months is 4.45 months.) Fig 1 is an illustration of this difference.

Results of the behavior and speech ratings were analyzed by age to determine the percent of children having each rating of the test behaviors and speech intelligibility described above. Two of the speech ratings, "Half Understandable" and "All Understandable" were also analyzed to determine 25%, 50%, 75%, and 90% passing norms, These two ratings subsequently were included as items in the language sector of the test.

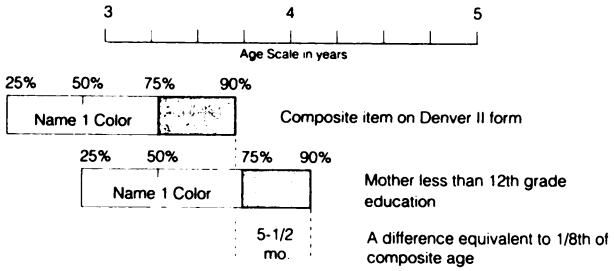

Fig. 1. Illustration of significant clinical difference.

RESULTS

Final Selection of Items

To select items for inclusion in the Denver II, the authors reviewed the following data for each of the potential 336 items:

1. Test materials needed for an item (preferring items requiring a minimum of elaborate materials).
2. Subjective ratings by examiners (preferring items which were easy to administer and score and which were liked by the examiners and children).
3. Percent of "refusal" scores (preferring items which had low refusal rates).
4. Percent of "no opportunity" scores (preferring items which had low "no opportunity" rates).
5. Reliability of items (preferring items which had high reliability; see "Reliability Study").
6. Differences between subgroups (race, gender, maternal education, and place of residence) in the ages at which 90% passed the item (preferring those items which did not show large subgroup differences).
7. Clinically significant difference between norms of a subgroup and composite norms (preferring items which did not show wide differences between the subgroup and the composite norms).
8. Items passable by observation or report (for those items readily observable, making them observation items; and for items either not readily observable or requiring a broader sample of behavior, making these passable by observation and/or report).
9. Distribution of items having 90% age cutoff within each of 12 age categories corresponding to the American Academy of Pediatrics periodicity schedule of 1 to 2 weeks, 2 to 4 months, 4 to 6, 6 to 9, 9 to 12, 12 to 15, 15 to 18, 18 to 24 months, and 2 to 3, 3 to 4, 4 to 5, and 5 to 6 years (attempting to have an equal distribution of 90% cutoffs for each of the age groups).

On the basis of these criteria, the authors selected 125 items for the Denver II. The Denver II test form is included here as Fig 2.

The *Denver II Screening Manual* contains instructions for test administration and interpretation, and recommendations for follow-up.[9] The *Denver II Technical Manual* contains tables depicting the standardization samples, specific characteristics of the potential items such as number of administrations for the standardization, number and percent of "refusal" and

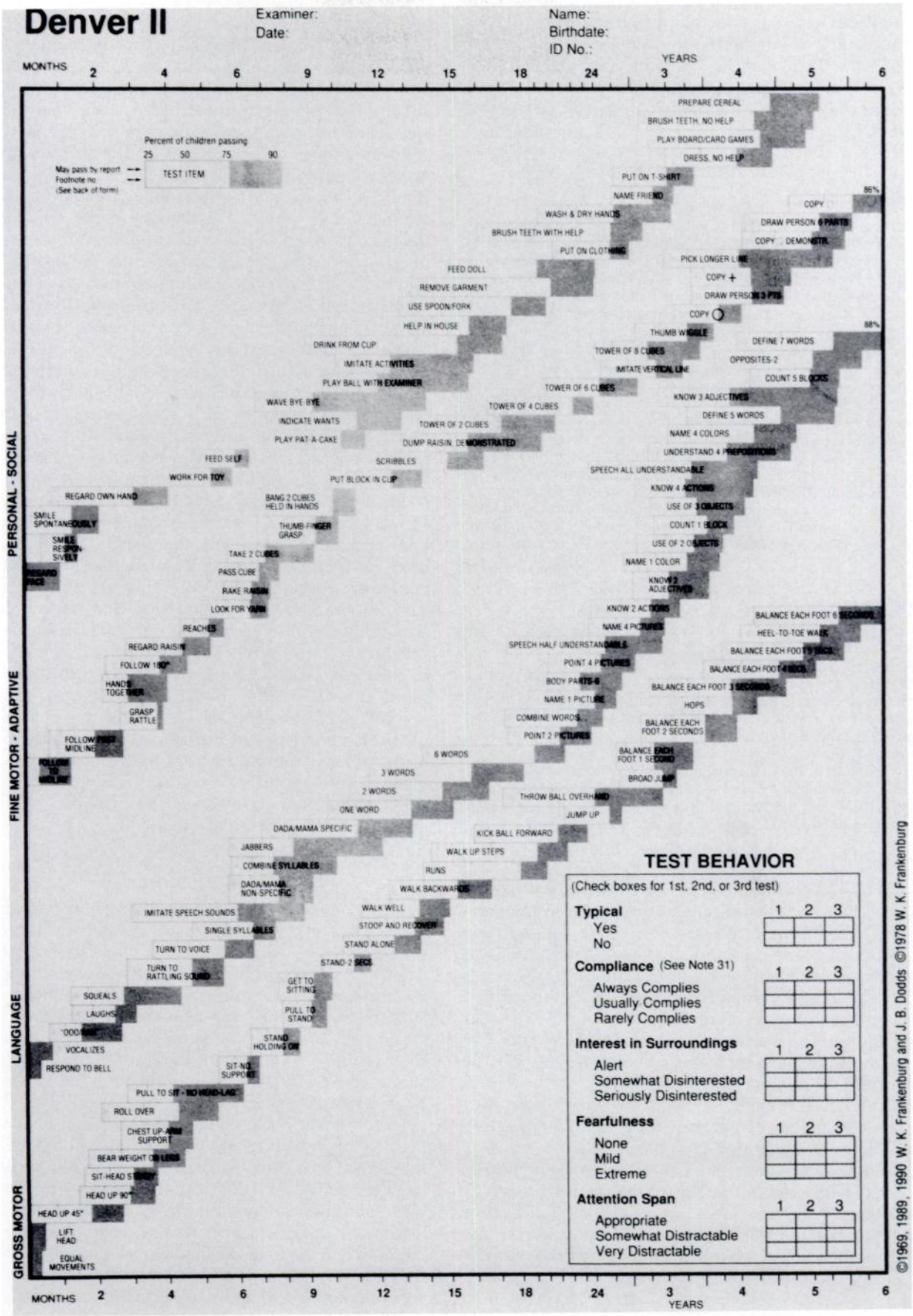

Fig. 2. Denver II test form.

"no opportunity" scores, subjective ratings, covariable differences, and reliability.[10] The technical manual also includes specifics of the sample distribution, behavior and speech ratings by age, κ statistics, and 25%, 50%, 75%, and 90% passing norms and standard errors of the items for which subgroup norms are clinically significantly different from the composite norms.

RELIABILITY STUDY

Method

One of the characteristics of a good test is that the items within the test show a high degree of reliability. A study was undertaken to evaluate the test-retest and inter-rater reliability of each of the potential items and to take the reliability data into consideration in the final selection of items. The reliability sample was divided into the same 10 age groups as those of the standardization sample. Children whose parents had volunteered to participate in the standardization study, but who were not needed for that study, were recruited for the reliability study.

Items assigned for administration in each group were those passed by 25% to 75% of the standardization sample. This was done to avoid spuriously high agreements generated by the administration of items that children of that age obviously passed or obviously failed. For example, the item "Sit No Support," which is passed by 25% at 5.4 months and 90% at 6.8 months, was not administered to 12-month-old infants because there is little chance of disagreement as to whether a 12-month-old child is sitting or not. Six examiners were trained until they achieved a minimum of 90% inter-rater agreement with the trainer on five consecutive children.

Procedure

Thirty-eight children were scheduled for testing on each of two occasions, which were separated by an interval of 7 to 10 days. At the first testing, an examiner administered and scored all items assigned to the age group to which that child belonged, while an observer also scored the items without knowledge of the examiner's scores. After a 5-minute break, another examiner administered and scored the same items while another observer scored. Seven to ten days later, the process was repeated with the order of examiners reversed. That is, the examiner and observer who administered and scored the second test the previous week administered and scored the first test, and the first examiner and observer of the previous week administered and scored the second test.

RESULTS

Of the 38 children tested, 34 came on both occasions. The four types of reliability assessed were inter-rater, 5- to 10-minute test-retest, 7- to 10-day test-retest (same examiner and same observer), and 7- to 10-day test-retest (inter-examiner and inter-observer). The number of comparisons used to determine the percentages of agreement ranged from 12 to 20 per age group. The number of items analyzed was 149 (the multiple parts of items and items passable by reported or observation scores were not analyzed separately). The agreement rate for each of the four types of reliability was generally high with a few exceptions as illustrated in Table 1.

To determine if agreement was better than chance, the κ statistic was calculated for the inter-rater and 7- to 10-day test-retest (same examiner) reliabilities for each of the items.[11] Of the items included in the Denver II, all had excellent inter-rater agreement ($\kappa \geq 0.75$). For the 7- to 10-day test-retest (same examiner) reliability, 59% of the Denver II items had excellent agreement ($\kappa \geq 0.75$), and 23% were in the fair to good range ($\kappa \geq 0.40$).

VALIDITY

The validity of a test is determined by the extent to which it measures what it purports to measure. Thus the validity of the Denver II is established by the precision with which the ages corresponding to 25%, 50%, 75%, and 90% passing for each item and subgroup have been determined. The manner in which the test was standardized on more than 2000 children assures a high degree of face validity in the age placement of the individual items. The Denver II is not a test of some hypothetical construct (e.g., intelligence or physical dexterity), it simply defines the ages at which children accomplish a broad variety of specific tasks. As such, construct validity is not applicable. The Denver II is comparable with a growth curve in that both specify the ages at which children achieve milestones such as development or growth.

With regard to validity of test interpretation and referral criteria, the criteria for Normal, Questionable, and Abnormal test results were established arbitrarily to satisfy demands for an overall test rating. Appropriate referral criteria await further study, because referral criteria will not only vary among areas such as language, gross motor, fine motor, or personal-social, but also on the basis of degree of delay. Under Public Law 99-457 Part H each state is to develop its own criteria. To date, few states use similar criteria; and, furthermore, it is anticipated that state criteria may not be applied uniformly throughout each state. The authors therefore recommend that communities, depending on their resources and aspirations, devise criteria that most efficiently identify children eligible to receive services. Although one can argue that referral criteria should not be based on the potential restriction of available services, screening for delays for which services are not available should only be undertaken if the parents are first told of this unavailability to avoid the unethical raising of parental expectations that services will be provided when in actuality they are unavailable.

DISCUSSION

Ideally, a child developmental screening test covers all areas of development (e.g., social, physical, cognitive, self-care, etc), has construct validity with more extensive standard measures of child development, is uniformly interpretable across all populations and subgroups, has optimum age-specific sensitivity and specificity, and has good predictive validity of future development. What is needed, in short, is a good single test of development similar to the screening test for phenylketonuria.

Unfortunately the developmental phenomena we are dealing with here are not unidimensional. Development of different skills can be quite independent of other skills, and development of a single skill is usually influenced by multiple interactive factors. Developmental rates are determined by a variety of factors such as heredity, biological intactness, emotional health, physical and psychosocial environment. Deviations in developmental profiles are usually the result of multiple etiological factors. For instance, a child suffering from a hearing loss will have a differ-

TABLE 1. Number of Items in Various Mean Percentage of Agreement Ranges by Type of Reliability

Type of Reliability	Mean Percentage of Agreement				
	100%	90–99%	80–89%	50–79%	<50%
Inter-rater	141	7	1	0	0
5–10 min test-retest	59	35	37	18	0
7–10 d test-retest (same examiner and observer)	54	32	39	23	1
7–10 d test-retest (inter-examiner and observer)	50	26	39	33	1

ent developmental profile than a child suffering from cerebral palsy or social deprivation. In brief, the state-of-the-art in measuring such complex phenomena as child development is still quite primitive. This study is a continuation of the beginning stage of developing the art.

The Denver II is designed to reflect the development of a broad range of heterogeneous skills. The data are presented as age norms, similar to a physical growth curve. As such, the Denver II is not designed to measure any single, or even a few, underlying hypothetical constructs such as intelligence, motor functioning, social facility, or communication skill. The authors specifically designed the Denver II (as the DDST) so that neither the total test nor any of the four sections of the test could be scored with a single number such as an intelligence quotient or developmental quotient.

Because the Denver II does not purport to measure unitary underlying constructs, convergent validity studies are meaningless. For example, the construct underlying standard intelligence tests, general intellectual ability, underlies only a few of the Denver II items. Many items which *might* reflect some underlying construct, such as gross and fine motor development, social development, and communication skill, do not have widely accepted standard measures, which means that, with our current state of child development measurement, construct validity studies are impossible.

A more heuristic way of studying the validity of a developmental screening test is to measure its sensitivity in detecting children who have significant deviation in one or more areas of development and its specificity in not generating false positive results. These authors have chosen to approach such research via definitions *a priori* of how many delay and caution items a given child has before being considered "abnormal" (highly suspect) or "questionable" (suspect) in development. These definitions were not set arbitrarily but were based on the clinical judgment of the authors and other screeners in diverse parts of the United States. To illustrate, if a child fails one item which 75% or even 90% of children of the same age pass, it may be a fluke; if the child fails 10 such items, one would naturally be more concerned. Thus the authors rated a child's test performance as "questionable" if the child had one delay (defined as a child failing an item which 90% of his/her age mates pass) and/or two or more cautions (defined as a child failing an item which between 75% and 90% of his/her age mates pass), and rated a child's test performance as "abnormal" if the child had two or more delays. Using the analogy of the growth chart, the level of delay between the third and tenth percentile may be "suspect" and level of delay less than the third percentile may be highly suspect. Applying these definitions in different communities with different populations yields different rates of "suspect" and "highly suspect." As in the case of most phenomena in nature, the prevalence of problems will vary in different populations. For example, the rate of children's language acquisition is related inversely to the number of years of their parents schooling. Thus, a good test of development will produce different rates of children manifesting "suspect" language development in different populations. Subsequent research and field use may lead to redefinition of "abnormal" and "questionable." To date this has been undertaken in three states in quite different parts of the United States with very different populations and yielded different rates of suspect and highly suspect.[10] This difference does reflect true differences in rates of development among these diverse populations.

Although the DDST has enjoyed widespread use, it has been strengthened with the development of the Denver II. One criticism of the DDST, low predictive accuracy, could be made of most infant and preschool tests of child performance. This is true because functions measurable at a young age are far more limited than those that can be measured at a later age. Another reason is that intervening events such as accidents, illness, and stimulation programs are likely to alter the rate of a child's development. It is for these reasons that one should interpret the developmental status of a child at one point in time with caution. It is more important to look at the rate of development over time and interpret the results together with what is known about the child's background. The authors of the test agree with Dworkin and the American Academy of Pediatrics, who suggest that a developmental screening test be utilized as an aid to surveillance of the child's ongoing development.[12,13]

It is important to emphasize that the Denver II is a screening test, the results of which should be integrated with everything else that one knows about the child; the family, the community, the educational experiences, and the culture in which the child has grown up. It is essential that those using this test interpret the test results in the context of the larger picture of the child.

To assist in this process, the *Denver II Technical Manual* contains the norms of items for which there are clinically significant differences between the norms of one or more subgroups and composite norms.[10] The reason that these items are not identified

on the screening form but are identified for health professionals in the technical manual is that a child's delay on any one item can be attributed to a variety of factors such as a hearing loss, brain damage, ethnic group, etc. It is the view of the authors, therefore, that determining a cause for a delay is a diagnostic process that should be conducted by professionals and not the nonprofessional staff who may be the screeners.

One caution is that the test be utilized only for the purpose for which it has been designed, namely to give a brief overview of the child's development reflecting the child's biological intactness and past experiences. It is *not* designed to yield a Developmental Quotient, nor is it designed to predict later learning disabilities, emotional problems, special education placement, etc. The Denver II is designed simply to identify children who are not "up to snuff" or not performing as their age mates, for whatever reason. If a child is not acquiring skills at the normal time (as are his/her age mates), then the child is considered to be at greater risk of having a biological or environmental condition which would interfere seriously with future development. Children who manifest developmental deviance early in their lives must be studied carefully to determine the etiological factors underlying such deviance.

Reliability data of the examiner's test behavior and speech intelligibility ratings were not collected because of the subjectivity of the ratings; they were not considered test items per se. In retrospect, the authors realize that collecting data on and examining the reliability of these ratings would help test users know more about them. Certainly the speech intelligibility ratings, which eventually were placed as two developmental items in the final test form, need reliability data similar to those collected for the other 123 final items. Hopefully such studies will be undertaken in the future.

Long-range research is planned to establish a series of valid criteria for scoring the Denver II to accurately and presumptively identify children who suffer from significant developmental deviations for each of a variety of etiologies. Final criteria for treatment and/or educational intervention need to be developed locally to maximize agreement between screening findings and the provision of educational/habilitative services since the availability of such services varies from one community to the next. This variation is attributed in part to service eligibility in most states that is based on a group decision and test results taking into account the child's background and ethnicity. Communities are being invited to help explore ways the Denver II can be most useful to them in detecting and helping children with developmental problems.

A question frequently asked pertains to the differences between the DDST and the new Denver II. The ten major differences follow:

1. The Denver II was standardized in 1988 and 1989 on 2096 children from all over Colorado: the DDST was standardized on 1036 Denver children in 1966.
2. The Denver II age scale corresponds to the periodicity schedule for health maintenance visits of the American Academy of Pediatrics; the age scale of the DDST does not correspond to this schedule.
3. The Denver II has 86% more language items than the DDST.
4. The Denver II has 20% fewer parent-report items than the DDST.
5. The Denver II has two speech intelligibility items; the DDST does not.
6. The Denver II does not contain items from the DDST that were difficult to administer and/or interpret.
7. The Denver II test form has a checklist for noting behavioral observations; the DDST does not.
8. The average inter-rater and test-retest reliabilities of the Denver II items were 0.99 (SD 0.01) and 0.90 (SD 0.12), respectively, higher than those of the DDST items.[14]
9. The Denver II designates "caution" items (failed items that 75% to 90% of children in the standardization sample passed). The DDST has no such designation.
10. The Denver II identifies items for which there is a clinically significant difference between the norms of one or more subgroups and the composite norms of the total sample; the DDST does not identify such items.

Concerns have been raised that the increase in the number of items from 105 in the DDST to 125 in the Denver II might increase testing time; this has not proven to be true in the initial field tests involving 300 to 400 children. There is no substantial increase in testing time required for the Denver II because many of the items are progressive; that is, one administration makes it possible to score more than one item, such as "Draw a Person, 3 Parts," Draw a Person, 6 Parts," and "Name 1 Picture," "Name 4 Pictures." As with the DDST, the test takes more time to administer to older children. Currently the authors are cross-validating an abbreviated administration of the test requiring only 4 to 7 minutes depending on the age of the child.

A concern of the authors raised at the beginning of this paper is the inaccurate way in which the DDST is sometimes administered and/or interpreted. To minimize such errors with the Denver II the following steps have been taken:

1. The *Denver II Screening Manual* has been developed for use by screeners to give details on the proper administration and interpretation of the tests.[9]
2. To aid those learning the test, The *Denver II Screening Manual* contains a self-evaluation with answers and references for the answers to each question.[9]
3. The *Denver II Technical Manual* contains, in addition to details of the standardization, chapters on training in the administration of the test and on

the establishment of a community screening program.[10]

4. A video instructional program and proficiency test have been developed for the Denver II.

SUMMARY

The DDST has undergone a major revision and restandardization, the Denver II. Changes include not only an update in norms, but the removal or modification of troublesome DDST items, an 86% increase in language items, and the addition of a subjective behavior rating scale. A number of steps have been taken to assure correct administration and interpretation of the test as well. Finally, field tests of the Denver II indicate that the time required for complete administration of the Denver II is not much longer than that of the DDST. The reader is reminded that this face-valid screening test is a first step in tackling the problems of early detection, diagnosis, and treatment of developmental deviations in children. Research is ongoing to develop an abbreviated administration of the test and to develop the best scoring system for specific communities using various specific criterion tests chosen by each community or state.

ACKNOWLEDGMENT

Work was done at University of Colorado Health Sciences Center.

REFERENCES

1. Frankenburg, WK. The Denver approach to early case finding. In: Frankenburg WK, Emde RN, Sullivan JW, eds. *Early Identification of Children at Risk*. New York, NY: Plenum Press; 1985:135–158
2. Borowitz KC, Glascoe JP. Sensitivity of the Denver Developmental Screening Test in speech and language training. *Pediatrics*. 1986;78:1075–1078
3. Frankenburg WK, Chen J, Thornton SM. Common pitfalls in the evaluation of developmental screening tests. *J. Pediatr*. 1988;113:1110–1113
4. Bayley N. Comparisons of mental and motor test scores for ages 1–15 months by sex, birth order, race, geographical location, and education of parents. *Child Dev*. 1965;36:379
5. Golden M, Birns B. Social class and cognitive development in infancy. *Merrill-Palmer Q*. 1968;14:139
6. Sandler L, Van Campen J, Ratner G, et al. Responses of urban preschool children to a developmental screening test. *J Pediatr*. 1970;77:775
7. Bureau of the Census, US Department of Commerce. General population characteristics: Colorado. In: *Characteristics of the Population*. PC 80-1-B7
8. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115:1
9. Frankenburg WK, Dodds J, et al. *Denver II Screening Manual*. Denver, CO: Denver Developmental Materials, Inc; 1990
10. Frankenburg WK, Dodds J, Archer P. *Denver II Technical Manual*. Denver, CO: Denver Developmental Materials, Inc; 1990
11. Fleiss JL. *Statistical Methods for Notes and Proportions*. 2nd Ed. New York, NY: John Wiley & Sons; 1981
12. Dworkin PH. Developmental screening—expecting the impossible? *Pediatrics*. 1989;83:619–622
13. American Academy of Pediatrics, Committee on Practice and Ambulatory Medicine. Recommendations for preventive pediatric health care. *Pediatrics*. 1988;81:466
14. Frankenburg WK, Camp BW, et al. Reliability and stability of the Denver Developmental Screening Test. *Child Dev*. 1971;42:1315–1325

MAGNETIC RESONANCE IMAGING OF SEVERELY DISTURBED CHILDREN—A PRELIMINARY STUDY

Robert L. Hendren, DO, Janet E. Hodde-Vargas, MS, Luis A. Vargas, PhD, William W. Orrison, MD, and Lance Dell, MD

Abstract. This study investigates the relationship between brain pathology and psychiatric disturbance in 37 psychiatric inpatients between 5 and 14 years of age referred for magnetic resonance imaging (MRI). Of 37 images, 13 were categorized as abnormal by neuroradiologists who were blind to the diagnoses of subjects. Three of six children with schizophrenia-related diagnoses had abnormal scans. In contrast, only one of 15 children with a primary disruptive behavior disorder diagnosis had an abnormal magnetic resonance image. A greater proportion of children with schizophrenia spectrum diagnoses had greater left than right frontal horns of the lateral ventricles than children with other diagnoses. *J Am Acad Child Adolesc Psychiatry*. 1991;30.3:466–470. Key Words: MRI, schizophrenia spectrum, structural abnormalities.

The Denver II: A Major Revision and Restandardization of the Denver Developmental Screening Test

William K. Frankenburg, Josiah Dodds, Philip Archer, Howard Shapiro and Beverly Bresnick

Pediatrics 1992;89;91-97