5-1986

The Stability and Predictive Validity of the Bayley Scales of Infant Development: A Meta-Analysis

Gary A. Goodrich
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Psychology Commons

Recommended Citation
https://digitalcommons.usu.edu/gradreports/892

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact rebecca.nelson@usu.edu.
THE STABILITY AND PREDICTIVE VALIDITY OF THE
BAYLEY SCALES OF INFANT DEVELOPMENT:
A META-ANALYSIS

by

Gary A. Goodrich

A report submitted in partial fulfillment
of the requirements for the degree
of
MASTER OF SCIENCE
in
Counseling Psychology
(Plan B)

Approved:
ACKNOWLEDGMENTS

I would like to thank Dr. Glendon Casto for his patient support. I very much appreciate the gentle but convincing suggestions he has made throughout this project. It has been a pleasure for me to work with him and to learn from his political knowledge, research acuity, and practical sensibility.

I would also like to thank Sheryl, my wife. She has managed the amazing task of financing and completing her own education while offering me much needed emotional and intellectual support. Her kind motivations and soft humor have been inspiring and delightful.

Gary A. Goodrich
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>OBJECTIVES</td>
<td>5</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>6</td>
</tr>
<tr>
<td>Stability, Predictive Validity, and Reliability</td>
<td>6</td>
</tr>
<tr>
<td>Summary of Definitions</td>
<td>9</td>
</tr>
<tr>
<td>Stability of Infant Development</td>
<td>10</td>
</tr>
<tr>
<td>Predictive Validity of Infant Scores</td>
<td>11</td>
</tr>
<tr>
<td>Meta-Analysis</td>
<td>14</td>
</tr>
<tr>
<td>Previous Reviews</td>
<td>15</td>
</tr>
<tr>
<td>METHODS</td>
<td>24</td>
</tr>
<tr>
<td>Coding Sheet Development</td>
<td>25</td>
</tr>
<tr>
<td>Conventions for Coding</td>
<td>27</td>
</tr>
<tr>
<td>Coding Reliability</td>
<td>30</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>31</td>
</tr>
<tr>
<td>RESULTS</td>
<td>33</td>
</tr>
<tr>
<td>Summary</td>
<td>58</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>61</td>
</tr>
<tr>
<td>Future Research</td>
<td>62</td>
</tr>
<tr>
<td>Summary</td>
<td>63</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>64</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>71</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MDI Cell Values</td>
<td>34</td>
</tr>
<tr>
<td>2. PDI Cell Values</td>
<td>36</td>
</tr>
<tr>
<td>3. Sample Size and MDI Correlations</td>
<td>38</td>
</tr>
<tr>
<td>4. Sample Size and PDI Correlations</td>
<td>38</td>
</tr>
<tr>
<td>5. Subject Attrition and MDI Correlations</td>
<td>39</td>
</tr>
<tr>
<td>6. Subject Attrition and PDI Correlations</td>
<td>39</td>
</tr>
<tr>
<td>7. Mean Pre-Test Scores and MDI Correlations</td>
<td>41</td>
</tr>
<tr>
<td>8. Mean Pre-Test Scores and PDI Correlations</td>
<td>41</td>
</tr>
<tr>
<td>9. Pre-test Standard Deviations and MDI Correlations</td>
<td>43</td>
</tr>
<tr>
<td>10. Pre-test Standard Deviations and PDI Correlations</td>
<td>43</td>
</tr>
<tr>
<td>11. Mean Post-Test Scores and MDI Correlations</td>
<td>44</td>
</tr>
<tr>
<td>12. Mean Post-Test Scores and PDI Correlations</td>
<td>44</td>
</tr>
<tr>
<td>13. Handicap and MDI Correlations</td>
<td>45</td>
</tr>
<tr>
<td>14. Handicap and PDI Correlations</td>
<td>45</td>
</tr>
<tr>
<td>15. Race and MDI Correlations</td>
<td>47</td>
</tr>
<tr>
<td>16. Race and PDI Correlations</td>
<td>47</td>
</tr>
<tr>
<td>17. Gender and MDI Correlations</td>
<td>49</td>
</tr>
<tr>
<td>18. Gender and PDI Correlations</td>
<td>49</td>
</tr>
<tr>
<td>19. Socio-Economic Status and MDI Correlations</td>
<td>51</td>
</tr>
<tr>
<td>20. Socio-Economic Status and PDI Correlations</td>
<td>51</td>
</tr>
<tr>
<td>21. Reported Intervention and MDI Correlations</td>
<td>52</td>
</tr>
<tr>
<td>22. Reported Intervention and PDI Correlations</td>
<td>52</td>
</tr>
</tbody>
</table>
23. Intervention Duration and MDI Correlations ............... 53
24. Intervention Duration and PDI Correlations ............... 53
25. Intervention Intensity and MDI Correlations ............... 54
26. Intervention Intensity and PDI Correlations ............... 54
27. Pre-Test Examiner and MDI Correlations ............... 56
28. Pre-Test Examiner and PDI Correlations ............... 56
29. Post-Test Examiner and MDI Correlations ............... 57
30. Post-Test Examiner and PDI Correlations ............... 57
31. Blind Post-Test Examiners and MDI Correlations ............... 59
32. Blind Post-Test Examiners and PDI Correlations ............... 59
INTRODUCTION

The Bayley Scales of Infant Development (Bayley, 1969) is the most widely used infant assessment instrument in the United States. Reasons for its widespread use include research into the nature and development of intelligence, studies of intervention efficacy, and treatment and placement of handicapped infants.

To this date, the stability and predictive validity of the Bayley scales have not been well established. According to Webster and Bates, "There remains a serious lack of research examining the stability of all general intelligence measures over extended intervals of time" (1977, p. 5). Anastasi (1976) suggests that tests like the Bayley are particularly useful for the early detection of neurological or sensory defects. Conversely, Caldwell concluded that tests given in infancy describe very well but are inadequate at diagnosis.

Bayley (1958) claimed that the use of infant tests for research in human development is appropriate and justified, despite low predictive validity. Cronbach (1967) noted that correlation matrices involving infant measures follow a simplex pattern (value highest near the major diagonal, decreasing uniformly with distance therefrom) and are therefore useless for factor analyses of qualitative stages of
mental development. Lewis and McGurk (1972) did not find a simplex pattern of correlations in their study, but argued that infant tests should not be used to assess intervention efficacy. "Simply stated, infant intelligence scales are unsuitable instruments for assessing the effects of specific intervention strategies" (Lewis & McGurk, 1972, p. 1176).

Bayley wrote, "It seems evident that the very nature of intelligence in children under two or three years is such that tests in these early years will have little if any predictive value" (1955, pp. 132-133). McCall (1979), Lewis and McGurk (1972), and Gannon (1968) have concluded that the Bayley and other measures of infant development are of very limited practical use for predicting later intelligence in normal children.

McCall, Hogarty, and Hurlburt (1972) reviewed eight studies, most published prior to 1960, and systematically examined the effect of age at pre-test, age at post-test, and the interaction of the two on pre/post correlations. They found that measures taken before seven months yielded negligible correlations with scores obtained after the age of four years. However, other important concomitant variables, such as instrumentation, restriction of range, and type of subjects, were not considered.

Erickson (1968) argued that standardization procedures in many studies have resulted in the underestimation of predictive validity for the population. "What can be concluded from the early studies is that prediction from infancy is not possible for normal middle class children" (Erickson, 1968, p. 728). The subjects in her study were children with suspected developmental delays. Correlations from
pre-tests on the Cattell Infant Intelligence Test at ages 7 and 19 months to Cattell and Stanford-Binet scores at 19, 31, and 43 months were .72, .80, and .80. She concluded, "The results of this study gave evidence that the Cattell Infant Intelligence Scale was useful for predicting the later IQ scores of children referred to a clinic for possible developmental problems" (1968, p. 732).

Similar results have been found for the Bayley scales. VanderVeer and Schweid (1974) studied 23 infants with serious developmental delays. The Bayley Mental Development Index (MDI) was given at a mean age of 24 months (range=18 to 30 months). In many cases ratio quotients were computed because raw scores were below the norms. Subjects were retested between one and three years later (mean=25 months), using either the Bayley or the Stanford-Binet. The resulting Pearson correlation was .97.

It should be noted that in many studies children with very low performance are omitted because Bayley norms only allow deviation scores at or above 50. An example of such a study is an investigation by Goffeney, Henderson, and Butler (1971) in which the Bayley Scales of Infant Development were administered to 621 eight-month-old infants from the Portland area. At seven years the children were retested on the Wechsler Intelligence Scale for Children and the Bender Gestalt. Correlations of the eight-month Bayley MDI with full IQ taken at seven years were .30 (Black females), .28 (White females), .01 (Black males), and .16 (White males). Slightly higher correlations were reported with the Bender. However, children with severe retardation or neurological impairment were excluded from the study. The authors
concluded that eight-month Bayley scores were minimally useful in predicting seven-year IQs.

Much research has been done with the Bayley. In the Mental Measurements Yearbooks (Buros, 1978) there are 59 references for the Bayley Scales of Infant Development. Thus, a large quantity of data exists which reflect on the stability and predictive validity of the Bayley.

The problem is that although a considerable body of data exists, the stability and predictive validity of the Bayley are not well established. If the Bayley is not predictive of later performance, its use in equating groups for longitudinal studies may be questionable. If the scores are not stable, they may be of limited value in documenting intervention effectiveness. If correlation matrices with Bayley scores follow a simplex pattern, they should not be used in factor analyses of qualitative developmental changes.

Meta-analysis is a review process that allows for the systematic and comprehensive collection of significant data from existing literature. Reported here is a study in which all available data pertaining to the stability and predictive validity of the Bayley Scales of Infant Development were systematically collected and analyzed, using the techniques of meta-analysis.
OBJECTIVES

The primary objective of this study was to analyze an extant data base to draw conclusions about the stability and predictive validity of the Bayley Scales of Infant Development. To accomplish this several tasks were completed.

Task 1: Available publications regarding stability and predictive validity of the Bayley scales were identified through an aggressive library search.

Task 2: Reviews of stability and predictive validity of the Bayley and other infant measures were examined. Noted from these reviews were variables which might influence correlations, common methodological weaknesses, and general conclusions.

Task 3: A coding sheet was developed, based on the factors identified in the review of reviews. Conventions for the use of the coding sheet were also developed. Then all obtained articles which contained original correlational data, with the Bayley as pretest, were coded according to the established conventions.

Task 4: The data which resulted from the coding of original primary research were analyzed using regression techniques and descriptive statistics.
Stability, Predictive Validity, and Reliability

Because the psychometric concepts of predictive validity, test/retest reliability, and stability are so crucial to this paper, they will be discussed in terms of theoretical and practical significance. Then operational definitions will be presented.

Validity is a characteristic of an instrument, often defined as the degree to which a test measures what it is purported to measure. One type of validity, predictive validity, is determined by the correlation of scores on a given instrument with later performance on some other measure. There is no standard amount of time that must elapse between tests in order for correlations between tests to be considered predictive. There must be sufficient time to allow for substantial changes in the individual, so that situational and random fluctuations are not the only sources of the variance between the sets of scores. Due to the extremely rapid development of infants, an interval of one month may be sufficient to meet this requirement.

Reliability is a characteristic of an instrument, often defined as whether a test measures the same thing repeatedly. Anastasi notes, "The concept of reliability is generally restricted to short-range, random changes that characterize the test performance itself rather than the entire behavior domain that is being tested" (1976, p. 112). Test/retest reliability may be obtained by calculating a correlation between pre- and post-test scores, with an interval between tests short enough to disallow substantial changes in the individuals being
tested. Test/retest reliability is distinct from predictive validity in that the interval between tests is shorter and one instrument is used for both assessments.

The term "stability" has two meanings, one theoretical and the other practical. Theoretically, stability is a characteristic of a construct, not of an instrument, and refers to the change or lack of change in that construct over time. Practically, stability "refers to the preservation of an individual's rank ordering within a group on some behavioral measure(s) when the measurements are made across time" (Dunst & Rheingrover, 1981, p. 50). The traditional statistical index of stability is a cross-time correlation coefficient computed from the scores obtained on two measurement occasions (Dunst & Rheingrover, 1981).

Emmerich (1964) identified two ways to characterize change in development. "The first considers behavioral continuity over time and asks if needs, acts, cognitive operations, etc., are essentially the same at various periods of development" (Emmerich, 1964, p. 311). This type of developmental consistency is often termed "continuity" in current discussions. "The other approach defines the continuity issue in terms of individual stability. Here, the essential question is whether the distinctiveness of the individual relative to others is maintained throughout development" (Emmerich, 1964, p. 312).

Dunst and Rheingrover (1981) argued in support of Emmerich's position. They contend that stability is an issue of interindividual rank, while continuity relates to intraindividual changes. McCall (1979) drew a distinction, based on Emmerich's 1964 suggestions,
between continuity, described as the amount or frequency of an attribute over time, and stability, the relative consistency of individual differences across ages. This position was recently restated by Ulvund (1984).

Kagan (1980) extended the classification scheme, defining four types of stability. The first is the persistence of a psychological quality, as reflected in minimal change over time. This is a theoretical definition of stability. Ipsative stability, the second type, is the persistence of a hierarchical relation between complementary dispositions within an individual. As such, this would be a subset of continuity in the more broad framework, described above. Normative stability, Kagan's third type, is the preservation of a set of individual ranks on a quality within a constant cohort. This is stability in a practical sense, based on correlations between two sets of scores on a given instrument. Kagan's fourth type of stability is the necessary and contingent relation between phenotypically different structures or functions at two points in time due to the operation of specifiable processes. This category is unique to Kagan and is not represented in other formulations.

Kerlinger (1973) and Borg and Gall (1979) present a simpler formulation of stability, describing it as being synonymous with test/retest reliability. However, this definition ignores the possibility that an instrument may well be reliable, while the construct it measures is highly variable. Furthermore, as Buss (1979) notes, to define stability as the consistency of interindividual differences on a single variable through time allows only for
inferences about patterns or shapes of multiple scores over time.

Thus, it would be possible to have absolute changes in scores yet obtain a high stability coefficient so long as there was relatively high invariance of the pattern of interindividual differences through time. (Buss, 1979, p. 50)

Cattell (1949) presented a coefficient of pattern similarity that takes into account pattern, mean, and scatter. This coefficient is computed from squared differences of standard scores, with a second factor derived from chi-square values. The resulting $r_p$ is intuitively similar to the Pearson $r$, with values ranging from 1 to -1.

The coefficient of pattern similarity, although potentially representing a very interesting contribution to psychometrics, has been essentially ignored by test constructors and mental measurements researchers. In conducting the library research necessary for this paper, 18 articles were identified which presented correlation coefficients between the 1969 revision of the Bayley and later measures. Numerous articles dealing with predictions from other infant tests were also reviewed. In none of the primary research publications was the coefficient of pattern similarity computed or discussed, nor were any other statistical alternatives to the traditional correlations found.

Summary of Definitions

Stability is a characteristic of a real or hypothetical construct, referring to the change or lack of change of that construct over time. In practice, however, stability is most frequently defined by a product moment correlation. It is distinct from test/retest
reliability in that the time intervals involved are greater. It is distinct from predictive validity in that predictions are made to the test itself. For the purposes of this report, stability is the correlation of two sets of scores from one instrument, for one group of individuals, with an interval between tests of at least one month. Predictive validity is the correlation between two sets of scores, from two distinct instruments, for one group of individuals, with an interval between tests of at least one month.

**Stability of Infant Development**

Anderson (1939) proposed an overlap model to account for obtained correlations between repeated administrations of developmental measures. The model predicts a curvilinear pattern of correlation coefficients to percent of occurred development, with correlations approximately equal to the square root of the ratio of intermediate measures (pre-test) to terminal status (post-test). Essentially the hypothesis states that development, once it has occurred, is set and that subsequent gains are unrelated to achieved status. Thus, the percent of variance explained by a correlation is also roughly equivalent to the percent of development which has occurred, relative to the post-test. This implies that there is no actual prediction of subsequent development; interage correlations merely reflect the percent of post-test development which had occurred by the time of the pretest.

To illustrate the hypothesis, Anderson created 96 cases of "scores". For each case playing cards were used to randomly generate
16 independent numbers, ranging in value from 1 to 12. The terminal score was the sum of all 16 values, with intermediate scores defined by the sum of all numbers to that position. Correlations were computed between sums at each of the 16 positions and the terminal scores. The results corresponded very closely to the predicted pattern \( r^2 = \% \text{ variance explained} \).

A similar pattern was predicted for cognitive scores, using the ratio of mental age-pre to mental age-post as the meter for percent of development which has occurred. Age ratios were plotted against interage correlations for 135 boys and 130 girls from the Harvard Growth Study (Dearborn, Rothney, & Shuttleworth, 1938) and an unspecified number of children from a study by Honzik (1938). In all cases the actual correlations fell at or below the level predicted. Anderson interpreted the results:

> Whatever question may be raised with reference to the accuracy with which the data obtained fits the formula, it is clear that the phenomenon of the increase and decrease of correlation coefficients as we move toward terminal status or away from initial status is one that is related to the per cent of overlap between the measures. (Anderson, 1939, p. 365)

The concept of overlap rests on the assumption of an additive, linear unfolding of intellectual maturity. Partly in response to this, many researchers have attempted to demonstrate qualitative stages in mental development.

**Predictive Validity of Infant Scores**

Researchers in human development caution that infant tests do not measure intelligence, at least not as the term is applied to adult
functioning. Honzik wrote:

Infant tests obviously do not measure what is measured by the Stanford-Binet, the Wechsler, or primary abilities tests; they measure abilities and skills that, to a large extent are the bases and precursors of later mental development. (1976, p. 91)

Anderson (1939) cautioned against the use of infant tests to make generalizations about the nature of intelligence. Infant behaviors, he suggested, reflect on something distinct from adult intelligence.

Bayley argued that infant tests cannot predict later intelligence, partly due to the qualitative differences in the phenomena observed at different ages. "The growth of intelligence... [appears] to be the maturing of a succession of partially overlapping functions which become increasingly complex as they approach adulthood" (Bayley, 1949, p. 166). Somewhat later she wrote, "I see no reason why we should continue to think of intelligence as an integrated (or simple) entity or capacity which grows throughout childhood by steady accretions" (Bayley, 1955, p. 807). Many other authors (Lewis, 1973; McCall et al., 1972; Ulvund, 1984) have also argued against equating infant test scores with later intelligence.

It is important to note the distinction between infant scores and intelligence, primarily because of the history of testing and the heavy emphasis on predictive validity. Bayley (1955) described some of the original reasons for testing infants. Some of them were to improve educational planning, to make better foster home placements, and to "put the feeble-minded into custodial care" (Bayley, 1955, p. 805). All of these intents are apparent in the operation of two orphanages, described by Skeels and Dye (1939). As Bayley (1955)
noted, the goals of early testing rested on the assumption of constancy of intelligence and the predictive powers of infant measures.

Brooks and Weinraub (1976), in a historical review of infant testing, wrote that assessment of infants gained momentum in the 1920s and 1930s and was followed by a series of studies on stability, reliability, and predictive validity. Prediction, initially the most valued criterion, resulted in much disillusionment with infant testing. "By the 1960s investigators were becoming reconciled to the fact that even improvements of existing tests would not lead to high predictive validity for normal children" (Brooks & Weinraub, 1976, p. 50).

The evaluation of infant tests by correlation studies has come under fire since the early 1960s. McCall observed that infant tests do not typically show highly stable individual differences and added:

Some have taken this conclusion to mean that infant tests are not useful assessments and have relatively little validity. Such a position testifies to the American psychologist's penchant for longitudinal prediction as a preferred research strategy and even as a criterion for evaluating the utility and validity of infant assessment techniques. (1979, p. 715)

Sameroff wrote, "Since qualitative shifts in development do occur during childhood, it might be thought to be surprising if children did show continuity in their intellectual performance" (1975, p. 279).

Aylward and Kenny echo the reasoning:

Although development is typically defined as growth or change over time, looking for predictability suggests stability. Development involves qualitative change, which, by definition, is discontinuous. (1979, p. 334)

Perhaps one reason that researchers concern themselves with predictive validity is some kind of cultural bent. Kagan (1980) wrote of a tendency of human development researchers to look for
connectedness in all changes. Rutter opened his review:

The study of predictions from infancy regarding later psychological development has received impetus from many different sources—not the least of which is the widespread curiosity concerning any attempt to forecast the future. (1970, p. 49)

McCall wrote, "Some things are sacred. For developmental psychology, predicting later behavior from early behavior is sacred, and no amount of evidence to the contrary will sway us from our appointed task" (1981, p. 41). Lewis (1973), in a similar vein, stated:

The concept of intelligence, the belief that it is relatively easily measurable, and that, as a monolithic construct, it is a useful predictor or subsequent human behavior, is firmly engrained in the mind of Western man. The consequence of this is to render a discussion of this construct difficult. (p. 108)

Finally, Ulvund asserted, "A requirement of predictive validity for infant tests may in many ways be held to be outdated" (1984, pp. 78-79).

Whatever the motives, many researchers have conducted studies regarding predictive validity of infant tests. For example, studies have been done which relate Bayley scores to Stanford-Binet IQs (Ireton, Thwing, & Gravem, 1970; McGowan, Johnson, & Maxwell, 1981; Ramey, Campbell, & Nicholson, 1973), with Wechsler IQs (Wilson, 1983; Roe, McClure, & Roe, 1983; Goffeney et al., 1971), and with numerous other measures (Berk, 1979; Ramey et al., 1973; Roe et al., 1983).

Meta-Analysis

The integration of research has recently received much attention (Carlberg & Wallberg, 1984; Pillemer, 1984). Inconsistencies in review results have led to a search for alternative ways of drawing
conclusions from existing research reports. One approach, meta-analysis, is based on the assumption that the procedures of primary research apply to reviews.

Meta-analysis (Glass, 1976) is a systematic, empirical review process. It has some characteristics similar to primary empirical research, but derives conclusions from data already published elsewhere. In both primary research and meta-analysis data are collected systematically across a group of subjects. In meta-analysis, however, the subjects are reports of research.

The essential task of the meta-analyst is to systematically analyze all (or a representative sample of) pertinent, available primary research on a given subject. To do this there must be some method of converting the reported results to one common metric. In traditional meta-analysis the effect size has served as the common metric. However, other measurements such as correlations or percents if applied consistently, may be used.

There are six stages in the research process. The first stage is to select and delimit the topic. This serves to clarify and define the problems that will be addressed. The topic addressed in this report is the lack of established stability and predictive validity of the Bayley Scales of Infant Development despite a large body of primary research.

Previous Reviews

The second stage of research is to review previous work. This has three purposes: to assure that new research will follow logically
from previous work, to identify an appropriate point of departure, and
to help the researcher learn from the mistakes and insights of those
who have gone before. In the present study, existing reviews were
used to identify primary research articles, to develop a coding sheet,
and to create coding conventions (described in METHODS). Location
procedures and criteria for inclusion will now be presented. Then the
evidence contained in previous reviews will be analyzed according to
the steps of sound research practice.

Previous reviews, per se, are of interest here only if they treat
the general theme of predictive validity and stability of infant
assessments. During the general literature search (described in
detail in the METHODS section of this paper), articles which included
at least a brief review the literature were separated for further
consideration. From those, any document regarding the predictive
validity or stability of infant measures was considered a review if
one or more of the following questions could be answered yes.

1: Is the word "review" in the title?

2: If original primary data are reported, do the authors cite
ten or more primary studies related to predictive validity/stability
of infant measures?

3: If the article does not report original data, are five or
more related primary articles cited?

4: Do the authors refer to it as a review in the text?

Two articles qualified as reviews using the second criterion
described above, 13 using the third. One article also met the fourth
criterion. (Review references are given in Appendix A). It should be
noted that not all articles which qualified as reviews were necessarily intended by the authors to be reviews.

Only seven of the identified reviews contained mention of previous reviews. That is, most authors completely failed to consider, in writing, previous work. Of those who did, the median number of cited reviews was one (range=1 to 6). McCall et al. (1972) were most frequently cited (n=3). Bayley (1970) was cited twice. Ten other publications were cited once as a review.

The third stage in conducting research is to select a sample. In a review this involves locating related literature and examining closely those articles which are relevant to the topic. If the conclusions of the review are to be generalizable, the sampled literature must be representative of or equal to the population of related research.

Subject selection procedures of previous reviews (i.e., how source materials were obtained) were coded according to reported location procedures and inclusion criteria. In no review were procedures for subject location given, although some did briefly describe criteria for inclusion. That is, no review reported sufficient information to allow replication of the literature search.

Within each review citations of original studies were counted. A reference was considered a primary study if, from the information in the review, the cited source appeared to be an original publication which contained primary research data regarding stability and/or predictive validity of any measure of infant development. The number of primary citations ranged from five (Bloom, 1964) to 27 (Brooks &
Weinraub, 1976), with a mean of 10.5 and a median of 8.5. Thus, some reviews were extensive while others were cursory.

The fourth stage in research is to systematically collect relevant data. In high-quality research reports the types of data collected for each subject and methods of data collection are given in sufficient detail to permit replication and accurate evaluation. In a review the process of examining research reports should be consistent, the variables studied should be stated, and detailed procedures must be given.

McCall (1979) examined the influence of pre- and post-test age on cross-age correlations. For a large body of studies, ages at testing and correlations were noted. The procedures were sufficiently well described to permit replication and accurate evaluation. In all other reviews, data collection was without explanation or procedural description. No data collection procedures, aside from McCall's review of age effects, could be replicated.

Data analysis is the fifth stage in research. Statistical methods of description and inference are invaluable in examining differences between groups or individuals. Similarly for reviews, statistical tools can be very useful in examining data. However, they must be stated explicitly so that the appropriateness of the methods can be assessed and to allow replication. Only McCall's (1979) analysis of age effect data was sufficiently detailed to permit replication. All other analyses in previous reviews were either subjective or not described.

The final stage of research is interpreting and reporting results. The interpretation of results in primary research must be
fairly rigorous. The author must be careful to draw conclusions that are well supported by the data. The report must be sufficiently detailed that readers can determine whether the interpretations are appropriate.

Numerous factors which might affect the strength of pre-post correlations were suggested in earlier reviews. Among them were age at pre-test, interval between tests, sex, socioeconomic status (SES), stability of the environment, test characteristics, developmental status, examiner competence, race, use in conjunction with clinical judgment or other assessment, and whether the scores are used on an individual or group level.

The factor of age was directly stated or implied in all reviews. Of the seven reviews which contained conclusions regarding the effect of age on correlations, all agreed that, other factors being equal, the older the subjects are at pre-test the higher will be the resulting coefficient (Anastasi, 1976; Bayley, 1949; Bloom, 1964; Dunst & Rheingrover, 1981; Honzik, 1976; McCall, 1979; Sattler, 1982). For other reviewers the effect of pre-test age seemed to be a foregone conclusion.

Interval between tests was an indicated factor in four reviews. The conclusion of all four was that the longer the interval between tests, the lower the correlation (Anastasi, 1976; Bayley, 1949; Dunst & Rheingrover, 1981). Honzik summarized the evidence on age at pre-test and interval length, "[The data] indicate what has become a truism in longitudinal studies of infants and children: the interage correlations are highly related to the age at testing and inversely related to the interval between tests" (1976, p. 68).
The influence of sex of subjects on interage correlations is not clear. Of five reviews which contained mention of sex as a variable, in only three were conclusions regarding that variable drawn. Honzik (1976), Bayley (1949), and McCall (1979) all concluded that girls' scores may be predictive slightly earlier than boys', although the influence of sex on correlations from infant measures is minimal at most.

Socioeconomic status has received some attention in reviews. Willerman, Broman, and Fiedler (1970) using original data, argued that SES interacts with developmental maturity. In their sample low scores were more predictive for low SES subjects than for high SES subjects. Honzik commented on their findings, "This is a provocative paper but there may be alternative interpretations" (1976, p. 75). McCall (1979) found that SES alone is a better predictor of later mental test scores than are infant assessment results. No reviewer concluded that strength of interage correlation is substantially influenced, in main effects fashion, by SES.

One factor mentioned by nearly all reviewers was developmental status. Scores reflecting handicap or developmental delay were seen as more predictive than normal or superior scores (Aylward & Kenny, 1979; Brooks & Weinraub, 1976; Dunst & Rheingrover, 1981; Honzik, 1976; Meier, 1976). McCall wrote:

Predictions to later test performance tend to be better for clinical groups and for samples involving children with mental deficiency than for "normal" infants (e.g., Knobloch & Pasamanick, 1960, 1963, 1967). Moreover, a low score on an infant test or a prognosis of risk made by a pediatrician at 20 months of age can predict low levels of mental test performance well into childhood. (1979, p. 712)
Anastasi (1976) suggested that infant tests are useful in the detection of organic pathology, of either environmental or hereditary origin, and that this accounts for the relatively high predictive value of subnormal scores. "In the absence of organic pathology, the child's subsequent development is determined largely by the environment in which he is reared. This the test cannot be expected to predict" (Anastasi, 1976, p. 334).

Rutter (1970) cautioned that even low scores are of limited use in predicting later individual performance:

It is of no use in differentiating within the above normal range of I.Q., but the developmental assessment is of value in detecting intellectual retardation.... Even so, it is important to realize that an appreciable minority of children will be completely misclassified. (p. 51)

It has been suggested that infant scores will be more predictive for children growing up in stable environments than for other subjects. Of seven reviews which mentioned the variable, only two drew conclusions regarding its effect on cross-age correlations. Both Honzik (1976) and Ulvund (1984) concluded, based on one study (Ramey, et al., 1973), that a relatively constant environment increases predictions from infancy scores.

In three reviews the distinction between group and individual prediction was made. McCall (1979) and Bloom (1964) concluded that predictions cannot be made from infant tests for individuals within normal populations. Aylward and Kenny argued, "Even in...clinical studies, predictability relates to the group and there is no way of knowing how the individual neonate or infant will fare" (1979, p. 333). Bhakoo, Kaur, Narang, and Verma (1977), in a primary study,
found that high cross-age correlations and relative invariance of mean scores were not incompatible with large changes in developmental quotients of individuals.

The fact that means remain relatively stable, does not necessarily imply that there will be no individual variations. The means are, like bikinis, sometimes known to hide more than they reveal. (Bhakoo et al., 1977, p. 62)

Other variables which might influence inter-age correlations are race (Sattler, 1982), examiner competence (Honzik, 1976), instrument characteristics (Bayley, 1949; Bloom, 1964; Brooks & Weinraub, 1976; Honzik, 1976; McCall, 1979; Sattler, 1982; Thomas, 1967; Ulvund, 1984), statistic type (Bloom, 1964), developmental irreversibility (Anastasi, 1976; Sattler, 1982), and combinations of assessment techniques, including clinical judgment (Anastasi, 1976; Aylward & Kenny, 1979).

The general conclusions reached by previous reviewers were generally consistent. Several reviewers concluded that infant tests are not very predictive, at least for normal subjects (Aylward & Kenny, 1979; Bayley, 1949; Bloom, 1964; Brooks & Weinraub, 1976; Dunst & Rheingrover, 1981; Honzik, 1976; Rutter, 1970; Sameroff, 1975; Sattler, 1982; Ulvund, 1984; Willerman et al., 1970). Reviewers also concluded that predictions may become more useful after 18 months of age (Anastasi, 1976; Honzik, 1976; McCall, 1979). Only Thomas (1967), in his review of methodological flaws in the research, concluded that the data are insufficient to draw any definitive statements.

Twelve reviews contained at least a brief discussion of problems in primary research. Among the difficulties were confusion of the
terms stability and developmental continuity (Dunst & Rheingrover, 1981; McCall, 1979), confusing correlations with cause/effect relationships (McCall, 1979), failure to report base rates when giving percent of correct diagnoses (Rutter, 1970; Thomas, 1967), and grouping data without apparent prior justification (Thomas, 1967).

Thomas (1967) argued that sample selection may be biased by exclusion of subjects who score below norms or basal levels, by using only cases which, by apparently subjective judgments, were "validly tested," and subject solicitation procedures. He noted that Bayley's (1949) longitudinal sample at no time had mean Binet IQs below 120--clearly not representative of the population. Failure to systematically treat SES (Willerman et al., 1970) and changes in the environment (Sameroff, 1975) and overdependence on test scores to the exclusion of other evidence (Sattler, 1982) were also mentioned.

The elements of high quality research have been described. Following specified procedures, a body of previous reviews of the stability and predictive validity of infant measures has been identified, examined, and found to have serious methodological flaws. Informal collection of data and discrepancies between the reviews in regard to specific conclusions are particular weaknesses. Factors which are considered important in the field of infant testing have also been specified.
METHODS

For this report the studies of interest are published primary research pertaining to the stability or predictive validity of the 1969 revision of Bayley Scales of Infant Development. Stability and predictive validity, as defined for this paper, refer to the correlation between two sets of scores for a group of individuals in which there is an interval between tests of at least one month. Since the Bayley scales have norms for ages 2 to 30 months, only studies with pretests administered in that age range were considered. Articles sought for this paper were those with cross-age correlations, with an interval of at least one month, using the 1969 Bayley as pretest.

Target articles were identified through several sources. All references in Mental Measurement Yearbook (Buros, 1978) for the Bayley were inspected to determine if sufficient information was available to permit coding. Psychological abstracts were reviewed, using the descriptors of Infant Development and Infants, for the years 1969 to the present. A computer library search was also conducted, using descriptors of age, test, and statistics. Relevant chapters from the texts Psychological Testing (Anastasi, 1976) and Assessment of Children's Intelligence and Special Abilities (Sattler, 1982), were reviewed. Finally, the bibliographies of all obtained articles published after 1968 were examined to determine if other appropriate articles are referenced.

To determine if a cited article was codable, many sources of information were considered. First, references often contained
descriptions of the sample, tests used, and type of analyses. The context in which an article was cited also provided information regarding its appropriateness for the meta-analysis. Abstracts were available for many references and were examined. An attempt was made to obtain all articles which appeared relevant to the stability/predictive validity of the Bayley. If the information from these sources did not allow a confident determination of whether an article was codable, a full effort was made to obtain it.

Resources for accessing source documents were the Merrill Library at Utah State University, the Early Intervention Research Institute Library, and Interlibrary Loan service. All articles thought to be codable were sought through these resources. At least 202 abstracts and 76 full articles were examined. From these, 15 reviews were identified and 18 primary studies (representing 23 separate articles) were obtained. The references for the primary articles are given in Appendix B.

Note that the manual for the 1969 Bayley is not included in the primary articles listing. This is because the only cross-age correlations reported in the manual involved the 1958-60 version of the Bayley.

Coding Sheet Development

A preliminary coding sheet and tentative conventions for its use were developed. Items were derived from factors identified in the review of reviews. Many of the conventions were taken from a meta-analysis of early intervention research and are of demonstrated effectiveness in yielding highly reliable coding.
The preliminary coding sheet and conventions were put to a test by coding five primary articles. Necessary changes were made, followed by a recoding of the original five articles on the revised coding sheet. Final changes were made at that point. Then all articles which had been deemed appropriate for the proposed meta-analysis were coded.

There are four categories of coding sheet items. The first category contains three items for identification of the coded correlation. The next 14 items pertain to sample characteristics. Items 4 and 5 are sample size at pre- and post-test, respectively. These two items, taken together, indicate the attrition rate between test administrations. Item 6 is the mean Bayley quotient at pre-test. Standard deviation, highest score, and lowest score at pre-test are items 7, 8, and 9. Post-test mean and standard deviation are items 10 and 11. Data from these items permit analyses of developmental status and sample variability. The next four items are the percentage of the coded sample which was white, black, Hispanic, and male. The final item of this section is sample SES.

The third category of items is Intervention. These items were included to assess the effects of preschool treatment and consistency of environment influences. Whether intervention occurred and, if so, the duration and hours per week were coded.

The last section is Measurement. Here the actual correlation, ages at testing, tests administered, examiner characteristics, source of correlation, and statistic type were coded. The total coding sheet contains 33 items. (A sample coding sheet is found in Appendix C.)
Conventions for Coding

All items involving mean values were coded with one of four measures of central tendency, the order of preference being mean, median, mode, and midrange. When exact data were not available, reasonable guesses were allowed. It was difficult to establish conventions governing guesses. However, a guess was considered "reasonable" if the coder felt 90% confident that it was within 10% of the true value or, in the case of categorical data, if the coder felt 90% confident that the given code matched the "true" value.

In cases for which the data were not provided, could not be reasonably estimated, and for which there were no specific conventions covering the situation, the item was coded "-9".

Ages were coded in whole month units, according to literal interpretation of data reports. If an author reported giving the Stanford-Binet to subjects when they were three years old, the mean age at testing would be coded 36 months, even though the description might refer to subjects 36 to 47 months of age. This approach was chosen because most tests in the meta-analysis were given to very young subjects (two months to three years) and most authors appeared to distinguish between subjects at, say, 1 year and subjects at 18 months.

A problem unique to meta-analysis is dealing with variable levels of reported information. The following hypothetical scenario will help illustrate the point. Smith conducts a correlational study with 100 infants. The Bayley MDI is given to the infants at age 18 months. At three years of age all infants are retested on the Stanford-Binet.
One correlation is reported. Jones conducts a similar study, with the Bayley MDI administered to 100 infants at age 18 months. At post-test the children are given the Stanford-Binet, the Slosson Intelligence Test, the Vineland Social Maturity Scale, and the Peabody Picture Vocabulary Test. Four correlations are reported. An uncritical coding of all reported correlations leads to a meta-analysis in which the Jones study is weighted four times as heavily as the study by Smith.

One method to control for this difficulty is to require independence of all observations, allowing only one correlation to be coded for each independent sample. Unfortunately this procedure drastically reduces the amount of information available to the meta-analysis. A compromise procedure is to require independence within designated cells. The variables which are likely to be most powerful can be used to limit coding. The result is that all studies which report data pertinent to a given cell receive equal weighting within that cell. Additionally, the influence of the limiting variables can be reduced within a given cell, while allowing for assessment of those variables by between cell comparisons.

In order to provide some independence of observations, a system of data collection, based on a matrix of age at testing, was used. Pre-test ages were grouped into four categories, as follows: 0-6 months, 7-12 months, 13-18 months, and 19-30 months. Post-test ages were grouped more broadly, into three categories: 0-24, 25-72, and 73 or more months. A matrix of 12 cells is formed by combining the pre-test columns with post-test rows.
Independent samples were defined as those which had no individual members in common. Only one correlation per cell was coded for an independent sample. For example, if an author reported three correlations between scores at ages 6, 8, and 12 months, only one correlation would be coded. If the data were reported separately for white males, white females, black males, and black females, then four correlations would be coded. This procedure insured that all coded correlations in a given cell were independent. Furthermore, the influence of age and interval length was reduced within cells, permitting more powerful analyses of other variables.

When a large group of subjects could be subdivided into smaller groups, such as males and females, blacks and whites, high and low SES, they were coded as subgroups if the resultant sample size were larger than 25 and no significant data were lost.

When multiple breakdowns were possible, but the resultant correlations would not be independent, breakdown priorities were, first sex, then race, and finally SES. When multiple outcome measures were reported, only the correlation most closely associated with the pre-test (MDI or PDI, whichever was being coded), most well standardized, yielding deviation (not ratio) scores, and which had the largest sample size was coded. If more than one correlation was still possible for a cell (for example, when tests were given at 8, 12, 18, and 24 months) the correlation with pre- and post-test ages closest to the midrange for that cell was coded.

To illustrate these conventions, consider a hypothetical study in which the Bayley MDI and Cattell Infant Intelligence Test were given
at 8, 12, 18, and 24 months. Only two correlations would be coded: the correlation between MDI scores at 8 and 12 months, and that for ages 18 to 24 months. All correlations involving the Cattell would be disregarded, as would the correlations between MDIs at 8 and 18, 8 and 24, and 12 and 24 months. Conventions for each item are given in Appendix D.

Coding Reliability

Coding reliability was assessed by interrater agreement. A psychology graduate student who had participated in a previous meta-analysis was trained by the primary coder (the author) in the use of the coding sheet and received and read a copy of the conventions. The two codable articles with the lowest ID numbers (Siegel, 1981; McGowan et al., 1981) were selected and independently coded by the author and the trained student.

Both coders completed coding on six correlations for each article. There was perfect agreement on all breakdowns by sex, tests, and age.

For 6 of the 12 correlations there was a discrepancy on the sample size at post-test. One coder estimated the samples at 100 and the other coded 103. The margin of error is 3%. The next disagreement was on item 17, SES. For 6 of 12 correlations there was disagreement between codes of "low" and "mixed" SES of sample. Following a discussion of the item, complete agreement was reached, with the final code being used in the data analysis.

There were two discrepant codes on "age change", which is the interval between tests in whole month units. The two codes were 22
and 23 months. The margin of error is 5%, within the acceptable range.

There were six disagreements for "pre-test examiner" and six for "post-test examiner." The different codes were "paraprofessional" and "other professional." Complete agreement was reached after a discussion. The agreed upon codes were used in the data analysis. No other disagreements occurred. Overall agreement was 93%.

**Data Analysis**

Because an attempt was made to code all documents relevant to the stability and predictive validity of the Bayley, the identified sample was equal to the identified population. As a result, no inferential statistics were used, as no inference to a sampled population was needed.

Results of the study will be presented in tables of mean correlation value for all coding sheet variables, with standard deviations and sample sizes specified. A table of mean correlations will also be presented for all age-by-interval cells.

In all analyses raw correlations were used rather than Fisher's Z transformation because most values were in the .15 to .85 range. Furthermore, the purpose of Fisher's Z transformation is to adjust correlation values into a normal curve, making possible traditional inferential methods. With no inferential statistical analyses, the advantages of Z transformations did not apply to the present study.

The original study was designed anticipating sufficient data in the twelve age-pre/age-post cells to permit separate analyses for
each. Unfortunately, the cells only contained from 1 to 19 correlations. When PDI and MDI scores are analyzed separately, some cells shrink to zero, while the largest is 13, not enough to permit regression analyses.

An alternative analysis was conducted. Data from all cells were pooled and multiple regressions, with age-pre and interval length as predictors and the coded correlation as the dependent variable, were conducted for MDI and PDI data. All coded correlations (hereafter CCs) were then converted to difference values (hereafter DVs) by subtracting the predicted coefficient (computed from coded values and the corresponding regression equation) from the actual correlation. All coding sheet variables were then reanalyzed using the adjusted correlation values.

This procedure introduces a threat to validity, since combining the data across cell boundaries results in nonindependent observations. However, according to Tracz and Elmore (1985), violation of independence when combining correlation coefficients in a meta-analysis does not affect the estimation of either the central tendency or the standard deviations for correlation coefficients. The analyses conducted here involve only means and standard deviations. Combining data from all cells, then is justified.
RESULTS

There were 75 MDI correlations coded. The sample sizes, means, and standard deviations of correlations are given in Table 1 for the twelve age-pre by age-post cells. Of the 75 MDI correlations, 72 were product moment, 2 were Spearman ranks, and one was a point-biserial coefficient. One point-biserial coefficient and 26 product moment correlations were coded for the PDI. One PDI and one MDI correlation were computed from raw data in a publication. All other coded correlations were values reported in a document.

A multiple regression equation, with age-pre and interval length as independent variables, accounted for 36% of the variance in coded MDI correlations. The solved multiple regression equation, with $Y$ being the predicted correlational value, $X1$ the age at pre-test (in months), and $X2$ the interval length (in months), was:

$$Y = 0.297 + 0.018xX1 - 0.0028xX2$$

The mean of all coded MDI correlations was .463, with a standard deviation of .246. The mean age at pre-test was 13 months (sd=6.18) and mean interval length was 23.7 months (sd=25.6).

There were 27 PDI correlations coded. The means, standard deviations, and sample sizes for the twelve cells of the matrix are given in Table 2. A multiple regression equation, with age-pre and interval ad independent variables, accounted for 26% of the variance in coded PDI correlations. The solved PDI multiple regression, with $Y$ as the predicted correlation, $X1$ the pre-test age, and $X2$ the
Table 1

MDI Cell Values

<table>
<thead>
<tr>
<th>Cell</th>
<th>Age-Pre</th>
<th>Age-Post</th>
<th>(n^a)</th>
<th>Mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2-6</td>
<td>0-24</td>
<td>5</td>
<td>.33</td>
<td>.28</td>
</tr>
<tr>
<td>2</td>
<td>7-12</td>
<td>0-24</td>
<td>13</td>
<td>.53</td>
<td>.21</td>
</tr>
<tr>
<td>3</td>
<td>13-18</td>
<td>0-24</td>
<td>10</td>
<td>.52</td>
<td>.25</td>
</tr>
<tr>
<td>4</td>
<td>19-30</td>
<td>0-24</td>
<td>3</td>
<td>.59</td>
<td>.02</td>
</tr>
<tr>
<td>5</td>
<td>2-6</td>
<td>25-72</td>
<td>3</td>
<td>.28</td>
<td>.22</td>
</tr>
<tr>
<td>6</td>
<td>7-12</td>
<td>25-72</td>
<td>12</td>
<td>.32</td>
<td>.28</td>
</tr>
<tr>
<td>7</td>
<td>13-18</td>
<td>25-72</td>
<td>11</td>
<td>.55</td>
<td>.18</td>
</tr>
<tr>
<td>8</td>
<td>19-30</td>
<td>25-72</td>
<td>9</td>
<td>.67</td>
<td>.14</td>
</tr>
<tr>
<td>9</td>
<td>0-6</td>
<td>73+</td>
<td>1</td>
<td>.30</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>7-12</td>
<td>73+</td>
<td>6</td>
<td>.19</td>
<td>.10</td>
</tr>
<tr>
<td>11</td>
<td>13-18</td>
<td>73+</td>
<td>1</td>
<td>.48</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>19-30</td>
<td>73+</td>
<td>1</td>
<td>.56</td>
<td>-</td>
</tr>
</tbody>
</table>

\(^a\)Number of coded correlations for the cell.
interval was:

\[ Y = 0.541 + 0.0046X_1 - 0.0064X_2 \]

The mean of all coded PDI correlations was \(0.472\) (\(sd = 0.238\)). The mean age at pre-test was \(11.4\) months (\(sd = 5.93\)) and mean interval length was \(18.9\) months (\(sd = 17.4\)).

A simplex-like pattern is revealed for the MDI, as correlations increase with age at pre-test and decrease with interval length. In fact, predictive validity/stability coefficients correlate \(0.53\) with age at pre-test and \(-0.41\) with interval length, giving further evidence of a simplex-type of pattern. A determination of whether the PDI correlations form a simplex pattern cannot be made, primarily due to lack of data. Coded PDI coefficients do, however, correlate \(0.24\) with pre-test age and \(-0.50\) with interval length.

The data for the dependent variable in this study are coded stability/predictive validity correlations. Other items on the coding sheet constitute the independent variables. The analyses were conducted by comparing mean correlations across levels of each independent variable. All independent data continuous in nature were converted to ordinal scales by grouping. Divisions for grouping were determined by logical breaks, convenience, or divisions which occurred in the collected data.

Each item for which there are sufficient data was examined with both coded correlations (CC) and with difference values (DV). As described in the Methods section, coded correlations were subtracted from values predicted by the regression equation, either of the PDI or MDI, to give the DV. The reader is advised to remember that DVs refer to the difference between actual and predicted coefficients. A DV of
Table 2

PDI Cell Values

<table>
<thead>
<tr>
<th>Cell</th>
<th>Age-Pre</th>
<th>Age-Post</th>
<th>( n^a )</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>1</td>
<td>2-6</td>
<td>0-24</td>
<td>3</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>7-12</td>
<td>0-24</td>
<td>6</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>13-18</td>
<td>0-24</td>
<td>4</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>19-30</td>
<td>0-24</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>2-6</td>
<td>25-72</td>
<td>2</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>7-12</td>
<td>25-72</td>
<td>6</td>
<td>0.32</td>
</tr>
<tr>
<td>7</td>
<td>13-18</td>
<td>25-72</td>
<td>3</td>
<td>0.52</td>
</tr>
<tr>
<td>8</td>
<td>19-30</td>
<td>25-72</td>
<td>2</td>
<td>0.51</td>
</tr>
<tr>
<td>9</td>
<td>0-6</td>
<td>73+</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>7-12</td>
<td>73+</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>11</td>
<td>13-18</td>
<td>73+</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>19-30</td>
<td>73+</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

\( ^a \) Number of coded correlations for the cell.
zero means that the obtained value was equal to the predicted value, given age at pre-test and interval length. A DV of zero does not signify a pre-post correlation of zero.

The first items of the coding sheet that were analyzed were sample sizes at pre- and post-test. When the sample sizes were only reported for the post-test, that number was coded for both post- and pre-test. Data were divided into three groups, according to sample size: 50 or fewer subjects, 51 to 100, and 101 and greater. Larger groups did not show higher correlations, as might be expected. (See Tables 3 and 4.) There was no consistent effect of sample size for either the MDI or PDI.

The attrition of subjects, in percent of original sample size, was calculated. Samples were placed into three groups, based on percent attrition: 0-25%, 26-50%, and 51% or greater. For the MDI, the group with low attrition had higher DV means than the medium and high attrition groups. This suggests that a selection bias may be working in some reported studies. For mean CCs, however, a difference is apparent only between medium and low attrition groups, again with low groups obtaining higher scores. (See Table 5.)

Only two PDI samples had attrition less than 26%. High attrition groups had lower CC and DV means than the medium attrition groups, suggesting that some type of selection bias may be operating in published reports. (See Table 6.)

The developmental status of each sample was coded, in part, by the mean pre-test scores. There were eight MDI correlations for which pre-test means were not available. Samples were divided into three groups: those with mean pre-test scores below 85, 85 through 100, and
Table 3

Sample Size and MDI Correlations

<table>
<thead>
<tr>
<th>Subjects</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>(n)(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-50</td>
<td>.53</td>
<td>.24</td>
<td>.00</td>
<td>.23</td>
<td>33</td>
</tr>
<tr>
<td>51-100</td>
<td>.51</td>
<td>.25</td>
<td>.03</td>
<td>.23</td>
<td>17</td>
</tr>
<tr>
<td>101+</td>
<td>.34</td>
<td>.20</td>
<td>-.02</td>
<td>.12</td>
<td>25</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.

Table 4

Sample Size and PDQ Correlations

<table>
<thead>
<tr>
<th>Subjects</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>(n)(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-50</td>
<td>.43</td>
<td>.26</td>
<td>-.07</td>
<td>.23</td>
<td>7</td>
</tr>
<tr>
<td>51-100</td>
<td>.61</td>
<td>.20</td>
<td>.11</td>
<td>.16</td>
<td>12</td>
</tr>
<tr>
<td>101+</td>
<td>.31</td>
<td>.14</td>
<td>-.10</td>
<td>.18</td>
<td>8</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.
### Table 5

**Subject Attrition and MDI Correlations**

<table>
<thead>
<tr>
<th>Attrition Percent</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-25%</td>
<td>.42</td>
<td>.24</td>
<td>.07</td>
<td>.19</td>
<td>10</td>
</tr>
<tr>
<td>26-50%</td>
<td>.55</td>
<td>.21</td>
<td>.03</td>
<td>.17</td>
<td>36</td>
</tr>
<tr>
<td>50+%</td>
<td>.38</td>
<td>.27</td>
<td>-.06</td>
<td>.22</td>
<td>29</td>
</tr>
</tbody>
</table>

\(^a\) CC = Mean correlation.  
\(^b\) DV = Mean difference between obtained and expected r.  
\(^c\) Number of coded correlations.

### Table 6

**Subject Attrition and PDI Correlations**

<table>
<thead>
<tr>
<th>Attrition Percent</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-25%</td>
<td>.50</td>
<td>-</td>
<td>.04</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>26-50%</td>
<td>.53</td>
<td>.22</td>
<td>.02</td>
<td>.20</td>
<td>18</td>
</tr>
<tr>
<td>50+%</td>
<td>.33</td>
<td>.28</td>
<td>-.05</td>
<td>.24</td>
<td>7</td>
</tr>
</tbody>
</table>

\(^a\) CC = Mean correlation.  
\(^b\) DV = Mean difference between obtained and expected r.  
\(^c\) Number of coded correlations.
above 100. (See Table 7.) An effect of DQ on cross-age correlation is not apparent from CC values. Groups with MDI pre-test means above 100 appear to be slightly more predictable/stable than groups with lower DQs. However, by adjusting correlations for age at pre-test and interval length, the apparent effect of MDI on correlation reverses. Groups with lower scores demonstrate higher cross-age correlations than would be expected, given age at pre-test and interval length.

A similar analysis was conducted for PDI scores. Eight PDI means were not coded, leaving only 19 values for the analysis. Only two coded samples had mean below 85, so no determination of the effect of low scores could be made. The samples were redivided into two groups: those with pre-test means at or below 100 and those above 100. Mean CC values are identical for the two divisions (.40 for samples at or below 100; .41 for samples above). Likewise, no clear effect of pre-test PDI on cross-age correlation is evident from DVs. Low PDI scores do not appear more stable/predictive than high scores. (See Table 8).

Other factors being equal, highly variable samples often yield greater correlations than more homogeneous samples. Four separate coding sheet items deal with variance of the coded sample: standard deviation at pre-test, highest score at pre-test, lowest pre-test score, and standard deviation at post-test. For PDI correlations, data for these items were rarely reported. Standard deviations were available for only nine PDI samples, and highest and lowest scores could only be coded for four. The collected data permit no conclusions regarding PDI variance and cross-age correlations. (See Table 10.)
Table 7
Mean Pre-Test Scores and MDI Correlations

<table>
<thead>
<tr>
<th>Score</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-84</td>
<td>.42</td>
<td>.38</td>
<td>.08</td>
<td>.21</td>
<td>6</td>
</tr>
<tr>
<td>85-100</td>
<td>.42</td>
<td>.25</td>
<td>-.02</td>
<td>.20</td>
<td>33</td>
</tr>
<tr>
<td>101+</td>
<td>.49</td>
<td>.21</td>
<td>-.02</td>
<td>.20</td>
<td>28</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.

Table 8
Mean Pre-Test Scores and PDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100</td>
<td>.40</td>
<td>.27</td>
<td>-.06</td>
<td>.22</td>
<td>13</td>
</tr>
<tr>
<td>101+</td>
<td>.41</td>
<td>.11</td>
<td>-.11</td>
<td>.16</td>
<td>6</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.
The standard deviation of the MDI pre-test scores does not have any clear effect on cross-age correlations. The results are presented in Table 9. Highest and lowest pre-test scores were reported or available in only 5 out of 75 cases, disallowing any conclusions regarding these variables.

The relationship between mean post-test scores and inter-age correlations was examined by grouping outcome scores into three groups: post-test means of 0-67, 68-100, and 101 or greater. The results do not indicate any distinct relationship for either MDI or PDI. (See Tables 11 and 12.)

Data regarding presence and type of handicap were available for all MDI and PDI correlations. For the MDI, samples described as mentally retarded/developmentally delayed had higher cross-age correlations than predicted from pre-test age and interval length. Mean DVs were -.03 for non-handicapped populations (n=38), .00 for high risk samples (pre-mature or low birth weight infants; n=10), and -.02 for 18 samples of disadvantaged subjects. But for subjects described as mentally retarded/developmentally delayed the mean DV was .16 (n=8). Although low mean MDI scores only weakly correspond with higher correlations, diagnoses of retardation or developmental delay do coincide with correlations that are markedly higher than expected, given age and interval between test. A summary of this variable on the MDI is given in Table 13. Mean CCs and DVs involving PDI as pre-test were also higher for samples described as mentally retarded or developmentally delayed. Data regarding this item are presented in Table 14.
Table 9
Pre-test Standard Deviation and MDI Correlations

<table>
<thead>
<tr>
<th>SD Range</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-12</td>
<td>.41</td>
<td>.24</td>
<td>-.02</td>
<td>.19</td>
<td>9</td>
</tr>
<tr>
<td>13-18</td>
<td>.38</td>
<td>.24</td>
<td>-.02</td>
<td>.20</td>
<td>24</td>
</tr>
<tr>
<td>19+</td>
<td>.41</td>
<td>.32</td>
<td>-.06</td>
<td>.29</td>
<td>8</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  
<sup>b</sup>DV=Mean difference between obtained and expected r.  
<sup>c</sup>Number of coded correlations.

Table 10
Pre-test Standard Deviation and PDI Correlations

<table>
<thead>
<tr>
<th>SD Range</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-12</td>
<td>.21</td>
<td>.23</td>
<td>-.18</td>
<td>.11</td>
<td>3</td>
</tr>
<tr>
<td>13-18</td>
<td>.38</td>
<td>.13</td>
<td>-.09</td>
<td>.18</td>
<td>5</td>
</tr>
<tr>
<td>19+</td>
<td>.88</td>
<td>-</td>
<td>.32</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  
<sup>b</sup>DV=Mean difference between obtained and expected r.  
<sup>c</sup>Number of coded correlations.
Table 11
Mean Post-test Scores and MDI Correlations

<table>
<thead>
<tr>
<th>Score</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-67</td>
<td>.90</td>
<td>-</td>
<td>.31</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>68-99</td>
<td>.48</td>
<td>.22</td>
<td>-.01</td>
<td>.15</td>
<td>30</td>
</tr>
<tr>
<td>100+</td>
<td>.43</td>
<td>.25</td>
<td>-.02</td>
<td>.22</td>
<td>31</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.

Table 12
Mean Post-test Scores and PDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-67</td>
<td>.88</td>
<td>-</td>
<td>.32</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>68-99</td>
<td>.43</td>
<td>.23</td>
<td>-.07</td>
<td>.21</td>
<td>9</td>
</tr>
<tr>
<td>100+</td>
<td>.33</td>
<td>.17</td>
<td>-.13</td>
<td>.14</td>
<td>9</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.
Table 13

Handicap and MDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>.38</td>
<td>.24</td>
<td>-.03</td>
<td>.18</td>
<td>38</td>
</tr>
<tr>
<td>MR/DD</td>
<td>.68</td>
<td>.18</td>
<td>.16</td>
<td>.09</td>
<td>8</td>
</tr>
<tr>
<td>Orthopedic Disorders</td>
<td>.82</td>
<td>-</td>
<td>.32</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>High Risk</td>
<td>.47</td>
<td>.31</td>
<td>.00</td>
<td>.28</td>
<td>10</td>
</tr>
<tr>
<td>Disadvantaged</td>
<td>.53</td>
<td>.17</td>
<td>-.02</td>
<td>.17</td>
<td>18</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected r. \(^c\)Number of coded correlations.

Table 14

Handicap and PDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>.33</td>
<td>.18</td>
<td>-.10</td>
<td>.18</td>
<td>11</td>
</tr>
<tr>
<td>MR/DD</td>
<td>.68</td>
<td>.11</td>
<td>.18</td>
<td>.07</td>
<td>7</td>
</tr>
<tr>
<td>Orthopedic Disorder</td>
<td>.88</td>
<td>-</td>
<td>.32</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>High Risk</td>
<td>.44</td>
<td>.24</td>
<td>-.06</td>
<td>.21</td>
<td>7</td>
</tr>
<tr>
<td>Disadvantaged</td>
<td>.38</td>
<td>-</td>
<td>-.12</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV-Mean difference between obtained and expected r. \(^c\)Number of coded correlations.
To assess whether the Bayley was differentially predictive/stable for races, three items regarding race were coded: percent white, percent black, and percent Hispanic. Each item was coded according to whether 0-33, 34-67, or 68-100 percent of the subjects in the sample were of the particular race. MDI CCs are higher for samples with approximately equal numbers of black and white subjects. However, mean DVs do not differ according to racial composition of the samples (.01 when mostly white; -.04 when mostly black; -.07 when mostly Hispanics). A complication in this analysis is that 26 of 75 correlations could not be coded for race. (See Table 15.)

Only 20 PDI correlations were coded for race. No sample was predominantly Hispanic. The mean DV of mostly white samples was .05, while for mostly black samples it was -.01. The difference is not meaningful. There were only two PDI samples of approximately equal white and black subjects composition. No effect of race on PDI is apparent. (See Table 16.)

The percent of males in a sample was coded in item 16. The resulting data were divided into three groups: 0-33% male, 34-67% male, and 68-100%. This item was coded in all but one PDI sample. However, 23 of 26 samples for which the data were available were approximately equally divided for sex, with males comprising 33-67% of the total. Because few samples were mostly male or mostly female, no conclusions can be made about whether males, females, or neither have more predictive/stable PDI scores. (See Table 18.)

The percent of males in the sample was coded in 74 of 75 MDI correlations. There were eight MDI samples with 0-33% male, 56 with
### Table 15
Race and MDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC</th>
<th>sd</th>
<th>DV</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>.44</td>
<td>.33</td>
<td>-.01</td>
<td>.28</td>
<td>15</td>
</tr>
<tr>
<td>Black</td>
<td>.28</td>
<td>.26</td>
<td>-.05</td>
<td>.13</td>
<td>5</td>
</tr>
<tr>
<td>Hispanic</td>
<td>.50</td>
<td>.09</td>
<td>-.07</td>
<td>.11</td>
<td>6</td>
</tr>
</tbody>
</table>

Note: CC=Mean correlation, DV=Mean difference between obtained and expected r, n=Number of coded correlations.

### Table 16
Race and PDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC</th>
<th>sd</th>
<th>DV</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>.38</td>
<td>.12</td>
<td>-.05</td>
<td>.18</td>
<td>7</td>
</tr>
<tr>
<td>Black</td>
<td>.31</td>
<td>.07</td>
<td>-.01</td>
<td>.21</td>
<td>3</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: CC=Mean correlation, DV=Mean difference between obtained and expected r, n=Number of coded correlations.
34-67% males, and 10 with 68-100% male subjects. The DVs of those groups, respectively, are -.07, .02, and -.09. This suggests that samples that are mostly male or mostly females have lower cross-age correlations than expected, given age at pre-test and interval length. This same pattern is visible in CC means. (See Table 17.) It appears, then, that Bayley scores of groups which are mostly male or mostly female may be less stable/predictive than scores of mixed samples.

An attempt was made to compare white male, black male, white female, and black female samples. Unfortunately there were only three samples which could be placed in one of these four categories. It cannot be determined from the data collected whether race and sex interact in influencing predictive validity/stability of Bayley scores.

The SES of samples was coded as either Middle, Low, or Mixed, depending on data or author description. No articles were found which reported Bayley correlations for high SES samples. This item was coded in all PDI correlations and all but one MDI correlation. For the MDI, the middle class subjects had higher cross-age correlations than lower SES and mixed samples, as indicated by both CC and DV means. (See Table 19.) Middle class subjects had higher PDI correlations than mixed samples. There were only three PDI Low SES groups coded. (See Table 20.)

Whether a sample received early intervention was coded. By convention, if there was no indication of intervention, it was assumed that there was none. Consequently, this item was always coded. MDI samples which received intervention had slightly higher CC means than
### Table 17
Gender and MDI Correlations

<table>
<thead>
<tr>
<th>% Male</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-33</td>
<td>.33</td>
<td>.22</td>
<td>-.07</td>
<td>.17</td>
<td>8</td>
</tr>
<tr>
<td>34-67</td>
<td>.49</td>
<td>.22</td>
<td>.02</td>
<td>.19</td>
<td>56</td>
</tr>
<tr>
<td>68-100</td>
<td>.36</td>
<td>.32</td>
<td>-.09</td>
<td>.20</td>
<td>10</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.

### Table 18
Gender and PDI Correlations

<table>
<thead>
<tr>
<th>% Male</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-33</td>
<td>.12</td>
<td>-</td>
<td>-.20</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>34-67</td>
<td>.49</td>
<td>.21</td>
<td>.02</td>
<td>.19</td>
<td>23</td>
</tr>
<tr>
<td>68-100</td>
<td>.21</td>
<td>-</td>
<td>-.25</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.
other samples, but there was no difference on DVs. For the PDI there was no difference between groups on either CC or DV means. Intervention per se does not appear to coincide with high or low cross-age MDI correlations. (See Tables 21 and 22.)

For samples which received intervention, the duration was coded. Intervened samples were divided into three groups: those with 1-5 months of intervention, 6-17 months, and 18 or more months. The group with 1-5 months of intervention had the lowest DV and CC values. The samples with 6-17 months were next, and subjects which received 18 or more months of intervention had the highest DV and CC means. There were only 4 brief and 4 long intervention groups. However, for the MDI the length of intervention appears to be positively related to strength of cross-age correlations. (See Table 23.) Data from the PDI were divided into two groups: 1 to 11 months of intervention and 1 year or more. Because of small sample size, no conclusions can be made regarding duration of intervention and PDI correlations. (See Table 24.)

When intervention did occur, the intensity of intervention was coded, defined by the number of hours per week of treatment. There were very few PDI data. (See Table 26.) No conclusions are justified. For the MDI there were 15 cases of coded intervention intensity. These were divided into two groups: those with 1 to 9 hours per week, and those with 10 or more. There is no difference between the groups for cross-age correlations, as indicated by either DV or CC means. (See Table 25.)

The pre-test examiner was coded as either psychologist, paraprofessional, or other. No reports of correlations were found in
Table 19
Socio-Economic Status and MDI Correlations

<table>
<thead>
<tr>
<th>SES</th>
<th>CC</th>
<th>sd</th>
<th>DV</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle</td>
<td>.67</td>
<td>.14</td>
<td>.19</td>
<td>.13</td>
<td>9</td>
</tr>
<tr>
<td>Low</td>
<td>.47</td>
<td>.20</td>
<td>-.02</td>
<td>.16</td>
<td>30</td>
</tr>
<tr>
<td>Mixed</td>
<td>.39</td>
<td>.26</td>
<td>-.04</td>
<td>.21</td>
<td>35</td>
</tr>
</tbody>
</table>

aCC=Mean correlation. bDV=Mean difference between obtained and expected r. cNumber of coded correlations.

Table 20
Socio-Economic Status and PDI Correlations

<table>
<thead>
<tr>
<th>SES</th>
<th>CC</th>
<th>sd</th>
<th>DV</th>
<th>sd</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle</td>
<td>.68</td>
<td>.11</td>
<td>.18</td>
<td>.07</td>
<td>7</td>
</tr>
<tr>
<td>Low</td>
<td>.50</td>
<td>.33</td>
<td>.14</td>
<td>.23</td>
<td>3</td>
</tr>
<tr>
<td>Mixed</td>
<td>.38</td>
<td>.21</td>
<td>-.10</td>
<td>.18</td>
<td>17</td>
</tr>
</tbody>
</table>

aCC=Mean correlation. bDV=Mean difference between obtained and expected r. cNumber of coded correlations.
Table 21

Reported Intervention and MDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC^a</th>
<th>sd</th>
<th>DV^b</th>
<th>sd</th>
<th>n^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Intervention</td>
<td>.45</td>
<td>.24</td>
<td>.00</td>
<td>.17</td>
<td>59</td>
</tr>
<tr>
<td>Intervention</td>
<td>.51</td>
<td>.28</td>
<td>.02</td>
<td>.27</td>
<td>16</td>
</tr>
</tbody>
</table>

^aCC=Mean correlation.  ^bDV=Mean difference between obtained and expected r.  ^cNumber of coded correlations.

Table 22

Reported Intervention and PDI Correlations

<table>
<thead>
<tr>
<th>Group</th>
<th>CC^a</th>
<th>sd</th>
<th>DV^b</th>
<th>sd</th>
<th>n^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Intervention</td>
<td>.47</td>
<td>.25</td>
<td>-.01</td>
<td>.21</td>
<td>21</td>
</tr>
<tr>
<td>Intervention</td>
<td>.49</td>
<td>.22</td>
<td>-.02</td>
<td>.22</td>
<td>6</td>
</tr>
</tbody>
</table>

^aCC=Mean correlation.  ^bDV=Mean difference between obtained and expected r.  ^cNumber of coded correlations.
### Table 23

**Intervention Duration and MDI Correlations**

<table>
<thead>
<tr>
<th>Months</th>
<th>CC(^{a})</th>
<th>sd</th>
<th>DV(^{b})</th>
<th>sd</th>
<th>(n)(^{c})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>0.35</td>
<td>0.22</td>
<td>-0.18</td>
<td>0.26</td>
<td>4</td>
</tr>
<tr>
<td>6-17</td>
<td>0.50</td>
<td>0.26</td>
<td>0.04</td>
<td>0.23</td>
<td>8</td>
</tr>
<tr>
<td>18+</td>
<td>0.71</td>
<td>0.32</td>
<td>0.17</td>
<td>0.31</td>
<td>4</td>
</tr>
</tbody>
</table>

\(^{a}\)CC=Mean correlation.  \(^{b}\)DV=Mean difference between obtained and expected \(r\).  \(^{c}\)Number of coded correlations.

### Table 24

**Intervention Duration and PDI Correlations**

<table>
<thead>
<tr>
<th>Months</th>
<th>CC(^{a})</th>
<th>sd</th>
<th>DV(^{b})</th>
<th>sd</th>
<th>(n)(^{c})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-11</td>
<td>0.68</td>
<td>-</td>
<td>0.13</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>12+</td>
<td>0.40</td>
<td>0.14</td>
<td>-0.09</td>
<td>0.19</td>
<td>4</td>
</tr>
</tbody>
</table>

\(^{a}\)CC=Mean correlation.  \(^{b}\)DV=Mean difference between obtained and expected \(r\).  \(^{c}\)Number of coded correlations.
Table 25

Intervention Intensity and MDI Correlations

<table>
<thead>
<tr>
<th>Hrs/Week</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-9</td>
<td>.49</td>
<td>.21</td>
<td>.00</td>
<td>.22</td>
<td>8</td>
</tr>
<tr>
<td>10+</td>
<td>.47</td>
<td>.34</td>
<td>-.01</td>
<td>.34</td>
<td>7</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.

Table 26

Intervention Intensity and PDI Correlations

<table>
<thead>
<tr>
<th>Hrs/Week</th>
<th>CC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sd</th>
<th>DV&lt;sup&gt;b&lt;/sup&gt;</th>
<th>sd</th>
<th>n&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-9</td>
<td>.38</td>
<td>-</td>
<td>-.12</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>10+</td>
<td>.52</td>
<td>.24</td>
<td>.00</td>
<td>.24</td>
<td>5</td>
</tr>
</tbody>
</table>

<sup>a</sup>CC=Mean correlation.  <sup>b</sup>DV=Mean difference between obtained and expected r.  <sup>c</sup>Number of coded correlations.
which teachers or other professionals did the pre-testing. In the absence of adequate information, "Other" was coded. For the PDI "paraprofessional" was never coded. Most articles simply did not report data regarding pre-test examiner and were coded "Other". Correlations were stronger for infants who received the pre-test from a psychologist than were groups tested by paraprofessionals or "Other". This finding holds for both PDI and MDI, using both DV and CC means. (See Tables 27 and 28.)

Two non-mutually exclusive explanations may account for this. First, psychologists may administer the Bayley more validly than other persons. Second, psychologists may be employed in projects involving more handicapped children. To check this possibility, the incidence of different codes on the "Handicap" item were summed for each examiner group. Interestingly, all mentally retarded/developmentally delayed samples were tested by psychologists. Whether this is a causal relation cannot be answered by the present study.

The post-test examiner was also coded, with very similar results. However, one PDI and three MDI post-tests were administered by "Other Professionals". Again the samples tested by psychologists had higher interage correlations than subjects tested by other examiners. (See Tables 29 and 30).

Whether an examiner was blind to pre-test results might be expected to influence the interage correlation. Whether the post-test examiner was blind to pre-test was coded "No", "Yes", or "Not Reported". Of 75 MDI correlations, 13 post-test examiners were blind to pre-test results, 3 were not blind, and in 59 cases no information was given. No substantial difference between "Yes and "Don't Know"
Table 27
Pre-Test Examiner and MDI Correlations

<table>
<thead>
<tr>
<th>Tester</th>
<th>CC^a</th>
<th>sd</th>
<th>DV^b</th>
<th>sd</th>
<th>n^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychologist</td>
<td>.65</td>
<td>.23</td>
<td>.18</td>
<td>.14</td>
<td>11</td>
</tr>
<tr>
<td>Para-Professional</td>
<td>.50</td>
<td>.09</td>
<td>-.07</td>
<td>.11</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>.42</td>
<td>.24</td>
<td>-.03</td>
<td>.20</td>
<td>58</td>
</tr>
</tbody>
</table>

^aCC=Mean correlation.  ^bDV=Mean difference between obtained and expected r.  ^cNumber of coded correlations.

Table 28
Pre-Test Examiner and PDI Correlations

<table>
<thead>
<tr>
<th>Tester</th>
<th>CC^a</th>
<th>sd</th>
<th>DV^b</th>
<th>sd</th>
<th>n^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychologist</td>
<td>.63</td>
<td>.18</td>
<td>.18</td>
<td>.07</td>
<td>8</td>
</tr>
<tr>
<td>Para-Professional</td>
<td>.41</td>
<td>.23</td>
<td>-.08</td>
<td>.20</td>
<td>19</td>
</tr>
</tbody>
</table>

^aCC=Mean correlation.  ^bDV=Mean difference between obtained and expected r.  ^cNumber of coded correlations.
### Table 29
**Post-Test Examiner and MDI Correlations**

<table>
<thead>
<tr>
<th>Tester</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychologist</td>
<td>.70</td>
<td>.17</td>
<td>.20</td>
<td>.12</td>
<td>10</td>
</tr>
<tr>
<td>Professional</td>
<td>.23</td>
<td>.06</td>
<td>-.12</td>
<td>.06</td>
<td>3</td>
</tr>
<tr>
<td>Para-Professional</td>
<td>.50</td>
<td>.09</td>
<td>-.07</td>
<td>.11</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>.43</td>
<td>.25</td>
<td>-.02</td>
<td>.20</td>
<td>56</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.

### Table 30
**Post-Test Examiner and PDI Correlations**

<table>
<thead>
<tr>
<th>Tester</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychologist</td>
<td>.68</td>
<td>.11</td>
<td>.18</td>
<td>.07</td>
<td>7</td>
</tr>
<tr>
<td>Teacher</td>
<td>.25</td>
<td>-</td>
<td>.20</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>.41</td>
<td>.23</td>
<td>-.08</td>
<td>.20</td>
<td>19</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.
codes were obtained, either for DV or CC means. (See Table 31.) Six PDI samples were post-tested by examiners blind to pre-test. For 21 PDI samples no information regarding post-test examiner was given. Correlations were slightly lower for blind examiners than for others. (See Table 32.)

Summary

An analysis of the data found in primary research indicates that inter-age correlations from the Bayley scales tend to follow a simplex pattern. Multiple correlations with age at pre-test and interval between tests account for 36% of MDI correlation variance and 26% of PDI correlation variance.

Many of the variables in this study have no clear or consistent relation with cross-age correlation strength. Sample size, race, and whether a sample received early intervention did not appear to covary with reported coefficients. Pre-test PDI means do not appear to be related to cross-age correlations. Among MDI samples which received intervention, intensity of treatment, measured in hours per week, does not coincide with inter-age correlational values.

For many variables there was a lack of coded information, making analysis uncertain or impossible. There were insufficient data to draw conclusions regarding blinding of examiners, race/sex interactions, and, for PDI correlations, pre-test standard deviations, sex, and intervention length and intensity.

Samples with high attrition tended to have higher cross-age correlations than other samples. Low MDI pre-test coincide with
### Test 31

**Blind Post-Test Examiners and MDI Correlations**

<table>
<thead>
<tr>
<th>Blind</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>.57</td>
<td>.12</td>
<td>.11</td>
<td>.04</td>
<td>3</td>
</tr>
<tr>
<td>Yes</td>
<td>.50</td>
<td>.32</td>
<td>.03</td>
<td>.29</td>
<td>13</td>
</tr>
<tr>
<td>Not Given</td>
<td>.45</td>
<td>.23</td>
<td>-.12</td>
<td>.18</td>
<td>59</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.

### Test 32

**Blind Post-Test Examiners and PDI Correlations**

<table>
<thead>
<tr>
<th>Blind</th>
<th>CC(^a)</th>
<th>sd</th>
<th>DV(^b)</th>
<th>sd</th>
<th>n(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>.39</td>
<td>.13</td>
<td>-.05</td>
<td>.20</td>
<td>6</td>
</tr>
<tr>
<td>Not Given</td>
<td>.50</td>
<td>.26</td>
<td>.16</td>
<td>.21</td>
<td>21</td>
</tr>
</tbody>
</table>

\(^a\)CC=Mean correlation. \(^b\)DV=Mean difference between obtained and expected \(r\). \(^c\)Number of coded correlations.
higher stability/predictive validity coefficients. For both PDI and MDI, mentally retarded/developmentally delayed samples had higher correlations than expected, given pre-test age and interval length. Samples mixed with regard to sex had higher MDI correlations than groups which were mostly male or mostly female. Among MDI samples which received intervention, those with greater duration tended to have relatively high correlational values. Finally, samples pre- and post-tested by psychologists had consistently higher cross-age correlations than other samples. Probably connected with this finding is the fact that all MR/DD samples were tested by psychologists.
DISCUSSION

Many factors related to cross-age correlation strength have been identified. Some of these, based on the data collected, do not appear to be substantially important. For example, race and sex do not appear to play a significant role in determining correlations.

Other variables do appear to be important. Samples of MR/DD subjects obtain consistently higher cross-age coefficients than would be expected, given age and interval. A related and interesting finding is that when examiners are psychologists, the stability/predictive validity coefficients are higher. A complicating fact here is that all MR/DD samples were tested by psychologists. Furthermore, only 3% of samples tested by psychologists were nonhandicapped.

For many of the variables in this study there simply were not enough data found to permit conclusions about their relation to cross-age correlations. For example, there were very few data which could be used to examine an interaction between sex and race on correlations.

The age matrix used in data collection proved to be valuable. Although the procedures led to a reduction of obtained data, they permitted an orderly and sensible means of limiting data collection. This takes on particular significance when one encounters studies like Hallowell (1941) in which more than 400 correlations were computed.

It might be argued that the analyses in this study should be conducted with squared correlations. This would change the nature of the data to a more truly interval scale. However, since there were a
few negative correlations encountered, the present method of analysis is most appropriate.

A few disclaimers are in order. As noted by several reviewers, prediction is only one way of assessing the utility and validity of an instrument. There was no intention for this analysis to provide answers regarding the value of the Bayley scales. Rather it is a description of the information in available primary research reports. Second, the correlations reported in this paper do not reflect on the nature of infant intelligence. This paper is only intended to be an assessment of two characteristics of the Bayley scales: predictive validity and stability.

Future Research

This review does not establish firmly that certain factors are important in determining stability and predictive validity of Bayley scores. Rather, it shows that some variables have, in the reported literature, coincided with differential correlational values. What remains to be done is primary research in which the factors found to be important in this meta-analysis are systematically treated. One research project which might be conducted would assess the influence of examiner characteristics on cross-age correlations, controlling for developmental level of subjects. More studies are also needed regarding the predictive validity and stability of the PDI. There is little information available pertinent to the factors which influence strength of PDI coefficients. Finally, there is a need for more information about diagnostic base rates, spontaneous changes, and
their interactions. Many other factors do not appear to exert any significant influence on correlations. For example, research of simple racial correlation differences does not appear to be either promising or interesting. Similarly, single variable studies with sex, age-pre, and interval length will have little to add to existing knowledge.

Summary

There is much existing information regarding the predictive validity and stability of the Bayley Scales of Infant Development. Many factors appear to be related to the strength of cross-age correlations from the Bayley. But only a small portion of the total variance in the correlations found for this study is explained by the identified variables. So we are still left with the question, "How well does the Bayley predict?" This review echoes Honzik's reply, "As in most areas of human behavior, it depends!" (1976, p. 67)
REFERENCES


Erickson, M. T. (1968). The predictive validity of the Cattell infant intelligence scale for young mentally retarded children. American Journal of Mental Deficiency, 72, 728-733.


McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. In J. Osofsky (Ed.),


Appendix A:  
Coded Review Articles


Appendix B
Coded Primary Articles


Appendix C: Coding Sheet

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Number:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R#:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Characteristics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test N:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-test N:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean DQ Pre:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD of DA Pre:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Score:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Score:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean DQ Post:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD of DQ Post:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type Handicap:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% White:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Black:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Hispanic:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Male:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SES:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1=High, 2=Middle, 3=Low, 4=Mixed)
Intervention

Intervention: __ , __ , __ , __
(Did intervention occur? 1=Yes, 0=No.)

Duration: ___ , ___ , ___ , ___

Intensity: ___ , ___ , ___ , ___
(No intervention=0, 1-9 hrs/wk=1, 10-19 hrs/wk=2, 20+ hrs/wk=3)

Measure

Age Pre: ___ , ___ , ___ , ___
Age Post: ___ , ___ , ___ , ___
Age Change: ___ , ___ , ___ , ___
Pre-test Code: ___ , ___ , ___ , ___
(Post-test Code: MDI=1; PDI=2)
Examiner:
(1=Psychologist, 2=Teacher, 3=Other Pro., 4=Parapro, 5=Other)

Type Post-test: ___ , ___ , ___ , ___
Post-test code: ___ , ___ , ___ , ___
Post-Examiner: ___ , ___ , ___ , ___
Blind: ___ , ___ , ___ , ___
Statistic: ___ , ___ , ___ , ___
Type of R: ___ , ___ , ___ , ___
*R*: ___ , ___ , ___ , ___
Source of R: ___ , ___ , ___ , ___
(Reported=1; computed from given data=2)
Appendix D: Coding Conventions

General Rules. Means: Give a measure of central tendency, in the following order of preference: mean, median, mode, midrange.

Guessing: When exact data are not available but there is sufficient data to permit a reasonable estimate, code the estimated value. A guess is "reasonable" if the coder is 90% confident that the coded value is within 10% of the true value. In the case of categorical data, the coder should be 90% confident that the coded value is correct.

DK: When the conventions do not cover how to handle missing data and the information is not provided or cannot be reasonably estimated, code "-9", for DK.

NA: If the item is not applicable and conventions don't cover the situation, code "-7".

It is hoped that as many correlations are coded as is possible. However, it is desirable to maintain some independence of the data. To achieve this, a matrix of age-pre by age-post will be used.
Matrix (Age in Months)

<table>
<thead>
<tr>
<th>Age Post</th>
<th>Age Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-6</td>
</tr>
<tr>
<td>0-24</td>
<td>Cell 1</td>
</tr>
<tr>
<td>25-72</td>
<td>Cell 5</td>
</tr>
<tr>
<td>73+</td>
<td>Cell 9</td>
</tr>
</tbody>
</table>

Only code one correlation per independent group in a cell of the matrix. This will assure that all correlations within a cell will be independent. When a large group of subjects can be broken down into smaller groups, such as males and females, blacks and whites, high and low SES, code them as subgroups if

. . . .(1) the resultant n's are larger than 25, and
. . . .(2) no significant data are lost.

When multiple breakdowns are possible, but the resultant correlations would not be independent, code, in order of preference, breakdowns by (1) sex, (2) race, (3) SES. When multiple outcome measures are reported, code only the correlation with the test which (1) is most closely associated with the MDI or PDI, whichever is being coded, (2) is most well standardized, (3) yields deviation (not ratio) scores, and (4) which has the largest n. If more than one correlation is still possible for a cell (for example, when tests are given at 8, 12, 18, and 24 months) code the correlation that has pre-test and post-test ages closest to the midrange for that cell. (Using the above example, one correlation of 8 to 12 months would be in cell 2 and another would be coded for 18 to 24 months in cell 3. The correlations between 8 and 18, 8 and 24, and 12 and 24 would not be coded.)
Identifiers. In this section there should be no "-9" or "-7" codes. All information must be reported.

ID: Put the four-digit code for the article. If two or more articles report data from the same study, code here the article which contains the most complete data.

Year: Write the year (two digits only). If no year is available, estimate by adding one year to the most recent reference listed in the article.

R#: Enter the correlation number for that study. If two or more articles report on the same data, continue sequencing the numbers as if they were from one article.

Sample Characteristics. Any item for which the information is not available or cannot be reasonably estimated, code "-9". If the item is Not Applicable (NA), code"-7".

"N" Pre: The number of subjects involved at the beginning of the study. This relates only to subjects involved with the correlation being reported. For example, if 100 children (50 males and 50 females) are tested at age 6 months, but only 75 (45 females and 30 males) are retested at 1 year, and the correlation reported is for females, the correct number to enter is 50, not 100.

"N" retest: Number of subjects in the actual correlation. Using the above example, the correct code for this item would be 45.

Mean DQ Pre: This item is designed to help analyze the effects of developmental handicap on stability. Enter the mean MDI or PDI
(whichever is being coded) for the group.

SD OF DQ Pre: This item is included to help examine the influence of restricted range. If SD is reported, enter it. If it can be calculated, use the "N-1" denominator.

Highest DQ Pre: If the range of scores is given, note the highest score here.

Lowest DQ Pre: If the lowest DQ score is reported, note it here. If the lowest score is not given, code "-7". If ratio quotients extrapolated scores are used to estimate MDI's or PDI's below 50, write the lowest result. If the information is not reported, code "-9".

Mean Post Score: Give the mean post score only if (1) it is closely related to DQ/IQ and (2) it is a standard score (mean=100, sd=approximately 15). If the results from a test are not closely related to DQ/IQ or no IQ is reported, code this "-9". If the test yields an IQ-equivalent (such as the PPVT), write that equivalent mean.

SD OF DQ Post: Enter the standard deviation of the post-test IQ or IQ-equivalent scores.

Type Handicap: Using the descriptions below, record the percentage of the sample which, at pre-test, exhibit a handicapping condition. If most of the subjects in the sample (P 50%) fit one of the categories, code that handicap. If most of the subjects are handicapped, but no single diagnosis accounts for 50% of the sample, code "14=Combination".
0=None: use this code if 50% or more of the sample are not handicapped. If no information is provided regarding handicaps, assume that there is no handicapping condition and code "0".

1=Multihandicapped: coded if there are concomitant impairments, such as MR and blind, MR and CP, etc.

2=Hearing Impaired. Code this if at least 50% of the sample have a hearing impairment of such severity that the subject, with or without amplification, is impaired in processing auditory/linguistic information.

3=Visually Impaired: a visual impairment which, even with correction, adversely affects educational performance or developmental advances. The term includes both partially seeing and blind children.

4=Mentally Retarded: significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior and manifested during the developmental period. Do not include autistic children in this category. Assume that Down Syndrome children are MR (or, if indicated, multihandicapped). If MDI is in MR range and adaptive behavior is not mentioned, code 4.

5=Speech/Language Impaired: a communication disorder, expressive or receptive.

6=Orthopedically Impaired: severe orthopedic impairment which adversely affects developmental progress or educational performance. The term includes impairments caused by congenital anomalies, disease (e.g., polio), cerebral palsy, and accidents (e.g., burns causing contractures or amputation).

7=Other Health Impaired: limited strength, vitality or alertness, due to chronic or acute health problems such as heart condition, tuberculosis, rheumatic fever, nephritis, asthma, sickle cell anemia, etc.

8=Emotionally disturbed: exhibiting one or more of the following characteristics over an extended period of time and to a marked degree, with adverse effects on developmental progress and/or educational performance. The term includes autistic and schizophrenic children. The term does not include social maladaptation unless it is an expression of emotional disturbance. Children referred to as hyperactive, hyperkinetic, or ADD should be included in this category.

9=General Developmental Delay: term used for very young children who have delays in more than one area of development, e.g., language, motor, cognitive, social-emotional, self-help. It is used when other labels are not clear-cut and definitive. Children referred for testing or screening may be coded here.

10=High Risk: includes only children determined to be at risk of being or becoming handicapped because of medical (e.g., low birth weight) or genetic (mother MR) reasons. Include here premature infants.
11=Disadvantaged: subjects from poverty, culturally or socially disadvantaged settings.
12=Other: if children exhibit a handicap not clearly included in one of the above codes, note it here.
13=Combination: if 50% or more of the sample are handicapped but no one condition is clearly most prevalent.

% White: What percent of the sample, at the time of post-test, were white? Use whole numbers.

% Black: What percent of the sample, at the time of the post-test, were black? Use whole numbers.

% Hispanic: What percent of the sample, at the time of the post-test, were Hispanic? Use whole numbers.

% Male: Enter, in whole numbers, the percent of the sample who were male. If the sample is described as having an equal or approximately equal number of boys and girls, code 50%.

SES: If SES of the subjects is reported, enter it here. If 80% or more of the subjects are from one level, enter that level. If at least 21% are in other levels, code "4-Mixed."

**Reported Intervention.** This section contains questions regarding reported intervention. Intervention is defined as any setting or services provided to a child beyond that which would normally be given to a healthy child of that age. Sensory stimulation, early education, nutritional supplements, home visits by trained personnel, and physical therapy are all forms of intervention. If no intervention is reported, assume there was none.

Intervention: This is a "Yes/No" item. If 50% or more of the sample received intervention between pre- and post-tests, code 1. If there was no intervention, or none was reported, code 0.
Duration: Record, in whole numbers, how many months the intervention was provided. If duration is reported in weeks, convert to months by the equation: Months=weeksX4.3. (No reported intervention--code 0.)

Intensity: Record here the intensity of the intervention. Code "0" if no intervention was given or reported. Code "1" if intervention occurred an average of 1 to 9 hours per week. Nutritional supplements which essentially do not involve any time directly with the child should be coded "1". Code "2" if intervention was from 10 to 19 hours per week, and "3" if hours per week were equal to or greater than 20.

Measure. Age Pre: Write the age in months at the time of the pre-test. Use means, medians, or midpoint of the range (in that order of preference) for heterogeneous groups.

Age Post: Record the age at the time of the post test, in months.

Age Change: Subtract Age Pre from Age Post.

Pretest Code: MDI= 1; PDI= 2.

Examiner: Code here the person who administered the test. 1=Psychologist (licensed or graduate student), 2=teacher, 3=Other professional, 4=Paraprofessional, such as student teachers, aides, or other persons trained to administer the tests, 5=Other. Code 5 if no information is given.

Type of Post test: Code here the type of post test used. The following list provides examples of the types of tests which should be included in each category. Brief definitions for each code are also
given.

1=Verbal Intelligence Test: Include tests like the verbal portion of the Wechsler Scales, Verbal Scale of the McCarthy, and the verbal part of the CAT.

2=Non-Verbal/Performance Intelligence Test: Include performance IQ of Wechsler's, Perceptual-performance of McCarthy, Progressive Matrices, most drawing tests (DAP, House-Tree-Person), Leiter, Pictorial Test of Intelligence, and Columbia Mental Maturity Scale.

3=Full Scale/General Intelligence Test: A psychological test designed to measure cognitive functions such as reasoning, comprehension, and judgment. Include Wechsler full scale IQs, Stanford-Binet, GCI of McCarthy, Slosson, and Otis-Lennon Mental Ability Test. Note: the Quick Test and PPVT should be coded #10 (Receptive language).

4=Developmental Quotient: Infant scales provide a basis for establishing the child's current status and any deviations from normal expectancy. Include Gesell, the Cattell, Infant Psychological Development Scale (Uzgiris-Hunt, Piagetian), the Griffiths, Bayley MDI, and Alpern Boll.

5=Fine Motor: Small muscle-dependent skills such as reaching, grasping, and eye-hand movement. Include Fine Motor Composite score of Bruininks Oseretsky.

6=Gross Motor: Large muscle-dependent skills such as walking, running, and throwing. Include the Gross Motor Composite of the Bruininks Oseretsky.
7=General Motor: Include Total Battery score on the Bruininks, the motor Scale of the McCarthy, and the PDI of the Bayley.

8=Perceptual Organization: Include perceptual-motor and visual-motor tests, such as the Bender, VMI, Purdue Perceptual-Motor Survey, Frostig, and Revised Visual Retention Test.

9=Expressive Language: Skills required to communicate ideas through language such as writing, gesturing, and speaking. Include tests like the Carros Elicited Language Inventory, Developmental Sentence Analysis, and the Parsons Language Sample.

10=Receptive Language: Language that is spoken or written by others and received by the individual. Include listening, reading, and understanding sign language. Include tests like the Assessment of Children's Language Comprehension, Language Comprehension Test, PPVT, Quick Test, and Vocabulary Comprehension Scale.

11=Articulation: The production of speech sounds. Include tests like Goldman-Fristoe Test of Articulation and the Templin-Darley Test of Articulation.

12=Language Combination or other language: Language tests which do not clearly fit into other categories. Include tests like the Houston Test of Language Development, Utah Test of Language Development, REEL, and the Sequence Inventory of Communication Development.
13=Social Functioning/Adaptive Behavior: Ability of an individual to interact appropriately and effectively with the environment. Include tests like the AAMD Adaptive Behavior Scale, Adaptive Behavior Inventory for Children, Balthazar Scales of Adaptive Behavior, Cain-Levine Social Competency Scale, Preschool Attainment Record, TMR School Competency Scales, and the Vineland Social Maturity Scale.

14=Interpersonal Interaction: Observations or ratings of the quality or frequency of an individual's interactions with others in his/her environment.

15=ITPA: The Illinois Test of Psycholinguistic Abilities is a psycholinguistic measure which is rather unique and should be coded alone.

16=Academic: Readiness tests and achievement tests. Include tests like Boehm Test of Basic Concepts, Key Math, Classroom Reading Inventory, Peabody Individual Achievement Test, WRAT, Woodcock Reading Mastery Test, and the Metropolitan Readiness Test.

17=Psychological or Emotional functioning: Includes behavioral checklists, projective tests, and personality inventories. Include tests like the Devereux Child Behavior Rating Scale, the Burks Behavior Rating Scale, and the Children's Apperception Test.

18=Self-Concept: The person's sense of self-worth, identity, or capabilities. Include tests like the Coopersmith Self-Esteem Inventory, Piers-Harris Children's Self-Concept Inventory.
Scale, and Lipsitt's Self-Concept Rating Scale for Children.

19=Attitude: Typically yield a total score indicating the direction and intensity of the individual's attitude toward a person, policy, program, or other stimulus category. An example is the Likert-type scales or Thurstone-type questionnaires.

20=Health Status/Physical Growth: Soundness/vigor of body and mind; freedom from defect or disease. Measurements of height, weight, and head size are examples.

21=Other: Specify.

Post test Code: From the following list, give the code of the post-test. If the test is not on the list, give it a new number and add it.

Bayley MDI=1
Bayley PDI=2
Cattell Infant Intelligence Test=68
Columbia Mental Maturity Scales=51
Detroit Test of Learning Aptitude=52
Griffiths=67
Hiskey-Nebraska=53
McCarthy=65
Mother/child interaction=64
Neurological exam=63
Northwestern Infant Intelligence Test=66
Pictorial Test of Intelligence=54
Porteus Maze Test=55
Stanford Binet=50
Thorpe Developmental Inventory=57
Vineland=61
WPPSI=56
WISC=60
WISC-R FS=58
WISC-R Verbal=62
WISC-R Perf.=61
WAIS-R=59
Examiner: Indicate who administered the post-test. Use the conventions from the "Examiner" item on pre-test.

Blind to Pretest: Due to halo effect and other sources of possible bias, it is important that the examiner was blind to the pre-test results. Code here "1" if examiner was clearly not blind to pre-test; code "2" if examiner was probably not blind to pre-test results; "3" if examiner was blind to pre-test; code "4" if no information is given.

Statistic Type: Record here the type of statistic used.
1=Pearson Product Moment correlation, 2=Spearman, 3=Kuder-Richardson, and 4=Other. If 4 is coded, note at the side what equation was used.

Type of 'R': 1=Test/Retest, 2=Correlation between one test and another given at a later date, 3=Other (specify in the margin what type), 4=Don't know. Codes 3 and 4 will probably seldom be used.

*R*: Write the correlation coefficient. Occasionally an author (Siegel for example, #2510) will report only the correlations which are significant. To code only the reported correlations biases the meta-analysis toward higher correlations, since only they were reported. It may be best not to code the article at all. If it is determined that the article should be coded, compute the smallest 'r' which would be significant for the sample and divide by 3. This will give a correlation which is fairly small. The reasoning is that nonsignificant correlations probably lie between zero and the computed lowest possible, but may also be negative. The procedure picks a conservative point in between.

Source of R: Code if the above correlation was reported in the article (code 1) or computed from individual scores (code 2).