A Meta-Analysis of Infant Habituation and Recognition Memory Performance as Predictors of Later IQ

Robert B. McCall and Michael S. Carriger
University of Pittsburgh

McCall, Robert B., and Carriger, Michael S. *A Meta-Analysis of Infant Habituation and Recognition Memory Performance as Predictors of Later IQ*. Child Development, 1993, **64**, 57–79. A meta-analytic review of the literature on infant habituation and recognition memory performance as predictors of later IQ suggests several conclusions: (1) Habituation and recognition memory assessments made on a variety of risk and nonrisk samples in the first year of life predict later IQ assessed between 1 and 8 years of age with a weighted (for *N*) average of normalized correlations of .36 or a raw median correlation of .45. (2) The size of the predictive correlation is essentially the same for habituation and for recognition memory paradigms. (3) This prediction phenomenon is not obviously associated solely with one laboratory, one particular infant response measure, or a few extremely disordered infants. (4) The level of prediction to childhood IQ is substantial given the reliability of the infant measures. (5) Predictions are somewhat higher for risk than for nonrisk samples. (6) Predictions are consistently higher than for standardized infant tests of general development for nonrisk but not for risk samples, and they are not consistently higher than predicting from parental education and socioeconomic status or a few other infant behaviors for nonrisk samples. (8) Coefficients may be higher when the predicting assessments are made between 2 and 8 months of age than earlier or later, but prediction coefficients are remarkably consistent across the observed outcome age period of 2–8 years.

Prior to approximately 1960, intelligence, at least as reflected in scores within the normal range on standardized IQ tests, was considered stable, if not fixed, across the lifespan (Hunt, 1961). In 1961, Hunt questioned this assumption, suggesting that mentality was not fixed, at least not early in life. Subsequently, in the 1970s and early 1980s, several students of mental development emphasized the lack of stability in early mental test performance for normal samples (e.g., McCall, 1979; McCall, Hogarty, & Hurlburt, 1972) as well as for samples including at-risk and neurologically disordered infants (Kopp & McCall, 1982). This led to the conclusion that standardized tests of infant development did not predict later intelligence at useful levels until after 18 to 24 months of age for either nonrisk, at-risk, or some obviously disordered groups. This lack of stability, it was thought, was partly the result of profound qualitative changes in the nature of intelligence from infancy through childhood to adulthood (e.g., Kopp & McCall, 1982).

But the search for early predictors of later IQ continued (McCall, 1981), and recently habituation and recognition memory measured during the first year of life have been found to predict later IQ scores (e.g., Bornstein & Sigman, 1986). This finding suggests that while the sensorimotor capacities measured by standardized infant tests are not related conceptually or empirically to childhood intelligence, the encoding, storage, retrieval, discrimination, and recognition presumably measured by habituation and recognition memory tests may be related to the vocabulary, abstract reasoning, and memory skills assessed on childhood intelligence tests (e.g., Bornstein, 1985; Bornstein & Sigman, 1986; Colombo & Mitchell, 1988, 1991; Fagan, 1988; Fagan & Singer, 1983). The previous failure to predict, it was assumed, was simply a consequence of measuring the wrong attributes in infancy.

The Prediction Phenomenon

More specifically, *habituation* is defined as the decrement of attention or re-

The authors thank John Colombo, Jerome Kagan, and several psychology faculty and graduate students at the University of Pittsburgh for their detailed comments on an early draft of this paper; an unknown reviewer who suggested plotting the raw *r*'s; Cathy Kelley for preparing the manuscript endless times; and Trisha McCall for preparing some of the figures. Reprint requests and inquiries may be directed to Robert B. McCall, Office of Child Development, 2017 Cathedral of Learning, University of Pittsburgh, Pittsburgh, PA 15260.

[*Child Development*, 1993, **64**, 57–79. 0009-3920/93/6401-0005$01.00]

sponsiveness to a repeatedly presented or continuously available stimulus (called the "standard" or "familiar" stimulus) that is not simply a consequence of sensory receptor fatigue. Currently, habituation is often measured by the number of presentations of the standard stimulus (presented for as long as an infant visually fixates it) that are necessary for the infant to look 50% as long as he or she looked on the first presentation(s) of that stimulus. However, the percent decrement in attention during familiarization, the total looking time during the familiarization period, or the response to a new stimulus presented after familiarization also have been used as the predicting measure (see Table 1). Presumably, rapid habituation reflects the infant's ability to quickly encode a stimulus into memory, to recognize the familiar stimulus when it is presented again, and to cease looking further at it.

Typically, *recognition memory* is assessed by presenting for 5 to 30 sec a pair of identical stimuli (i.e., the "familiar" or "standard" stimulus) for familiarization, followed by two very brief test trials (e.g., 5 to 20 sec each) consisting of a simultaneous presentation of the familiar (i.e., standard) and a completely novel stimulus with the left-right positions reversed on the second trial. The predicting variable typically is the ratio of looking time to the novel divided by the total looking time to both familiar plus novel stimuli during the two test trials. Presumably, recognition memory reflects the infant's ability to encode a stimulus into memory, to recognize that stimulus as familiar or an alternative stimulus as not being familiar, and to cease looking at the familiar and/or to look longer at the novel stimulus.

The mental functions required by habituation and recognition memory collectively have been interpreted to reflect "information processing" capacities, such as the speed, accuracy, and completeness of encoding a stimulus into memory; the ability to recognize a familiar stimulus; and the propensity to avoid looking at the familiar and to study the novel stimulus (Bornstein & Sigman, 1986; Colombo & Mitchell, 1988, 1991; Fagan & McGrath, 1981; Lewis & Brooks-Gunn, 1981; McCall & Carriger, 1991; O'Connor, 1980; O'Connor, Cohen, & Parmelee, 1984; S. Rose, Feldman, & Wallace, 1988; D. Rose, Slater, & Perry, 1986; S. Rose & Wallace, 1985; Ruddy & Bornstein, 1982; Sigman, Cohen, Beckwith, & Parmelee, 1986). These skills would seem to be more obviously related to capacities measured by childhood standardized tests of intelligence (Bornstein & Sigman, 1986) than would the sensorimotor, action-consequence, and imitation skills represented on standardized infant tests (McCall, Eichorn, & Hogarty, 1977).

Regardless of the actual processes involved, individual differences in habituation and recognition memory do seem to predict later intelligence test scores better than those associated with standardized tests of general developmental level. Bornstein and Sigman (1986) reviewed the early prediction literature and found that habituation and recognition memory assessed between birth and 7 months of age correlated approximately .47 (median of raw *r*'s) with childhood intelligence (usually IQ) assessed between 2 and 8 years of age, and the degree of relation was the same for each paradigm. In comparison, the average correlation between scores on standardized infant tests given between 1 and 6 months of age and childhood intelligence measured between 5 and 7 years of age for nondisordered samples is .09, although this correlation is somewhat higher when predicting from 7 to 12 months ($r = .20$) and at every infant age for risk ($r = .54–.57$) and disordered ($r = .26–.51$) samples (Kopp & McCall, 1982). Presumably, habituation and recognition memory predict better because those measures reflect abilities that are more similar to those underlying intelligence test performance than do standardized infant tests.

Early Criticisms

However, while this newly found predictability has generated considerable excitement, some scholars have been more tempered, if not skeptical, in their response (e.g., McCall, 1981). Their concerns have taken three separate but related forms: (1) the underlying nature of habituation and recognition memory may not be what it appears, (2) the reliability and generality of the infant measures are poor, and (3) the observed phenomenon may be artifactual.

The nature of habituation and recognition memory performance.—Some researchers have disputed that habituation or recognition memory reflect any interesting "cognitive" process (Lecuyer, 1988, 1989; Malcuit, Pomerleau, & Lamarre, 1988). Malcuit et al. (1988), for example, argued that habituation procedures (and, by implication, recognition memory procedures) are nothing more than an operant paradigm involving a synchronous reinforcement schedule. Spe-

TABLE 1

Studies and Samples Used in the Meta-Analyses

Authors (Lab Letter)	Sample Size	Included Preterm	Included Low SES	Included Disordered	Infant Age	Child Age	Predictor Measure	Outcome Measure	Correlation Coefficient
Habituation:									
Slater, Cooper, Rose, & Morrison, 1989 (S)	11	No	. . .	. . .	6 mos.	8 yrs.	Mean length of fixation to novel stimulus (response recovery)	Wechsler	.61
	21	No	. . .	. . .	3 mos.	1 yr. 6 mos.		Bayley	.43
Sigman, Cohen, Beckwith, & Parmelee, 1986 (P)	91	Yes	Yes	Yes	Birth	8 yrs.	Total length of fixation to a redundant stimulus	Wechsler	.29
Bornstein, 1985 (B)	18	No	. . .	. . .	5 mos.	2 yrs.	Index-first fixation, total fixation, decrement in fixation to a redundant stimulus	Other	.55
O'Connor, Cohen, & Parmelee, 1984 (P)	28	Yes	Yes	Yes	4 mos.	5 yrs.	Change in heart rate to novel stimulus (response recovery)	Binet	.60
Bornstein, 1984 (B)	14	No	. . .	. . .	4 mos.	4 yrs.	Total length of fixation to a redundant stimulus	Wechsler	.54
Cohen & Parmelee, 1983 (P)	96	Yes	Yes	Yes	Birth	5 yrs.	Total length of fixation to a redundant stimulus	Binet	.29
Ruddy & Bornstein, 1982 (B)	20	No	No	No	4 mos.	1 yr.	Percent response decrement to a redundant stimulus	Bayley	.46

TABLE 1 (*Continued*)

Authors (Lab Letter)	Sample Size	Included Preterm	Included Low SES	Included Disordered	Infant Age	Child Age	Predictor Measure	Outcome Measure	Correlation Coefficient
Lewis & Brooks-Gunn, 1981 (L)	22	No	No	. . .	3 mos.	2 yrs.	Fixation to a novel stimulus (response recovery)	Bayley	.52
	57	No	Yes	. . .	3 mos.	2 yrs.		Bayley	.40
Miller et al., 1979 (M)	29	No	. . .	. . .	1,2,3 mos.	3 yrs. 3 mos.	Response decrement in first fixation	Language Comprehension	.39
Lewis, Goldberg, & Campbell, 1969 (L) ..	40	No	. . .	. . .	12 mos.	3 yrs. 8 mos.	Response decrement to a redundant stimulus	Binet	.39
Recognition memory:									
Rose, Feldman, Wallace, & McCarton, 1989 (R)	46	Yes	Yes	Yes	7 mos.	5 yrs.	Percentage of looking time to novel stimulus	Wechsler	.61
	45	No	Yes	No	7 mos.	5 yrs.		Wechsler	.54
Rose, Feldman, & Wallace, 1989 (R)	40	Yes	Yes	Yes	7 mos.	4 yrs.	Percentage of looking time to novel stimulus	Binet	.58
	42	No	Yes	No	7 mos.	4 yrs.		Binet	.45
Gottfried, Guerin, & Bathurst, 1989 (G)	130	No	No	No	6 mos.	8 yrs.	Percentage of looking time to novel stimulus	Wechsler	.25
DiLalla & Fulker, 1989 (D)	58	. . .	. . .	. . .	8 mos.	3 yrs.	Visual anticipation of the side a redundant stimulus would appear[a]	Binet	.23
Colombo, Mitchell, Dodd, Coldren, & Horowitz, 1989 (C) ..	23	No	. . .	No	4 mos.	1 yr. 4 mos.	Mean percentage of looking time to novel stimuli across five tasks	Binet Spatial Task	.53

Rose, Feldman, & Wallace, 1988 (R)	84	Yes	Yes	Yes	6 mos.	3 yrs.	Percentage of looking time to novel stimulus	Binet	.46
Fulker et al., 1988 (D)	51	No	No	No	9 mos.	3 yrs.	Percentage of looking time to novel stimuli	Binet	.32
	143	No	No	No	9 mos.	2 yrs.		Bayley	.01
Rose & Wallace, 1985 (R)	35	Yes	Yes	Yes	6 mos.	6 yrs.	Mean proportion of looking time to novel stimuli across six tasks	Wechsler	.56
Fagan, 1984 (F)	36	No	No	. . .	7 mos.	5 yrs.	Mean percentage of looking time to novel stimuli across three tasks	PPVT	.42
Caron, Caron, & Glass, 1983 (A)	31	No	No	Yes	5 mos.	3 yrs.	Percentage recovery of fixation to a novel stimulus[b]	Binet	.42
Fagan & McGrath, 1981 (F)	19	No	No	No	5 mos.	4 yrs. 4 mos.	Percentage of looking time to novel stimulus	WISC Vocab	.33
	20	No	No	No	4 mos.	6 yrs. 6 mos.		WISC Vocab	.66
	19	No	No	No	5 mos.	7 yrs. 6 mos.		WISC Vocab	.46
O'Connor, 1980 (O)	17	Yes	No	No	4 mos.	1 yr. 6 mos.	Change in heart rate to novel stimulus[c]	Bayley	.61
	12	Yes	No	No	4 mos.	1 yr. 6 mos.		Bayley	.06
Yarrow, Klein, Lomonaco, & Morgan, 1975 (Y)	39	No	. . .	. . .	6 mos.	3 yrs. 7 mos.	Percentage of looking time to novel stimulus	Binet	.35

NOTE.—. . . signifies that this information could not be determined from the article.

[a] This was a procedure suggested by Marshall Haith. While it is neither traditional habituation (decrement in responding to a redundant signal is not measured) nor recognition memory (response to novel relative to familiar stimuli is not measured), it was included under recognition memory because presumably the infant was required to learn an alternating left-right position sequence which was displayed by choosing between two simultaneously available alternatives (i.e., left vs. right) which characterizes recognition memory paradigms.

[b] Caron, Caron, and Glass (1983) presented two familiar faces followed by two novel faces and measured fixation to the novel divided by the fixation to novel plus familiar.

[c] O'Connor (1980) presented alternating 400 and 1,000 Hz pure tones, replaced the first with a 200 Hz pure tone, and measured cardiac deceleration to the change in tone.

cifically, a particular behavior (an ocular movement or head turn) leads to a particular consequence (a stimulus enters the visual field), which reinforces the attending behavior. With repeated exposure, the consequence (i.e., the stimulus) loses its reinforcing properties, causing the behavior (i.e., looking) to stop. Therefore, habituation reflects nothing more than an operant behavior plus the well-known decline in the reinforcing potency of sensory/perceptual stimulation.

Lecuyer (1989) also suggested that individual differences in habituation may not index information processing. For example, he observed that infants can distinguish novel from familiar stimuli prior to having habituated to the familiar stimulus (Kagan, 1989; Lecuyer, 1989). Presumably, then, it is not necessary to habituate to a stimulus to process information about that stimulus, so the important information processing may not be reflected in habituation measures at all. Rather, habituation may simply reflect the ability to allocate attention to the various familiar and novel aspects of the environment. Although this ability may have importance in its own right, this argument potentially impugns the conceptual link between habituation and childhood measures of intelligence. Generally, researchers in the field have argued against these extreme conceptual criticisms (see Kuhn, 1989).

Reliability and generality problems.—The short-term test-retest reliability of habituation and recognition memory behavior has been found to be quite low—approximately .30 to .45 (e.g., Bornstein, 1989; Bornstein & Sigman, 1986; Cohen, 1988; Fagan, 1984; Fagan & McGrath, 1981; Fagan & Singer, 1983; Lecuyer, 1989; McCall, 1989; Slater, 1988), varying with the individual measure (Colombo, Mitchell, & Horowitz, 1988; Colombo, Mitchell, O'Brien, & Horowitz, 1987). This level of reliability is no higher and often lower than the size of the predictive correlations. In addition, the generality of habituation and recognition memory performance across different stimuli and sensory modalities is similarly low (see Kagan, 1989; McCall, 1989). These observations, which are not disputed, raise questions about the nature of the endogenous thread that presumably ties together these different behaviors across the early years of life.

Prediction artifact.—Some investigators have suggested that the relation between measures of habituation and recognition memory assessed during infancy and later childhood IQ scores may be an artifact of small sample sizes, the inclusion of a small number of organically damaged or at-risk infants with extreme values (Kagan, 1989; Lecuyer, 1989), or the presence of a subgroup of fast habituators who show high intelligence in childhood due to environmental enrichment (Kagan, 1989).

To demonstrate that this is at least possible, Gottfried (1988, personal communication) and Lecuyer (1989) computed the correlation between sample size and the predictive correlation between habituation/recognition memory and later childhood intelligence as reported in Bornstein and Sigman (1986). They found a correlation of −.60. This indicates that studies with a smaller sample showed higher predictive correlations, raising at least the possibility that the relation was produced or increased by one or two extreme infants who have a more substantial effect on the correlation in a small than in a large sample.

Other criticisms.—Other students of infancy have voiced a variety of additional concerns when discussing this literature. For example, they worry that (1) the evidence for predictability for certain subgroups (e.g., at-risk or disordered infants assessed with the recognition memory paradigm) comes from one or two laboratories, (2) some samples contain infants having a great variety of syndromes known to be associated with retardation, (3) criteria for eliminating subjects from a sample are not always clear, (4) the variability of recognition memory scores is too high relative to the mean and the total exposure duration to permit much meaningful prediction, (5) the variability of infant performance is atypically large in the studies producing some of the best predictions, and (6) predictions do not occur for all measures of infant habituation and the best predicting measure is not consistent from one study to another.

Each of these criticisms is aptly leveled at some studies, and collectively they have fueled considerable skepticism about the validity of the prediction phenomenon. The question is whether the entire literature can be explained away by such concerns, and no one has seriously and publicly attempted to do so.

The Current Paper

The literature on the prediction of childhood IQ from habituation and recognition memory was reviewed by Bornstein and Sigman as recently as 1986, and numerous subsequent reviews have been published (e.g., Bornstein, 1989; Colombo & Mitchell, 1991; Rose, 1989; Rose & Feldman, 1991). Why another review?

First, many of the recent reviews are selective, covering only habituation, only recognition memory, or mainly the work of one laboratory. A major purpose of the present paper was to compare habituation and recognition paradigms with respect to the level of prediction, as did Bornstein and Sigman (1986), and with respect to certain parameters of prediction (e.g., *N*, risk status, age), some of which have been given potential importance by the recent criticisms described above.

Second, the current literature is substantially larger. Early findings generated more research, and at least 77% more studies and 63% more samples are now available than when Bornstein and Sigman (1986) wrote their influential review. In addition to providing a larger database for review, the expanded literature now permits a more detailed meta-analysis of several parameters of this prediction phenomenon. Therefore, although only a few years have elapsed since Bornstein and Sigman's 1986 paper, this literature is worthy of being reviewed again.

Specifically, we asked several questions that could not be addressed previously because of the limited number of studies then available:

1. Do habituation and recognition memory tasks display similar patterns of correlation with childhood intelligence across several potential parameters?

2. What is the relation between the sample size and the predictive correlations for habituation and recognition memory tasks? Could this signal an artifactual effect based upon one or two extreme scores?

3. Is the risk status of the sample associated with the level of prediction for habituation and recognition memory?

4. What is the pattern of predictive correlations for habituation and recognition memory as a function of the age at assessment in infancy and the age at assessment in childhood?

Meta-Analytic Review

Sampling

Relevant articles on the predictive validity of habituation and recognition memory were generated by a computerized search of *Psychological Abstracts* for the years 1974 to 1989, a review of the bibliographies contained in this literature, and an examination of the personal files of the first author. The search produced a total of 23 studies (see Table 1), a substantial increase over the 13 studies reviewed by Bornstein and Sigman (1986).

A study was included in the review if (1) it was reported in full in a standard empirical report or with sufficient methodological detail in a chapter, (2) the sample did not contain frankly disordered infants having known syndromes associated with retardation, (3) it employed habituation, recognition memory, or both measures assessed between birth and 12 months, and (4) it measured childhood intelligence with some measure of general mental performance (usually IQ but sometimes vocabulary or memory) at least 1 year after the infant measure was taken (i.e., between 1 and 8 years of age).

These criteria meant that certain studies that appear relevant were not included in this analysis. For example, some unpublished studies and most data reported in book chapters were typically not included because they often lacked parametric details needed for the meta-analyses. Also, some studies of habituation or recognition memory used unique infant assessments or childhood mental behaviors (e.g., play) but not general IQ or a major component of it (e.g., vocabulary or memory subtests) as the outcome measure. These studies were not included because sufficient variability and numbers of studies employing these independent and dependent variables were not available on which to conduct meta-analyses and because they might introduce irrelevant variance if these procedural differences influenced the predictive correlations. Sometimes more than one predicting variable was available, and we tended to favor the most common variable first or, subsequently, that which produced the highest or most consistent predictive *r*. These several choices meant that our set of samples and *r*'s were not always the same as those used by Bornstein and Sigman (1986). The potential bias that this could introduce did not seem to matter (see below).

Seven of the 23 studies that were included employed multiple (2 or 3) samples (e.g., preterm and full-term, younger and older infants), and these independent samples, not the studies, constituted the unit of analysis here. Although it might be argued that multiple samples within a study are not independent, we adopted the strategy of using sample as the unit of analysis, because it would increase the size of the meta-analytic sample and because Bornstein and Sigman (1986) also used sample, not study, as the unit of analysis. It must be acknowledged, however, that most meta-analytic strategies require independent samples, which assumption is violated to an unknown extent when more than one sample comes from the same study. We examined this possibility informally, feeling that the number of instances of multiple samples was too few for more formal statistical treatment.

Thirty-one samples from 23 studies were analyzed, 12 samples from 10 habituation studies and 19 samples from 13 recognition memory studies. This is a substantial increase over the previous review (Bornstein & Sigman, 1986) of 19 samples, 8 habituation and 11 recognition memory.

Variables

Once the samples were selected, several potential parameters of the subjects, procedures, outcome measures, and results were identified (see Table 1).

Independent variables.—For each sample we recorded paradigm (habituation or recognition memory), birth status of the infants (i.e., the sample included some preterm infants or it included only full-term infants), socioeconomic status of the families (i.e., the sample included some low-SES families or the sample included only middle- to upper-SES families), health status of the infants (i.e., the sample included neurologically disordered infants or the sample included only neurologically healthy infants), the nature of the infant and childhood measures, sample size, age of infants at the initial assessment, and age of children at the outcome assessment.

Dependent variable.—The dependent variable for the meta-analyses is the correlation coefficient (typically the z-transformed correlation coefficient, see below) between infant and outcome measures. We used one correlation from each sample. Data from a single sample were sometimes reported in more than one paper, in which case only one r from each sample was used, and, as a result, some reports are not cited. Unfortunately, the lack of details in reports about previous publications on the same sample or changes in a sample over time sometimes made these decisions uncertain.

In those cases in which more than one correlation was available, the correlation between the youngest infant age and the oldest childhood age was selected. This was done because it was possibly (see immediately below) the most conservative approach and because some interest focuses on the ability of the infant measures to predict "mature intelligence."

Ordinarily, as stated above, this approach would seem to bias the results toward lower predictions, because the longer the intertest interval the lower the correlation for standardized tests (Kopp & McCall, 1982; McCall, 1989). But this decreasing trend may not occur when predictions are made with habituation or recognition memory. For example, as indicated above, the test-retest reliabilities and short-term stabilities of these infant measures are often lower than their prediction coefficients, and predictions from these infant measures to intelligence tests given in the first 8 years of life may not show the typical decline in r with increasing childhood age (see below). So it is an empirical question whether this approach actually biases the size of the prediction and by how much.

Statistical considerations.—Because the sample correlation is a biased estimator of the population correlation, all sample correlations were subjected to an r-to-z transformation (normalizing). The z-transformed coefficients were then weighted by the size of the sample minus 3, and this quantity was then divided by the sum of all sample sizes minus 3 for the set of samples included in the analysis (e.g., for the entire set of samples, for habituation samples only, and for recognition memory samples only). This has the effect of normalizing the sample correlations and rendering them unbiased estimators of the population correlation. This is the most common method of estimating the population correlation using independent samples (Hedges & Olkin, 1985).

This weighted normalized correlation coefficient, then, was the main dependent variable analyzed below. However, for comparison purposes, we also report in parentheses following the weighted normalized values the nonnormalized (i.e., untransformed) value corresponding to the weighted normalized average r. Note that this is merely transforming the weighted normalized average to a nonnormalized value (z_r to r transformation); it is *not* the mean unweighted nonnormalized value. We also report distributions and scatterplots of raw (i.e., untransformed, unweighted) prediction correlations as well as the median raw r of these distributions. Such plots help to depict graphically certain results, they demonstrate that the transformation and weighting procedures did not distort or produce the findings (except the central tendency of the predictions), and they provide a look at certain parameters that could not be analyzed statistically because of a limited number of cases.

The average transformed, weighted r and the median raw r each have certain advantages and limitations. The average transformed, weighted r is an unbiased estimator of the population average, and the weighting process roughly eliminates sample size as a potential contributor to estimates of central tendency. The median treats each sample as a unit and weights each sample equally, but it minimizes the influence of extreme results. While the weighted average r is a good estimate of the *population,* the median raw r is a good estimate of the typical *reported r.*

Using a single r per sample was an attempt to satisfy the statistical requirement that independent r's be used in a meta-analysis. However, dependencies could exist among r's derived from different samples but from the same laboratory, sometimes reported in the same article. While it is possible to deal with such potential dependencies in a statistical manner, the number of such samples was small. So we observed these possibilities more informally by examining scatterplots (see below).

Analytic strategy.—Meta-analyses were conducted in two steps. First, the weighted normalized correlation coefficients were analyzed to determine their difference from zero and their homogeneity within a set of samples. Second, the weighted normalized coefficients were entered into separate correlational and difference-between-means analyses to determine if they varied as a function of subject and/or procedural variables. However, the *un*weighted normalized correlation coefficients were entered into correlational analyses with sample size (which otherwise is used to determine the weights). Then graphic displays of raw r's are presented to illustrate and extend the statistical results.

The Size and Significance of the Prediction of Later IQ from Infant Habituation and Recognition Memory

The first task was to test the size and significance of the relation between habituation and recognition memory assessed during infancy and childhood IQ. This was done for all samples regardless of paradigm and then separately for habituation and recognition memory samples. Further, tests of homogeneity were conducted to determine if the set of correlations could be viewed as being sampled from a single population or whether the correlations might derive from two or more populations, suggesting the influence of sampling or procedural parameters that varied within the set of samples (Hedges & Olkin, 1985).

All samples.—The average weighted normalized correlation coefficient over all samples was .39 (comparable to a weighted but nonnormalized $r = .36$), which was significantly different from zero ($z = 13.82$, $p < .001$).

Across all samples, the hypothesis of homogeneity of the weighted normalized correlation coefficients was rejected, $Q(30) = 47.99$, $p < .001$. Therefore, one might infer that one or more sampling or procedure parameters that varied among samples in the total set influenced the size of the predictive relation.

Habituation versus recognition memory.—The first potential parameter that might influence the predictive relation was type of infant assessment paradigm (i.e., habituation vs. recognition memory).

The separate average coefficients for the habituation and the recognition memory samples were each significantly different from zero. The average weighted normalized correlation coefficient for habituation was 0.41 ($r = 0.39$; $z = 8.40$, $p < .001$), and the average for recognition memory was 0.37 ($r = 0.35$; $z = 10.60$, $p < .001$). The weighted normalized correlation coefficient for the habituation samples was not significantly greater than the comparable coefficient for recognition memory samples. These and subsequent results are summarized in Table 2.

Although the weighted normalized correlation coefficients were not different for habituation and recognition memory paradigms, the significant heterogeneity result suggests that other parameters of the samples and procedures may influence the size of the predictions, perhaps interacting with paradigm. To evaluate this possibility in general, tests of homogeneity were conducted separately on the habituation and recognition memory samples. The coefficients for the habituation samples were homogeneous within sampling error, $Q(11) = 6.77$, $p < .10$, but the coefficients for the rec-

TABLE 2

MEAN WEIGHTED, NORMALIZED CORRELATION COFFICIENTS

Variable	Full Sample	Habituation	Recognition Memory
Paradigm:			
Habituation	.41 (r = .39)[a]		
Recognition memory	.37 (r = .35)		
Preterm infants:			
Included	.47 (r = .43)*	.35 (r = .33)[b]	.58 (r = .52)***
Not included	.35 (r = .34)	.48 (r = .45)	.30 (r = .29)
Low socioeconomic status:			
Included	.48 (r = .45)***	.35 (r = .33)[b]	.58 (r = .52)***
Not included	.27 (r = .26)	.54 (r = .49)	.24 (r = .24)
Disordered infants:			
Included	.45 (r = .42)*	.30 (r = .29)[b]	.58 (r = .52)***
Not included	.31 (r = .30)	.61 (r = .55)	.28 (r = .28)

[a] Values in parentheses represent the nonnormalized correlation corresponding to the weighted normalized r.

[b] Groups differ by sample size, and therefore differences in weighted, normalized correlation coefficients may be confounded by sample size. Groups did not differ by unweighted normalized correlation coefficients.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

ognition memory samples were not homogeneous, $Q(18) = 40.68$, $p < .01$.

The above test is not equivalent to a test of the *difference* in homogeneity between paradigms, and the closeness to significance of the habituation test guards against the conclusion that the paradigms are different in this way. So, we related the weighted normalized correlation coefficients to various potential parameters separately for both the habituation and recognition memory samples.

Raw data.—The use of transformed and weighted correlations is appropriate for the statistical meta-analysis, but some would argue that such procedures may distort results. Consequently, we present simple distributions and scatterplots of *untransformed, unweighted* r's as a complement to the formal meta-analyses. Specifically, Figure 1 presents the distribution of r's separately for the habituation and recognition memory paradigms taken from Table 1 in which the r's are represented by a Lab Code Letter corresponding to each laboratory (also given in Table 1). These distributions illustrate and amplify the statistical results reported above.

First, the median raw r is exactly .45 for each paradigm and for the combined sample. This value is higher than the average weighted transformed r of .39 (or its nonnormalized equivalent of .36) in part because of sample size. As reported below, r's were higher for small samples. The weighting process essentially combines subjects across samples, ignoring individual sample sizes, whereas the median allows sample size to operate. The former value (.36) is a better population estimate, but the latter (.45) is more typical of values reported in the literature.

Second, the variability is only slightly greater for recognition memory than for habituation paradigms. This corresponds to the statistical results for homogeneity reported above.

Third, while in Figure 1 there appears to be greater similarity among r's within than between laboratories, and some laboratories tend to report higher r's than others, the general prediction phenomenon is not obviously and exclusively tied to one or even two laboratories.

While the predictive response measure for recognition memory paradigms is usually percent of looking time to the novel relative to the sum of looking to novel plus familiar stimuli (see footnotes to Table 1 for exceptions), a much greater variety of response measures has been used in habituation paradigms. These measures fall roughly into three types: *Habituation (H)*—response decrement over familiarization trials, *Fixa-*

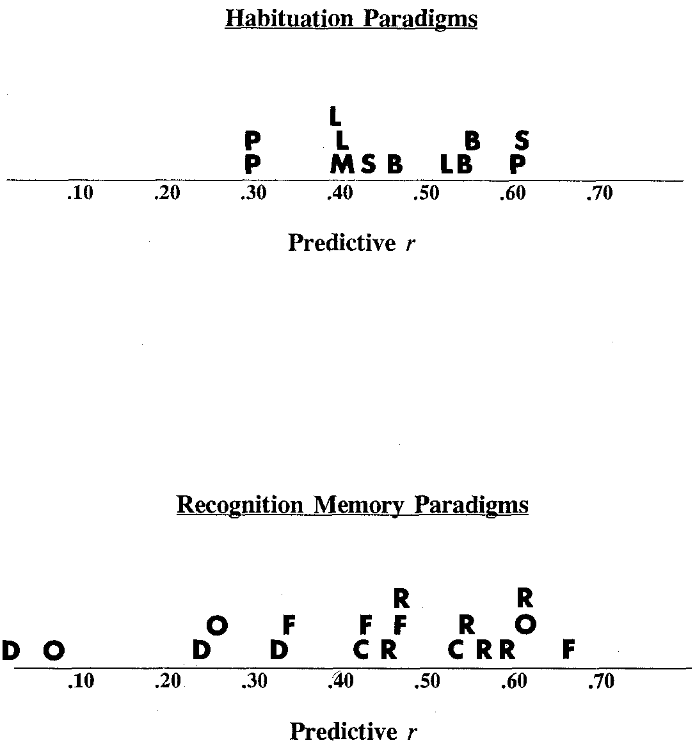

FIG. 1.—The distribution of raw predictive *r*'s for habituation and recognition memory paradigms designated by the Lab Codes given in Table 1.

tion (F)—the average or total looking at the standard stimulus during the familiarization phase, and *Recovery (R)*—the amount of response recovery to the novel stimulus relative to the last familiar stimulus.

Figure 2 presents the same distribution of untransformed *r*'s for habituation paradigms as in Figure 1, but this time each *r* is represented by the type of infant predictive response measure (i.e., *H, F, R*). While the number of *r*'s is too small for formal statistical treatment, a few trends for future systematic study are apparent.

First, the prediction from habituation paradigms does not rest solely on any one of the three infant response measures. All three types of predictors can be found around the median *r* of .45, and no type of predictor

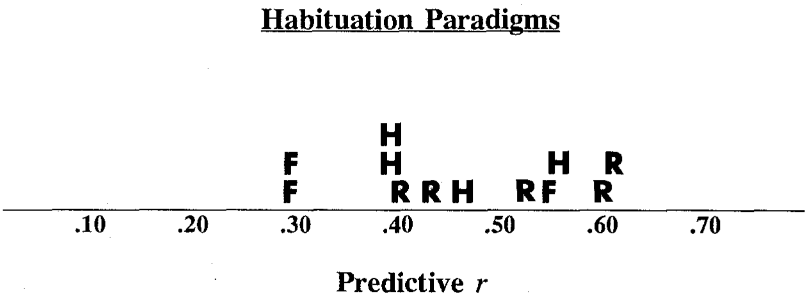

FIG. 2.—The distribution of raw predictive *r*'s for habituation paradigms designated by the type of predictor response: *H* = habituation or response decrement during familiarization, *F* = average or total fixation time during familiarization, and *R* = recovery of response to the presentation of a new stimulus following familiarization.

measure always falls above or below the median. Second, and notwithstanding the first point, there is a slight tendency for the lowest predictive *r*'s to be based on average or total fixation times during familiarization (*F*) and the highest predictive *r*'s to be based on the recovery of response to a novel stimulus presented immediately after the familiarization phase. However, in addition to the small size of this trend, it should be observed that the two lowest *r*'s for Fixation both derived from the same laboratory, and two of the three highest *r*'s for Recovery come from the same (but another) laboratory. More tantalizing, though, is the observation that the same laboratory that reported the two lowest values for Fixation reported one of the highest values ($r = .60$) for Recovery. It should be noted, however, that sometimes studies assessed several predictor measures, only one of which is presented in the current review. Much more systematic study of the possible differential predictability of various response measures, especially from previously reported data, might help to focus interpretation on a particular mental process that mediates this predictive relationship (see McCall & Carriger, 1991).

Sampling and Procedural Parameters

The weighted normalized correlation coefficients were entered into a series of correlational analyses and *z* tests for the difference between average normalized correlation coefficients to determine their potential association with the inclusion of preterm infants, the inclusion of low socioeconomic status infants, the inclusion of disordered infants, sample size, and age of subject at infant and at outcome assessment.

Risk samples.—Three series of *z* tests for the difference between weighted normalized correlation coefficients for the various sampling groups (i.e., risk factors of the presence of preterm, low SES, or disordered infants) were conducted, one for the entire set of samples and separately for samples employing a habituation and a recognition memory assessment (see Table 2 for details).

For *all samples,* the weighted normalized correlation coefficients were significantly *greater* for samples that included preterm infants, low socioeconomic status families, and neurologically damaged infants.

For the *habituation samples,* the weighted normalized correlation coefficients were not significantly related to any subject variable, but *r*'s tended to be *lower* in atypical samples.

For the *recognition memory samples,* however, the weighted normalized correlation coefficients were significantly *greater* for samples including preterm infants, low SES families, and neurologically damaged infants.

Therefore, it appears that atypical samples produce higher predictions to later IQ, especially for recognition memory paradigms (but see next section).

Figure 3 presents these data graphically, but for untransformed, unweighted *r*'s and for a single risk dimension. Specifically, the three indices of risk in Tables 1 and 2—the presence of preterm, low SES, or disordered infants—were not independent, because seven out of 13 samples that had one risk factor had all three. So samples were divided simply into No-Risk and Risk, the latter being any sample that had at least one of the three risk-factor groups included. Second, as one might expect, some laboratories concentrated on risk samples, and because only 13 risk samples are available from only five laboratories, confounds between laboratory and risk samples are likely at this early stage of the literature. So, distributions of predictive correlations represented by Lab Code Letters are graphed in Figure 3 separately for Risk and No-Risk samples within habituation (top) and recognition memory (bottom) paradigms.

For habituation paradigms, the four available *r*'s for risk samples span the entire range but include the two *lowest* *r*'s, both of which derive from the same laboratory (*P* = Parmelee), but one which also contributed the highest Risk *r*. For recognition memory paradigms, the nine risk samples also span the range of predictive *r*'s, but they predominate at the *high* end, again largely due to one laboratory (*R* = Rose). Therefore, there is a slight tendency for risk samples to produce higher predictive *r*'s, at least within recognition memory paradigms (but see next section).

Sample size.—Three series of correlational analyses were conducted, one each for the entire set of samples, the habituation samples, and the recognition memory samples. Sample size was the independent variable, and the *un*weighted but normalized predictive correlation coefficient was the dependent variable. The *un*weighted values

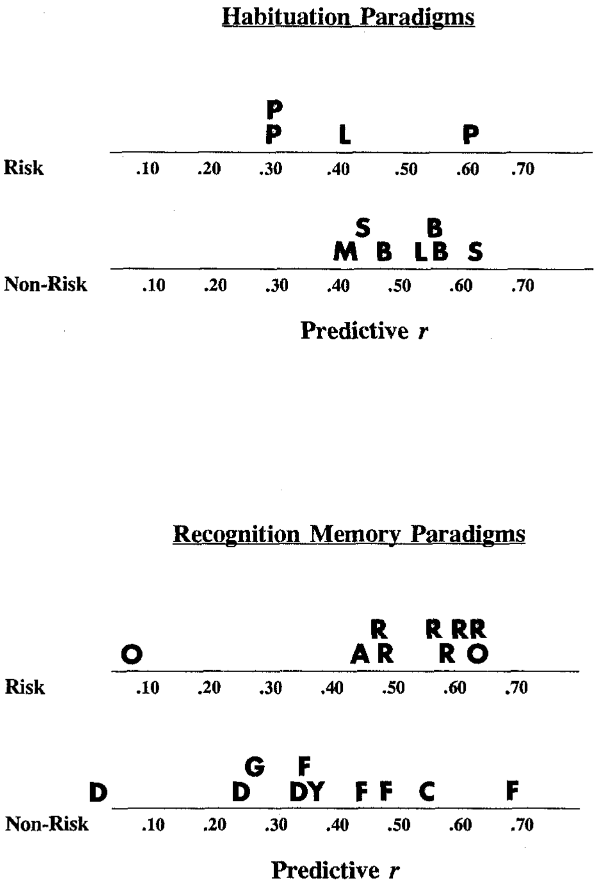

FIG. 3.—The distribution of raw predictive *r*'s for habituation and recognition memory paradigms as a function of Risk vs. Non-Risk samples designated by the Lab Codes given in Table 1.

were used only in the analyses on sample size, because *r*'s should not be weighted with the variable they are being correlated with (i.e., sample size).

For *all samples,* the unweighted normalized predictive correlation coefficients were significantly inversely related to the sizes of the samples ($r = -.56$, $p < .001$), a result similar to the $-.60$ observed by Gottfried (1988, personal communication) and Lecuyer (1989) for the samples reviewed by Bornstein and Sigman (1986). Samples employing smaller numbers of subjects tended to show larger predictive relations.

This significant inverse relation between sample size and predictive coefficient was also present for both the *habituation* and *recognition memory* samples. Specifically, *r* equaled $-.79$, $p < .01$, for the habituation samples, and *r* was $-.46$, $p < .05$, for the recognition memory samples.

Figure 4 presents a slightly different view of this issue using the untransformed, unweighted *r*'s. Samples were divided as above into Risk (*R*) and Non-Risk (*O*), and then the predictive *r* was plotted as a function of *N* ignoring paradigm (since the relation occurred statistically in both paradigms).

Several points may be observed. First, two samples are clearly extreme, both recognition memory samples. O'Connor's (1980) risk sample of $N = 12$, $r = .06$ is very much off trend (i.e., an "outlier"; Belsley, Kuh, & Welsch, 1980), although the same study produced an on-trend $N = 17$, $r = .61$ (i.e., an "influential observation"). Fulker et al.'s (1988) nonrisk recognition memory sample of $N = 143$, $r = .01$ is extreme in both *N* and *r*, but it is basically an on-trend influential observation.

Second, the weak general trend for Risk

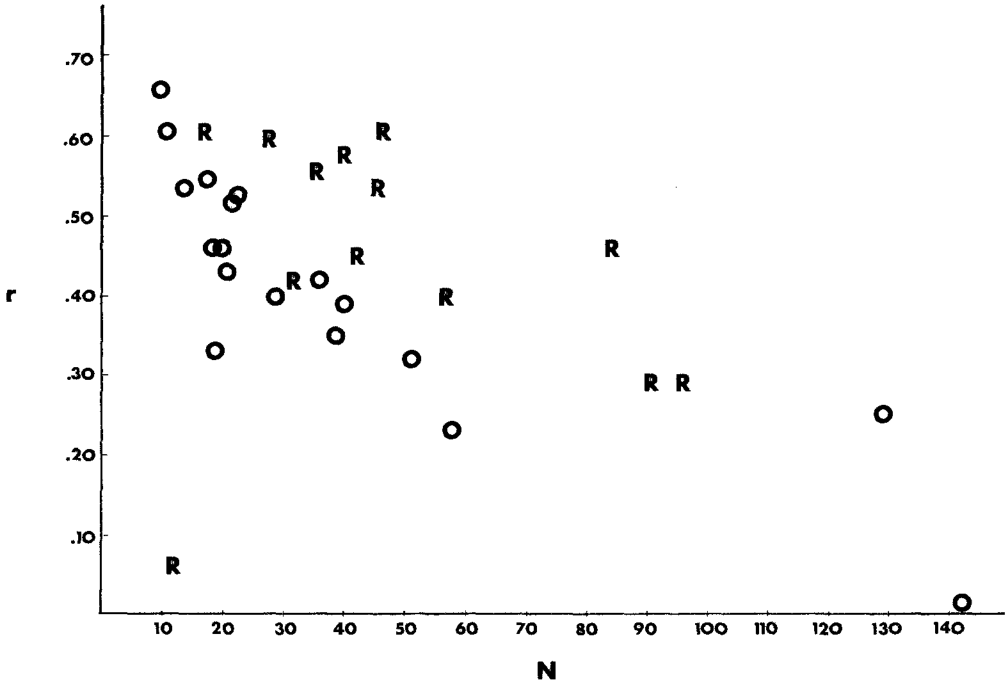

FIG. 4.—The scatterplot of predictive *r* as a function of *N* for Risk (*R*) and Non-Risk (*O*) samples ignoring paradigm.

samples (*R*) to have higher predictive *r*'s than Non-Risk samples (*O*) is more clearly present when viewed conditional upon sample size. That is, collapsed across *N* onto the ordinate of Figure 4 (see also Fig. 3), only a slight tendency can be seen for risk samples to have higher *r*'s as reported above from the statistical analyses. But when viewed within samples of particular sizes (i.e., conditional distributions), the risk sample *r*'s are clearly above the regression line of *r* on *N*, and this is true for both paradigms (separate scatterplots were examined but not presented here). Therefore, the trend toward higher predictive *r*'s for risk samples *of comparable sizes* is general across paradigms, a theme consistent with the literature predicting later IQ from standardized infant tests (Kopp & McCall, 1982).

Third, even excluding the extreme values, a trend toward higher *r*'s for smaller samples exists for Risk (*R*) and Non-Risk (*O*) samples and for both paradigms (not separately shown graphically).

A major issue, however, is whether the inverse relation between *N* and *r* reflects a few extreme points *within individual samples* that produce the high predictive correlations, especially within small samples (e.g., Kagan, 1989). While it is impossible to determine the answer to this without examining scatterplots for individual samples, a case can be made that this potential artifact is unlikely to explain away the entire predictive phenomenon. If this hypothesis were true, one would expect risk samples to be more likely to contain extreme scores and therefore have higher *r*'s and, crucially, a larger inverse *r*-to-*N* relation. While the predictive correlations are higher for risk samples, the uniformity of this relation across all levels of *N* suggests it is not produced solely by extreme scores, which should have less influence in larger samples. Furthermore, the *r*-to-*N* relation exists equally in risk and nonrisk samples. More specifically, this relation occurred for samples having no preterm ($r = -.79, p < .001$), low SES ($r = -.61, p < .01$), no low SES ($r = -.64, p < .001$), disordered ($r = -.59, p < .01$), and no disordered ($r = -.62, p < .001$) infants. The relation was weakest for samples having preterm infants ($r = -.28$), opposite to what one would expect if developmentally persistent disordered subjects were producing these predictions.

In short, while higher predictive *r*'s are obtained for smaller samples, this does not appear to be simply produced by a few extreme scores, presumably from disordered infants (but no scatterplots for individual samples are presented in the literature).

Age of assessment in infancy and childhood.—For all samples, the weighted normalized correlation coefficients were not found to be correlated significantly with either the age of the infants at the initial assessment or the age of the children at the outcome assessment. This also was true within both the habituation and recognition memory sets. However, infants responding to habituation tasks were younger (average of 3.83 months of age) than infants responding to recognition memory tasks (6.10 months of age), $F(1, 29) = 7.17$, $p = .01$.

While the meta-analyses did not reveal an effect of age at infant testing, this analysis is limited in two ways. First, in only three samples was the recognition memory assessment given at ages older than 7 months (i.e., at 8 and 9 months), and in only one sample was the habituation assessment administered at an age older than 6 months (i.e., at 12 months). Similarly, infants in only two samples in either paradigm were assessed before 2 months of age. Therefore, there is little power in our analyses to detect trends that might occur before 2 months or after 8 months. Furthermore, such analyses could not reveal any combined effect of age at infant and age at childhood assessments.

To look at the age data more closely, at least for heuristic purposes, Figure 5 presents the predictive correlations (ignoring paradigm and risk status) as a function of the age of infant and childhood assessments, with the marginal median correlations given within parentheses along each axis.

Notice first the marginal medians along the ordinate. They indicate a remarkably consistent level of prediction between $r = .40$ and $.56$ as a function of the age of the childhood assessment between 1 and 8 years. The apparently lower value of .29 at 8 years is fragile because of the great variability at that age (e.g., .25, .29, and .61) and

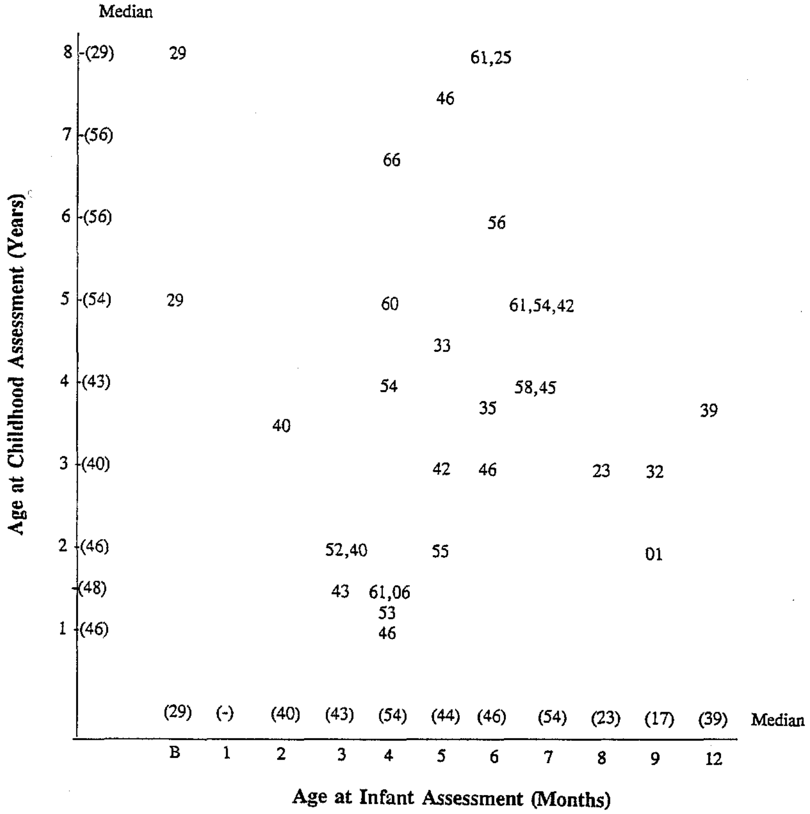

FIG. 5.—The predictive correlations as a function of the age of infant and childhood assessments. Marginal median r's are given in parentheses along each axis.

because the .48 at 7 years, 6 months was arbitrarily included in the calculation of the 7-year rather than the 8-year median. The level prediction pattern between 1 and (at least) 8 years is remarkable, because most longitudinal data display declining prediction coefficients with increases in outcome age and intertest interval. It also means that selecting correlations that span the longest intertest intervals did not bias the results toward lower values.

Now turn to the marginal values in parentheses along the abscissa. While the number of samples is small at crucial points, a trend is suggested, namely, that prediction is best from infant assessments made between approximately 2 and 8 months and poorer from assessments made earlier or later than this period. While the data are very fragmentary, it would not be surprising to find low predictions from the first month of life, since low correlations are typical from such ages in other domains, reliability of measurement is likely poorer, state plays a greater interfering role, etc. But the four relatively lower predictions from 8 months and older are more notable. They represent three different paradigms and measures (r = .01 and .32 for recognition memory, Fulker et al., 1988; r = .23 for the anticipation of left-right stimulus position, DiLalla & Fulker, 1989; r = .39 for response decrement in habituation, Lewis, Goldberg, & Campbell, 1969). They also constitute a prediction pattern that is opposite to the usual increase in correlations with increasing age at infant testing and shorter intertest intervals (note that the intertest intervals for these r's are 2–3 years, whereas most r's for longer intertest intervals are actually *higher*).

Conclusions and Analysis

The meta-analysis, coupled with other elements of the literature, lead to several conclusions about the prediction phenomenon, its parameters, and the processes that mediate the relation.

The Prediction Phenomenon

1. *Habituation and recognition memory assessments made on a variety of risk and nonrisk samples of infants in the first year of life predict later IQ assessed between 1 and 8 years of age with a weighted (for* N*) average of normalized correlations of .36 or a raw median correlation of .45.* The literature now includes at least 31 samples representing a substantial diversity of ages, stimuli, subject characteristics, laboratories, and specific procedures, making it unlikely that *the entire set of predictions is simply a collection of chance phenomena or produced* by a few procedural artifacts. While methodological issues may plague individual studies (e.g., McCall, 1981), it is becoming increasingly unlikely that extraneous factors can explain away the entire literature.

2. *The size of the predictive correlation is essentially the same for habituation and for recognition memory paradigms.* The average of the weighted normalized r's was .39 for habituation and .35 for recognition memory studies, and the median raw correlation was .45 for both paradigms. While comparable prediction coefficients do not necessarily mean the same mechanism operates in each paradigm, this is the simplest and most parsimonious hypothesis. The median raw r's are nearly identical to those reported in 1986 by Bornstein and Sigman.

3. *This prediction phenomenon is not obviously associated solely with one laboratory or one particular infant response measure.* While predictive r's from single laboratories are often higher or lower than average, this is not consistently the case even within a laboratory, and the data from one laboratory are not solely responsible for the entire prediction phenomenon. Similarly, while predictions are slightly higher for some infant measures than others (especially within the habituation paradigm in which a greater variety of measures has been used), no one measure is solely responsible for the general prediction. At the same time, some measures may predict better than others, and more comparisons of different predicting measures within studies are needed.

4. *It seems unlikely that such predictions are simple products of extreme scores, presumably those of a few extremely disordered infants who remain low scoring or retarded in childhood.* Some scholars had raised this as a potential explanation, partly because a correlation of −.60 was found in Bornstein and Sigman's (1986) review between sample size and the size of the predictive correlation (Gottfried, 1988, personal communication; Lecuyer, 1989). Further, Kagan (1989) and Lecuyer (1989) speculated that it was possible that a few organically damaged or at-risk infants alone might produce these correlations, and one might expect this to be especially true in small samples in which one or two individuals would have a more substantial effect.

In the present meta-analysis, this inverse relation between sample size and the level of prediction was −.56 for all samples, −.79 for habituation samples, and −.46 for recognition memory samples. However, contrary to what might have been expected by the extreme-score hypothesis, the negative relation between sample size and predictive *r* was remarkably similar in subsamples containing or not containing at-risk infants. Of course, the only real way to evaluate the extreme-score hypothesis is for individual studies to publish scatterplots for their predictive correlations, which has not been done in the past.

It is crucial to point out, however, that it is still possible, perhaps likely, that *low-scoring infants, especially those from relatively unstimulating and unsupportive home environments, carry a disproportionate amount of the prediction load.* That is, the scatterplot may not reflect a uniformly linear relation or homogeneous oval; the predictive slope may be produced disproportionately by cases in the lower left quadrant. For example, standardized infant tests predict somewhat better for low-scoring than for high-scoring infants (McCall, 1979). Furthermore, infants with prenatal problems and depressed performance on standardized infant tests are more likely to remain low scoring on tests of mental performance during early childhood if they are reared in impoverished or other environments that are less likely to support mental development (Sameroff & Chandler, 1975). It is interesting to note in this regard that the heredity × environment reaction surface for IQ (Turkheimer & Gottesman, 1991) shows that extremely poor environments interrupt the otherwise consistent and fairly high genetic correlations with childhood IQ. Finally, although not always reported or systematically examined in this meta-analysis, the correlation between general measures of socioeconomic status is typically higher with childhood IQ (e.g., .50, McCall, 1979) than with the infant measures (e.g., *r* is approximately .40; Cohen & Parmelee, 1983; Fagan, 1984; Gottfried, Guerin, & Bathurst, 1989; O'Connor, Cohen, & Parmelee, 1984; Rose, Feldman, Wallace, & McCarton, 1989).

These observations collectively support a "Sameroff-Chandler lower-left quadrant prediction phenomenon." That is, while poorly scoring infants may be more likely to come from low SES homes, others score poorly because of temporary medical problems and, with extensive medical care and a rich environment, recover to have average or above IQs as children (top left quadrant of the prediction scatterplot). But those poor scoring infants reared in very mentally unstimulating circumstances may not recover, and it is their presence in the lower-left quadrant that produces much of the prediction correlation (but see Fagan & Knevel, 1989). They are not flukes, because there are more than one or two (especially in risk samples which do produce higher correlations) and because they could be meaningfully identified and explained.

Unfortunately, such a possibility has rarely been examined directly in research using either paradigm. When SES variables have been examined, they typically are combined with the infant assessment in a multiple regression, a statistical procedure that would not be as sensitive as other analyses to nonlinear relations and the specific combinations of circumstances hypothesized above to mediate this relation. Old data could be reexamined to test this hypothesis.

The fact remains, however, that a relation does exist between small samples and large predictive correlations in the total sample set ($r = -.56$), in recognition memory samples ($r = -.46$), and in habituation samples ($r = -.79$). What explains this consistent relation? The answer is not clear, but one possibility may lie in the interplay between statistical significance, *N*, and publishing practices. Specifically, correlations, like most other statistics, are more variable for smaller samples. Therefore, extremely high as well as extremely low correlations are more likely to be found in small than in large samples, but only extremely high correlations will be significant and therefore likely to get published. In contrast, smaller correlations are more likely to be significant in large samples and will be published. The total result is a negative relation between sample size and prediction level in the published literature. If this publishing bias has any explanatory power, it should apply to other behavioral domains as well, and its validity assumes that small data sets with nonsignificant *r*'s repose unpublished in the files of researchers.

5. *The level of predictions to childhood IQ is substantial given the reliability of the infant measures.* The short-term test-retest reliability of habituation and recognition memory scores has been found to be quite low (Bornstein, 1989; Bornstein & Sigman, 1986; Cohen, 1988; Fagan, 1984; Fagan &

McGrath, 1981; Fagan & Singer, 1983; Lecuyer, 1989; McCall, 1989; Slater, 1988). Reliabilities typically are between .30 and .45, varying with the test-retest interval and with the particular measure (Colombo et al., 1987, 1988). Note that the reliabilities are approximately the same or lower than the predictive correlations. It is technically possible to have a predictive correlation that is higher than the reliability of the predictor if the reliability of the outcome variable is very high, because the maximum predictive *r* is the square root of the product of the two reliabilities (Ghiselli, 1964). The year-to-year reliability of the childhood IQ test is approximately .85 (McCall, 1989), sufficiently high to allow predictions to be higher than the much lower reliability of the infant assessment.

Many explanations for the poor reliability of the infant measures have been offered. Younger infants, for example, may process a different aspect of the stimuli presented in these tasks than they do when older (Cohen, 1988). Therefore, reliability measured across age may be low because the habituation (or recognition memory) procedures are actually measuring the processing of different elements of a single physical stimulus at the two ages. One may find higher test-retest reliabilities across two ages in infancy using different stimuli than across the same two ages using the same stimuli (Cohen, 1988), if the two different stimuli are similarly matched to the infants' abilities at each age.

Alternatively, poor short-term reliability may be due to the structure of the habituation and recognition memory assessments (McCall, 1989). In general, relatively few habituation or recognition memory tasks are given to the subjects in a session (Bornstein & Sigman, 1986; McCall, 1989). Therefore, only a brief sampling is taken of the infant's behavior, which is likely to be unreliable. A single memory task may be analogous to a single item on a paper and pencil test, and the reliability of .30–.45 for these infant measures is roughly similar to the correlation between single items on paper and pencil psychometric tests for older children and adults. Therefore, the reliability of the habituation and recognition memory scores may be adequate, but the number of "items" usually assessed is psychometrically insufficient (for a general discussion of this issue, see Rushton, Brainerd, & Pressley, 1983).

This explanation receives only partial empirical support. For example, Rose, Feldman, and Wallace (1988) found a median interitem correlation (i.e., "alternative forms" reliability) for 12 recognition memory tasks given at 6, 7, and 8 months of age to be −.10, with the range of average intertask correlations to be −.16 to .13. The mean test-retest correlation for single tasks across the 1-month interval was .18, with a range of −.16 to .47. However, combining the 12 tasks into summary scores at each age increased the age-to-age reliability to .30 to .49. Colombo et al. (1988) also raised the reliability of novelty preference scores to .50 from .24 by combining tasks. However, while reliability is increased by adding more tasks, even having 12 tasks, which is more than in most studies, does not produce a very reliable measure.

The lack of reliability, as well as the low generality of habituation and recognition memory measures across tasks using different stimuli (Kagan, 1989) or different stimulus modalities (McCall, 1989), suggests that whatever is being measured by these assessments may not be a single process and/or may not be solely or even largely an endogenous process. Instead, it may be influenced substantially by particular stimuli and other exogenous aspects of the total assessment situation.

Nevertheless, long-term predictions are obtained, and their level is substantial, especially when the low reliability of the predictor is considered. For example, with reliabilities ranging between .30 and .45 for the infant measures and a conservative .85 for the childhood IQ test, the maximum possible prediction is approximately .50–.62. Observed weighted predictions average .36, which means that the predictions account for 34%–52% of the "reliable" variance in this assessment system. This figure is higher (e.g., 53%–81%) if the median raw prediction of .45 is used. It is in this relative sense, more than the absolute level of prediction (see next section), that the infant measures potentially reflect a powerful predictor of later mental performance.

It should be noted that the calculations in this section are based on classical true-score test theory, which assumes that a "true score" remains constant over settings, occasions, and age. This assumption may not be valid (e.g., Lumsden, 1976), especially for developmental phenomena, and other approaches, such as Cronbach's Generalizability Theory (Cronbach, Rajasatnam, & Glaser, 1963; Shavelson & Webb, 1981), may be more appropriate.

6. *Predictions to childhood IQ from habituation and recognition memory are consistently higher than for standardized infant tests of general development for nonrisk but not for risk samples, and they are not consistently higher than predicting from parental education and socioeconomic status or a few other infant behaviors for nonrisk samples.* The average weighted correlation of .36 and certainly the median raw correlation of .45 for habituation and recognition memory is notably higher than correlations from standardized tests of general infant development, which display a median (raw) value of .09 for nondisordered samples for predicting 5–7-year IQ from the first 6 months of life (Kopp & McCall, 1982). On the other hand, the correlation for standardized infant tests to later IQ is .54 for risk samples (Kopp & McCall, 1982), and parental education and other measures of socioeconomic status tend to correlate approximately .40–.60 with children's IQ depending on age and the prediction interval (McCall et al., 1972). Also, while not often measured or reported, race and birth order have sometimes predicted almost as well as recognition memory in samples containing severely disordered infants (Fagan & Singer, 1983). Finally, and more startling as well as forgotten, the early onset of vocalization during the first year of life was observed to correlate .71 with IQ at age 26 years (Cameron, Livson, & Bayley, 1967), a result also found (r = .50) for similar behaviors by Moore (1967). Curiously, this result has only been reported for females. Therefore, while predictions from early habituation and recognition memory to later IQ are consistent, exist for both sexes (although sexes are not always reported separately; see O'Connor, 1980, for a very large sex difference), and are higher than from standardized tests of general infant development in nonrisk samples, they are not necessarily higher than for infant tests for risk samples, not the highest predictors ever found, and certainly not the easiest to obtain.

This observation has several implications. First, from the standpoint of sheer prediction accuracy, which may be of practical importance in screening, selection of infants for intervention programs, and counseling of parents, habituation and recognition memory are still of limited utility. Second, the field has become preoccupied with habituation and recognition memory, forgetting that other behaviors, such as early vocalization, may also predict later IQ. Third, this observation highlights the proposition that the importance of the habituation/recognition memory prediction phenomenon lies more in what it may reveal about the process of mental development than in the fact of the prediction per se or its size. Unfortunately, nearly all the empirical effort has been expended demonstrating the prediction, which is the first step; but very little research has been directed at discerning the nature of the processes or mechanisms reponsible for the correlations (see below), which now should receive more emphasis.

Parameters

7. *Predictions are somewhat higher for risk than for nonrisk samples.* While this difference was observed statistically only for recognition memory samples (the averages of weighted r's were .52 vs. .27), the plot of prediction r's as a function of sample size (Fig. 4) clearly shows the correlations for risk samples to be uniformly higher within an N, and this is true for both paradigms. This result conforms to a similar finding for the prediction from standardized infant tests (Kopp & McCall, 1982). The presumption in both cases is that some low-scoring infants, likely those who are neurologically disordered *and* who are reared in intellectually unstimulating environments, score poorly as infants and as children, and such cases are more prevalent in risk samples and thus more likely to increase the size of the prediction for risk samples. This interpretation is consistent with the "Sameroff-Chandler lower-left quadrant hypothesis" described above as well as the notion that standardized infant tests predict better for poor-scoring than for average- or above-average-scoring infants (McCall, 1979).

8. *Predictions from habituation and recognition memory may be stronger when such assessments are made between 2 and 8 months of age than earlier or later.* The data for this proposition are fragmentary, involving only six of the 31 samples in Figure 5; and longitudinal data, while consistent with this theme, are also sparse (e.g., Cardon & Fulker, 1990; Rose et al., 1988). Nevertheless, the trend in Figure 5 is obvious, especially in contrast to the consistently high r's between 2 and 8 months, and is worthy of future study.

It is not unusual to find cross-age correlations to increase from birth to 2–3 months of age, as they do for standardized infant tests (McCall, 1979; McCall, Eichorn, & Hogarty, 1977). At the very least, the infant measures before 2 months may be less reliable than after 2 months. But the possible decline in predictability after 8 months, if substantiated by further study, would be more unusual and provocative. It would suggest, for example, that the simplest model of stability and continuity—that is, that the same behavior correlates consistently with itself across age—is not the case for this situation, and, indeed, two reports suggest that infant recognition memory predicts later IQ better than it predicts itself in childhood (Fagan, 1984; Slater, Cooper, Rose, & Morrison, 1989). This might imply that some qualitative transformation of the mediating mechanism occurs during development, in the manner theorized by Piaget and observed by McCall et al. (1977) for the same ages.

Alternatively, the mechanism may remain essentially the same, but the stimuli, behaviors, and measurement of it might change with development. For example, perhaps habituation and recognition memory to simple stimuli can be accomplished very quickly by the time the infant reaches approximately 8 months of age, and salient individual differences thereafter are not revealed unless the task is cognitively more difficult, presumably as in cross-modal transfer, which Gottfried et al. (1989) found to predict later IQ better than habituation at 12 months.

In any case, if these curious observations are supported by future research, something more complicated—and interesting—than simple stability in the same behavior may be involved in this prediction phenomenon.

9. *The level of prediction coefficients is remarkably consistent across the observed outcome age period of 2–8 years.* The marginal distribution along the ordinate of Figure 5 shows that the sizes of the *r*'s do not obviously vary with the age at which the childhood IQ test is given, and no statistically significant relation was found between weighted prediction *r*'s and age at outcome. What little longitudinal data are available (not reported here) also support the proposition of minimum effect for outcome age within studies (e.g., Fagan, 1984; Fagan & McGrath, 1981; Gottfried et al., 1989; Rose, Feldman, & Wallace, 1989; Slater et al., 1989), although some longitudinal studies (Colombo, Mitchell, Dodd, Coldern, & Horowitz, 1989; Thompson, Fagan, & Fulker, 1991) suggest that predictive *r*'s increase through the first 3 years.

The consistency of the predictive *r*'s between 2 and 8 years (and perhaps to age 12; Sigman, Cohen, Beckwith, Asarnow, & Parmelee, 1989) is unusual but not unprecedented in the longitudinal literature, and that exception is provocative. Specifically, the typical pattern is for longitudinal predictions to decline with increasing age at outcome, which is the case, for example, for predictions to later IQ from total scores on standardized infant tests (McCall, 1979). But this trend for total scores actually masks a pattern for subsets of items on these tests that mimics the persistent high predictions observed for habituation and recognition memory through the childhood years. Specifically, the first principal component at and following approximately 21 months of age predicts at very high levels IQ assessments given throughout childhood and adolescence (McCall et al., 1977).

The parallel observations that habituation and recognition memory in the first year of life and consensual vocabulary and symbolic thought in the second year both predict later IQ between 2 and 8 years and do so at the same, undecreasing level regardless of the age of the child at outcome assessment might be a homologous coincidence. But recent studies have suggested that habituation and recognition memory predict certain language and memory functions in young children (Bornstein & Sigman, 1986; Colombo et al., 1989; Fagan & Knevel, 1989; Rose, Feldman, Wallace, & McCarton, 1989), even when IQ is covaried from such skills (Thompson et al., 1991). Could habituation and recognition memory in the first year, consensual vocabulary and early symbolic functions in the second and third years, and IQ, verbal skill, and certain memory functions in childhood all be threads of the same developing mental fabric?

Conceptual Mechanism

The overt behaviors that correlate across development, such as those mentioned immediately above, are clues to the underlying mechanisms that mediate the predictions. While a few have wondered if habituation, at least, required any serious cognitive processes (e.g., Lecuyer, 1988, 1989; Malcuit et al., 1988), most students of the field (e.g., see Kuhn, 1989) have argued that performance in these paradigms reflects "information processing," that is, the ability to encode the familiar stimulus, remember it, compare a presented stimulus with the remembered engram, recognize the familiar stimulus when it is represented, discriminate a new stimulus from the familiar or its engram, and encode the new stimulus into memory (e.g., Bornstein, 1985; Bornstein & Sigman, 1986; Colombo & Mitchell, 1988, 1991; Fagan, 1988; Fagan & Singer, 1983; Lewis et al., 1969). Presumably, those infants who perform these tasks most rapidly turn out to have higher IQs, and that relative performance on these processes is stable from infancy to childhood.

Without question, accomplishing these tasks is required in the habituation or recognition memory paradigms, but it is not necessarily the case that such processes are sufficient to make the prediction or that the speed of their execution is the primary component of the predicting measures. McCall and Carriger (1991) make the case that they are not. While they acknowledge these information processing tasks must occur, they speculate that these processes are conducted very rapidly, and variance in the predicting measures reflects in substantial part another disposition—the disposition to *inhibit* responding to familiar stimuli and to stimuli of minor prominence (e.g., low energy, static, etc.). Similarly, Dempster (1991) recently argued that mature intelligence is typically discussed in terms of speed of information processing, the quality or quantity of information represented, executive processes, and processing capacity, but that it cannot be understood without reference to inhibitory processes, which have been largely ignored.

Whatever the mechanism, it seems clear that the prediction phenomenon has been established, and that much future research, including reanalyses of existing data, should be directed at crafting tasks and measures that differentiate these several skills during infancy and early childhood and discerning how they become woven into the fabric of mature intelligence.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Bornstein, M. H. (1984, April). *Infant attention and care-giver stimulation: Two contributions to early cognitive development.* Paper presented at the International Conference on Infant Studies, New York.

Bornstein, M. H. (1985). Habituation of attention as a measure of visual information processing in human infants: Summary, systematization, and synthesis. In G. Gottlieb & N. H. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 253–300). Norwood, NJ: Ablex.

Bornstein, M. H. (1989). Stability in early mental development: From attention and information processing in infancy to language and cognition in childhood. In M. H. Bornstein & N. A. Kresnegor (Eds.), *Stability and continuity in mental development* (pp. 197–170). Hillsdale, NJ: Erlbaum.

Bornstein, M. H., & Sigman, M. D. (1986). Continuity in mental development from infancy. *Child Development,* **57,** 251–274.

Cameron, J., Livson, N., & Bayley, N. (1967). Infant vocalizations and their relationship to mature intelligence. *Science,* **157,** 331–333.

Cardon, L. R., & Fulker, D. W. (1990). *Sources of continuity in infant predictors of later IQ.* Unpublished manuscript, authors.

Caron, A. J., Caron, R. F., & Glass, P. (1983). Responsiveness to relational information as a measure of cognitive functioning in nonsuspect infants. In T. Field & A. Sostek (Eds.), *Infants born at-risk: Physiological, perceptual, and cognitive processes* (pp. 181–209). New York: Grune & Stratton.

Cohen, L. B. (1988). The relationship between infant habituation and infant information processing. *European Bulletin of Cognitive Psychology,* **8,** 445–454.

Cohen, S. E., & Parmelee, A. H. (1983). Prediction of five-year Stanford-Binet scores in preterm infants. *Child Development,* **54,** 1242–1253.

Colombo, J., & Mitchell, D. W. (1988). Infant visual habituation: In defense of an information-processing analysis. *European Bulletin of Cognitive Psychology,* **8,** 455–461.

Colombo, J. C., & Mitchell, D. W. (1991). Individual differences in early visual attention: Fixation time and information processing. In J. Colombo & J. W. Fagen (Eds.), *Individual differences in infancy: Reliability, stability, and prediction* (pp. 193–227). Hillsdale , NJ: Erlbaum.

Colombo, J., & Mitchell, D. W., Dodd, J., Coldren, J. T., & Horowitz, F. D. (1989). Longitudinal correlates of infant attention in the paired-comparison paradigm. *Intelligence,* **13,** 33–42.

Colombo, J., Mitchell, D. W., & Horowitz, F. D. (1988). Infant visual attention in the paired-comparison paradigm: Test-retest and attention-performance relations. *Child Development,* **59,** 1198–1210.

Colombo, J., Mitchell, D. W., O'Brien, M., & Horowitz, F. D. (1987). The stability of visual habituation during the first year. *Child Development,* **58,** 474–487.

Cronbach, L. J., Rajasatnam, N., & Glaser, B. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology,* **16,** 137–163.

Dempster, F. N. (1991). Inhibitory processes: A neglected dimension of intelligence. *Intelligence,* **15,** 157–173.

DiLalla, D. L., & Fulker, D. W. (1989). Infant measures as predictors of later IQ: The Twin Infant Project (TIP). *Behavior Genetics,* **19,** 753–754.

Fagan, J. F. (1984). The relationship of novelty preferences during infancy to later intelligence and later recognition memory. *Intelligence,* **8,** 339–346.

Fagan, J. F. (1988). Evidence for the relationship between responsiveness to visual novelty during infancy and later intelligence. A summary. *European Bulletin of Cognitive Psychology,* **8,** 469–475.

Fagan, J.F., & Knevel, C. (1989, April). *The prediction of above-average intelligence from infancy.* Paper presented at the meeting of the Society for Research in Child Development, Kansas City.

Fagan, J. F., & McGrath, S. K. (1981). Infant recognition memory and later intelligence. *Intelligence,* **5,** 121–130.

Fagan, J. F., & Singer, L. T. (1983). Infant recognition memory as a measure of intelligence. In L. P. Lipsitt & C. K. Rovee-Collier (Eds.), *Advances in infancy research* (Vol. **2,** pp. 31–79). Norwood, NJ: Ablex.

Fulker, D. W., Plomin, R., Thompson, L. A., Phillips, K., DiLalla, L. F., Fagan, J. F., & Haith, M. M. (1988). *Rapid-screening of infant predictors of adult IQ: A study of infant twins and their parents.* Unpublished, authors.

Ghiselli, E. E. (1964). *Theory of psychological measurement.* New York: McGraw-Hill.

Gottfried, A. W., Guerin, D., & Bathurst, K. (1989). *Infant predictors of intelligence and achievement: A comparative analysis of sensorimotor, environmental, and recognition memory measures.* Unpublished, authors.

Hedges, L. L., & Oklin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hunt, J. McV. (1961). *Intelligence and experience.* New York: Ronald.

Kagan, J. (1989). Commentary: Does infant boredom reflect intelligence? *Human Development,* **32,** 172–176.

Kopp, C. B., & McCall, R. B. (1982). Predicting later mental performance for normal, at-risk, and handicapped infants. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. **4,** pp. 33–61). New York: Academic Press.

Kuhn, D. (Ed.). (1989). Special issue. *Human Development,* **32,** 177–186.

Lecuyer, R. (1988). Please infant, can you tell me exactly what you are doing during a habituation experiment? *European Bulletin of Cognitive Psychology,* **8,** 476–480.

Lecuyer, R. (1989). Habituation and attention, novelty and cognition: Where is the continuity? *Human Development,* **32,** 148–157.

Lewis, M., & Brooks-Gunn, J. (1981). Visual attention at three months as a predictor of cognitive functioning at two years of age. *Intelligence,* **5,** 131–140.

Lewis, M., Goldberg, S., & Campbell, H. (1969). A developmental study of information processing within the first three years of life: Response decrement to a redundant signal. *Monograph of the Society for Research in Child Development,* **34**(9, Serial No. 133).

Lumsden, J. (1976). Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. **27,** pp. 251–280). Palo Alto, CA: Annual Reviews.

Malcuit, G., Pomerleau, A., & Lamarre, G. (1988). Habituation, visual fixation, and cognitive activity in infants: A critical analysis and attempt at a new formulation. *European Bulletin of Cognitive Psychology,* **8,** 415–440.

McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. In J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 707–740). New York: Wiley.

McCall, R. B. (1981). Early predictors of later IQ: The search continues. *Intelligence,* **5,** 141–147.

McCall, R. B. (1989). Commentary: Issues in predicting later IQ from infant habituation rate and recognition memory performance. *Human Development,* **32,** 177–186.

McCall, R. B., & Carriger, M. S. (1991). *Infant habituation and recognition memory performance as prediction of later IQ: A review and conceptual analysis.* Unpublished paper, authors.

McCall, R. B., Eichorn, D. H., & Hogarty, P. S. (1977). Transitions in early mental development. *Monographs of the Society for Research in Child Development,* **42**(3, Serial No. 171).

McCall, R. B., Hogarty, P. S., & Hurlburt, N. (1972). Transitions in infant sensorimotor development and the prediction of childhood IQ. *American Psychologist,* **27,** 728–748.

Miller, D. J., Ryan, E. B., Aberger, E., McGuire,

M. D., Short, E. J., & Kenny, D. A. (1979). Relationships between assessments of habituation and cognitive performance in the early years of life. *International Journal of Behavioral Development,* **2,** 159–170.

Moore, T. (1967). Language and intelligence: A longitudinal study of the first eight years: I. Patterns of development in boys and girls. *Human Development,* **10,** 88–106.

O'Connor, M. J. (1980). A comparison of pre-term and full-term infants on auditory discrimination at four months and on Bayley Scales of Infant Development at eighteen months. *Child Development,* **51,** 81–88.

O'Connor, M. J., Cohen, S. L., & Parmelee, A. H. (1984). Infant auditory discrimination in pre-term and full-term infants as a predictor of 5-year intelligence. *Developmental Psychology,* **20,** 159–165.

Rose, D. H., Slater, A., & Perry, H. (1986). Prediction of childhood intelligence from habituation in early infancy. *Intelligence,* **10,** 251–263.

Rose, S. A. (1989). Measuring infant intelligence: New perspectives. In M. H. Bornstein & N. A. Krasnegor (Eds.), *Stability and continuity in mental development* (pp. 171–188). Hillsdale, NJ: Erlbaum.

Rose, S. A., & Feldman, J. F. (1991). Infant cognition: Individual differences and developmental continuities. In J. Colombo & J. W. Fagen (Eds.), *Individual differences in infancy: Reliability, stability, and prediction* (pp. 229–246). Hillsdale, NJ: Erlbaum.

Rose, S. A., Feldman, J. F., & Wallace, I. F. (1988). Individual differences in infants' information processing: Reliability, stability, and prediction. *Child Development,* **59,** 1177–1197.

Rose, S. A., Feldman, J. F., & Wallace, I. F. (1989). *Language: A partial link between infant attention and later intelligence.* Unpublished manuscript, authors.

Rose, S. A., Feldman, J. F., Wallace, I. F., & McCarton, C. (1989). Infant visual attention: Relation to birth status and developmental outcome during the first five years. *Developmental Psychology,* **25,** 560–576.

Rose, S. A., & Wallace, I. F. (1985). Visual recognition memory: A predictor of later cognitive functioning in pre-terms. *Child Development,* **56,** 843–852.

Ruddy, M. G., & Bornstein, M. H. (1982). Cognitive correlates of infant attention and maternal stimulation over the first year of life. *Child Development,* **53,** 183–188.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principal of aggregation. *Psychological Bulletin,* **94,** 18–38.

Sameroff, A. J., & Chandler, M. J. (1975). Reproductive risk and the continuum of caretaking casualty. In F. D. Horowitz (Ed.), *Review of child development research* (Vol. 4, pp. 187–244). Chicago: University of Chicago Press.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology,* **34,** 133–166.

Sigman, M., Cohen, S. E., Beckwith, L., Asarnow, R., & Parmelee, A. H. (1989, April). *Infant attention and sleep organization and cognitive skills in early adolescence.* Paper presented at the meeting of the Society for Research in Child Development, Kansas City.

Sigman, M., Cohen, S. E., Beckwith, L., & Parmelee, A. H. (1986). Infant attention in relation to intellectual abilities in childhood. *Developmental Psychology,* **23,** 788–792.

Slater, A. (1988). Habituation and visual fixation in infants: Information processing, reinforcement, and what else? *European Bulletin of Cognitive Psychology,* **8,** 517–523.

Slater, A., Cooper, R., Rose, D., & Morrison, V. (1989). Prediction of cognitive performance from infancy to early childhood. *Human Development,* **32,** 137–147.

Thompson, L. A., Fagan, J. F., & Fulker, D. W. (1991). Longitudinal prediction of specific cognitive abilities from infant novelty preference. *Child Development,* **62,** 530–538.

Turkheimer, E., & Gottesman, I. I. (1991). Individual differences and the canalization of human behavior. *Developmental Psychology,* **27,** 18–22.

Yarrow, L. J., Klein, R. P., Lomonaco, S., & Morgan, G. A. (1975). Cognitive and motivational development in early childhood. In B. X. Friedlander, G. M. Sterritt, & G. E. Kirh (Eds.), *Exceptional infant* (Vol. 3, pp. 491–502). New York: Brunner/Mazel.