Testing Infant Intelligence

Lewis and McGurk (*1*) have presented evidence to persuade us that infant testing is of doubtful validity. Because of the implications of their article, which is likely to be read by persons unfamiliar with infant testing, some qualifications are in order.

Lewis and McGurk establish the basis of their criticisms of infant testing by providing part of a sentence written by Bayley (*2*, p. 1174), which I quote in its entirety: "The findings of these early studies of mental growth of infants have been repeated sufficiently often so that it is now well established that test scores earned in the first year or two have relatively little predictive validity (in contrast to tests at school age or later), although they may have high validity as measures of the children's cognitive abilities at the time." Even this statement should be viewed in conjunction with another statement Bayley and her colleagues made regarding "results which strongly suggest that developmental psychologists need to rethink their previous conclusion that infant developmental test scores are unrelated to later measures of intelligence" (*3*, p. 332).

Because giving infant tests once to normal infants produces scores that correlate poorly with measures of future intelligence, Lewis and McGurk state that, in clinical populations, "use of infant intelligence scales is justified only if, in interpreting the resultant scores, the scores are regarded solely as measures of present performance and not as indices of future potential" (*1*, p. 1175). Yet it is precisely in such populations that one finds infant intelligence scores most useful for diagnostic appraisals and most valid as measures of future intelligence. Whether the actual scores from the tests are used or are categorized into ranges, there are numerous studies on the clinical evaluation of infants that show respectable, and even robust, interage correlations (*4*). These correlations can be improved by combining sex differences, neurological status, and test-taking behavior with the test scores. Moreover, since intervention programs are often directed toward populations that contribute proportionately more infants "at risk" for future developmental disabilities, one would very much want to apply some type of wide-ranging infant scales to assess the effects of intervention. Although intervention programs might be satisfied with the improvement of specific skills, such as the ability to play peek-a-boo or to smile, some effort should be made to ascertain whether more generalizable abilities have improved. Unfortunately, the constant reminder that investigators should search for social relevance may eventually lead to assessments of little at all—a trend exemplified by Lewis' own statement that "we are doomed in our evaluation of intervention programs if we ask for the presence or absence of whatever we are looking at" (*5*, p. 4).

There is a need for better measures of infants' behavior that can help us to predict their future abilities. For that reason alone, we should continue to test babies. Until those measures are obtained, however, we should not casually dismiss the usefulness of the ones we have. A leading question like "Infant intelligence scores—true or false?" and inadequate data purportedly answering that question cannot help much.

ADAM P. MATHENY, JR.
Child Development Unit, University of Louisville, Louisville, Kentucky 40202

References

1. M. Lewis and H. McGurk, *Science* **178**, 1174 (1972).
2. N. Bayley, in *Carmichael's Manual of Child Psychology*, P. H. Mussen, Ed. (Wiley, New York, 1970), pp. 1163–1209.
3. J. Cameron, N. Livson, N. Bayley, *Science* **157**, 331 (1967).
4. H. Knobloch and B. Pasamanick, in *Child Development and Child Psychiatry*, C. Shagrass and B. Pasamanick, Eds. (Psychiatric Research Report No. 13, American Psychiatric Association, Washington, D.C., 1960), pp. 10–31; R. Illingworth, *J. Child Psychol. Psychiat.* **2**, 210 (196➡ E. Werner, M. Honzik, R. Smith, *Child Develop.* **39**, 1063 (1968); M. Erickson, *Amer. J. Ment. Defic.* **72**, 728 (1968); A. Simon and L. Bass, *Amer. J. Orthopsychiat.* **26**, 340 (1956); A. Matheny, *Amer. J. Ment. Defic.* **71**, 371 (1966).
5. M. Lewis, in *Studies in Socio-emotional Development in Infancy*, I. Gordon, Ed. (Selected Documents in Psychology No. 210, American Psychological Association, Washington, D.C., 1972), pp. 3–9.

26 January 1973; revised 18 May 1973

The article by Lewis and McGurk (*1*) on evaluation of infant intelligence brings to mind the advertising adage that it is the packaging, not the product itself, that swings the market. In this case, the packaging is supplied by the joint themes of social relevance and intervention programs, both of which have high market appeal at this point, while the product is a collection of test data purportedly measuring infant intelligence. The results drawn from these data are used to support a conclusion that infant mental development lacks continuity and reveals no systematic age-to-age relationship from 3 to 24 months. The authors draw additional conclusions about the lack of predictive validity of infant intelligence scales and the inadequacy of these scales to reflect any presumed changes in intellectual ability that might arise from intervention programs. These are wide-ranging negative conclusions, and since Lewis and McGurk present their data as proof of these negative conclusions, the data require close scrutiny.

The quoted material, which appears to confirm these negative relationships, also demands a careful appraisal because the selections give a distorted view of the actual data and conclusions in the original sources. In particular, the careful and thorough work done by Bayley, culminating in the first really well-constructed and well-standardized test of infant mental development (*2*), is simply mocked by the patchwork of results and conclusions published by Lewis and McGurk. The nature of infant mental development—and, indeed, the design and evaluation of intervention programs—is much too important an issue to be dealt with as a vehicle for preformed opinions.

Lewis and McGurk begin by selecting an archaic and quaintly phrased definition of intelligence from Burt (*3*), which they find defective for its implication that intelligence is not subject to qualitative change or to environmental influence. A more appropriate and recent selection from Burt's work might have been chosen from his final paper (*4*), in which he synthesized the results of his entire research career concerning the determinants of mental development. His conclusions follow (*4*, p. 188):

> The hypothesis of a general factor entering into every type of cognitive process . . . is fully borne out by the statistical evidence; and the contention that differences in this general factor depend largely on the individual's genetic constitution appears incontestable.
>
> From this it is tempting to infer that each individual's innate capacity sets a fixed upper limit to what in actual practice he is likely to achieve under existing conditions. [However] A given genetic endowment is compatible with a whole range of developmental reactions and consequently of acquired attainments. All that a knowledge of a child's genetic endowment permits us to infer are the limits of that range, where "limit" is defined in terms of probability.

Lewis and McGurk then turn to Bayley's 1933 paper, citing a series of zero-order correlations between developmental scores obtained at 1 to 3 months of age and those obtained at 18 to 36 months. The correlations are introduced as evidence that "high predictive validity from one age to another . . . is singularly lacking in every scale used to assess intelligence during early infancy" (*1*, p. 1174). Aside from the fact that the infant tests used in that early study were hardly the equal of the recently published *Bayley Scales of Infant Development* (*2*), the authors might have referred to Bayley's 1949 article (*5*), in which the more extensive age-to-age correlations from the Berkeley Growth Study were reported. During infancy, the correlations between ages followed a coherent simplex pattern, and within the second year (13 to 18 to 24 months), the correlations ranged from .47 to .60. This hardly justifies a conclusion that predictive validity is singularly lacking. Developmental status in the first 3 months of life has very low predictive power for developmental scores in the second and third years, but it is misleading to cite this particular zero-order correlation as typical for all other ages of infancy.

Bayley's correlations were obtained for a small, homogeneous sample, but they have been confirmed recently in a much larger sample, with 200 or more cases in each correlation (*6*). The between-age correlations for the latter sample are shown in Table 1. The correlations are moderate, but they are neither insignificant nor without a coherent simplex pattern. The values are highest for adjacent ages and decline as the age span increases. Further, there is other evidence from the twin data showing that the age-to-age changes in developmental status that suppress these correlations are themselves systematic in origin. The changes are brought about by a genetic spurt-lag factor that may cause an infant to be precocious at one age and average at the next. Once this factor is fully understood, in terms of how it relates to the changing capabilities being measured between 3 and 24 months, the age-to-age progression in mental development will reveal even greater consistency and coherence than implied by the moderate correlations in Table 1.

Lewis and McGurk turn to the long-range predictive power of infant intelligence scales and stress repeatedly that the scales are invalid as measures of future potential. This statement is clearly inaccurate for cases of clinical retardation, and it becomes progressively less accurate for normal infants as they reach their second birthday. In Bayley's longitudinal study, the predictive power of each infant test was appraised by correlating it with later measures of intelligence up to age 18, and, as she remarks, "By two years the *r*'s with tests at later ages hold up fairly well, rarely dropping below .50" (*5*, p. 185). Data from the Fels longitudinal study showed a comparable pattern; the correlations between 2-year Gesell scores and 10-year Binet IQ scores exceeded .50 for both boys and girls and were substantially higher for girls at the earlier Binet ages (*7*). Considering the age span involved, the predictive correlations from both studies are substantial and do not lend themselves to casual dismissal.

Table 1. Intercorrelations between Bayley MDI scores at each age level. [$N \geq 200$; see (*6*) for full description.]

Age (months)	Age (months)				
	6	9	12	18	24
3	.53	.39	.29	.08	.23
6		.45	.35	.27	.28
9			.52	.37	.32
12				.41	.30
18					.42

Incidentally, Lewis and McGurk are extremely selective in choosing the results from Bayley to support their contention of no predictive validity. They assert that the use of "infant intelligence scales is justified only if . . . the scores are regarded solely as measures of present performance and not as indices of future potential. What this [present] performance may mean is questionable, since it is possible that 'superior' performance may be indicative of poor performance later. For example, Bayley shows a correlation of $-.30$ between males' earlier test behavior and their IQ . . . at ages 16 to 18" (*1*, p. 1175).

I am unable to find this correlation in the reference cited by Lewis and McGurk (*8*). It does appear in another, uncited article (*9*) as the correlation between the composite 4-, 5-, and 6-month developmental score and the 16- to 18-year IQ score for males. It happens to be the most extreme negative correlation in the entire series (although statistically nonsignificant), and perhaps the extent of the authors' selectivity may be appreciated by noting that the predictive correlation for females at the next developmental stage (7, 8, and 9 months) is both positive and larger ($r = .40$) than the cited correlation. Is this a less important finding? The fact is that measures of developmental status during the first year do not give consistent predictions of school-age or adult IQ. But citing only the largest negative correlation does violence to the original data and creates unjustified doubts about the validity of the infant scales at any age.

Lewis and McGurk then comment on the use of infant intelligence test scores as a criterion for evaluating intervention programs. Their data are subsequently presented as a test of whether infant intelligence scales are capable of reflecting "any improvement in competence that results from a specific enrichment experience" (*1*, p. 1175).

Since their article includes no report of an intervention program actually being conducted, it is presumptuous to assume that their data furnish a test of this question. The more critical issue, however, in evaluating the Lewis-McGurk data is whether the results accurately reflect the course of infant mental development during the first 24 months of life. This is particularly important because their results are largely negative, and as a consequence they conclude that "there is no reliable relation between successive measures of infant intelligence during the first 24 months of life" (*1*, p. 1176). Later, they add, "The implications . . . seem clear. Simply stated, infant intelligence scales are unsuitable instruments for assessing the effects of specific intervention procedures" (*1*, p. 1176).

What characteristics of the sample and the test scores can be inferred from the information supplied by Lewis and McGurk? The sample is reported to include approximately 20 infants, with approximately equal number of males and females, although a footnote advises that not all the infants were able to complete all the tests. The sample itself is very small (the Bayley standardization sample, in contrast, included more than 85 infants at each age), and there are conflicting reports about its composition. At one point, Lewis and McGurk state that "the sample was heterogeneous with

respect to social class, although it was slightly skewed toward the upper-middle classes" (*1*, p. 1175). In the following paragraph, however, they attribute the elevated Mental Development Index (MDI) scores for their infants to "the relatively high socioeconomic composition of our sample" (*1*, p. 1175). Aside from this ambiguity, the study begins with a serious deficiency in sample size; and since infant testing is unusually demanding in terms of obtaining valid scores, any error in testing or scoring will have a disproportionate effect on the scores for such a small sample.

The MDI means and standard deviations (S.D.'s) reported by Lewis and McGurk present a curious pattern—the means extend from 101.6, at age 3 months, to 126.4, at age 24 months; the standard deviations follow an inconsistent pattern, ranging from 11.6 to 20.6. By comparison with the standardization sample ($\bar{X} = 100$; S.D. = 16), the Lewis and McGurk infants are equal in developmental status at 3 months, but by 18 months they have moved up to the 80th percentile, and at 24 months their *mean* developmental score has jumped to the 95th percentile of the standardization group.

Such a trend is extremely deviant if the sample is anywhere near being representative. And while Lewis and McGurk attribute the elevated scores to the high socioeconomic composition of their sample, there is no research evidence to support this view. Other studies have found a zero-order correlation between socioeconomic status and first-year MDI scores; and even at 24 months the correlation is so low ($r = .23$) that it could not affect the scores of upper-class infants by more than 5 points (*6, 9, 10*).

Further, after Lewis and McGurk present their longitudinal data, they add in a reference that "Data from an additional 120 infants, seen cross-sectionally at 3, 6, 9, 12, 18, and 24 months (20 infants at each age), were essentially similar to those reported here" [reference 7 in (*1*)]. If this is the case, with cross-sectional data showing the same peculiarities as longitudinal data, then there is no explanation other than unreliability in test administration and scoring. The haphazard pattern of correlations between ages would follow as a natural consequence of unreliable data, and the fact that most of the correlations are insignificant and conform to no evident pattern is more of a statement about this set of data than about the nature of infant mental development.

The results reported for the object permanence scale of the Escalona-Corman Scales of Sensori-Motor Development display an unusually restricted range of scores, and the scoring on this scale is quite different from the standardized scoring on the MDI. Each infant appears to have been given a score equal to the number of items passed, and in view of the very narrow standard deviations at certain ages, it would be helpful to have the actual frequency distribution of scores at each age. It is inappropriate to compute or try to interpret correlations for very restricted distributions. The data simply are not strong enough to support inferences about the presence (or absence) of relationships between MDI scores and object permanence scores.

By way of final assessment, the data reported by Lewis and McGurk are so seriously lacking in measures of internal consistency that the absence of significant and coherent relationships must be regarded as a result of unreliable data. Yet the data have been used to support sweeping generalizations about the inconsistent nature of infant mental development and the lack of utility for the infant scales. The authors lend an air of authenticity to their data by stating that "our data tend to support the view, advanced by Bayley, that at each stage of development intelligence comprises a set of relatively discrete abilities, or factors" (*1*, p. 1176). But the Lewis-McGurk data are so erratic and inconsistent that they cannot be said to support any particular conception of infant mental development. Bayley's interpretation was built on data showing moderate correlations at adjacent ages but declining as the age span increased—the simplex pattern. As Lewis and McGurk note, however, this pattern is absent from their data: "Moreover, the data fail to reveal either simplex or other patterns of correlation" (*1*, p. 1175). Their appraisal is certainly correct—for example, the 6-month MDI scores did not correlate with the 9-month MDI scores ($r = .08$), but then correlated significantly with the 24-month MDI scores ($r = .54$) [table 3 in (*1*)]. No other investigator has reported such a pattern, which amounts to a reverse simplex, and it is hard to imagine what conception of infant mental development would follow from such results. Even this anomalous reversal was not consistently duplicated at other ages.

However, Lewis and McGurk then use these data to dismiss infant intelligence scales as unsuitable for measuring the effectiveness of intervention programs. A full discussion of the deficiencies in their approach to this topic would require a separate technical comment. In brief, a collection of negative results such as these cannot be used to discredit the Bayley scale as a measure of infant mental development, whether enhanced by intervention or not.

Nor can these negative results weigh in favor of an alternative conclusion—namely, that intervention programs do improve intellectual ability but the scale is unable to detect the improvement. Data unreliability aside, the Lewis-McGurk results do not begin to reach this issue, since no intervention program was conducted, and there is a sense of a conclusion prepared in advance when the authors write, "Even more serious is the possibility that, by using the wrong instrument of evaluation [a standard infant intelligence scale] in a large number of programs, one would erroneously conclude that intervention in general is ineffective in improving intellectual ability, thereby supporting the view that environment is ineffective in modifying intelligence" (*1*, p. 1177). There is clearly an implicit assumption here that intervention programs will be effective, and any failure to find supporting evidence is the fault of the scale, not of the program or the original assumption. What better way to make this conclusion seem plausible than to publish a collection of test results that stigmatize the scale as inadequate?

However, my comment is principally concerned with inadequacies in the data and bias in the selection of supporting references, which not only annul the authors' conclusions, but also discredit the careful work done on infant mental development by Bayley and her colleagues. The issues here are too important to be evaluated with poor data that have been given slick packaging in a theme of social relevance.

RONALD S. WILSON
Child Development Unit,
University of Louisville,
Louisville, Kentucky 40202

References

1. M. Lewis and H. McGurk, *Science* **178**, 1174 (1972).
2. N. Bayley, *Bayley Scales of Infant Development* (Psychological Corporation, New York, 1969).
3. C. Burt, E. Jones, E. Miller, W. Moodie, *How the Mind Works* (Appleton-Century-Crofts, New York, 1934).
4. C. Burt, *Amer. Psychol.* **27**, 175 (1972).
5. N. Bayley, *J. Genet. Psychol.* **75**, 165 (1949).
6. R. S. Wilson and E. B. Harpring, *Develop. Psychol.* **7**, 277 (1972).
7. R. B. McCall, P. S. Hogarty, N. Hurlburt, *Amer. Psychol.* **27**, 728 (1972).
8. N. Bayley, *ibid.* **10**, 805 (1955).
9. ——— and E. S. Schaefer, *Monogr. Soc. Res. Child Develop. No. 29* (1964), ser. 97.
10. N. Bayley, *Child Develop.* **36**, 379 (1965).

26 February 1973; revised 5 June 1973

Both Matheny and Wilson seem to feel that (i) our data (*1*) represent some unique result in the study of infant "intelligence tests" and that (ii) we were out to get Bayley and her scales.

In fact, our results were nothing more than another example of a now frequently reported finding. Wilson, in a rather excitable tone, accuses us of misrepresenting Bayley. It was ungracious to impute to us any such motivation.

In a 1955 article, "On the growth of intelligence," she reported an attempt "to find predictive items from the first year scale on the Berkeley Growth Study children" (*2*, p. 805). She was able to select 31 items that distinguish a group of the "6 children at each extreme of intelligence as measured at the 14 to 16 year tests." Using these items, she studied the relationship between the scores obtained at 6, 9, and 12 months and the intelligence sigma scores at ages 16, 17, and 18 years. "We were unable to get significant correlations even though our sample was composed in large part of the cases on whom the items were selected, including all of the extreme cases that would determine a relationship" (*2*, p. 807). And finally, "So far, none of these efforts has been successful in devising an intelligence scale applicable to children *under two years that will predict their later performance* [italics added]" (*2*, p. 807).

Thus it would appear that, rather than misrepresenting Bayley, we were in total agreement. Let us quote a later section (*2*, p. 807):

> These findings give little hope of our being able to measure a stable and predictable intellectual factor in the very young. I am inclined to think that the major reason for this failure rests in the nature of intelligence itself. 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 thus agrees with our earlier statements, which argued that intelligence is "not a general, unitary trait, but is, rather, a composite of skills and abilities that are not necessarily covariant" (*1*, p. 1176).

One might argue that it is unfair to quote Bayley's 1955 article. In a review published in 1970, "Development of mental abilities" (*3*), Bayley refers to her study of 1933 (*3*, p. 1171):

> Tests were given at regular intervals, monthly 1 to 15 months then tri-monthly to 3 years. Their scores during the first 6 months were entirely independent of scores at 2 or 3 years, and even at 10 to 12 months the correlation with 3-year intelligence was only .45. Since that time other studies (Honzik, Macfarlane, and Allen, 1948; Hindley, 1960; P. Cattell, 1940) repeatedly show that scores on tests during the *first 2 years are correlated very little or not at all with scores earned at 4 years or later. These findings raised a series of questions about the reliability and validity of infant tests and about the nature of the developing mental functions in infancy and early childhood* (Stott and Ball, 1965) [italics added].

And finally (*3*, p. 1174):

> The findings of these early studies of mental growth of infants have been repeated sufficiently often so that it is now *well-established* that test scores earned in the first year or two have relatively little predictive validity (in contrast to tests at school age or later), although they may have high validity as measures of the children's cognitive abilities at the time [italics added].

Bayley did not consider this failure to be due to unreliability of the measuring instrument, nor do we: "Thus the lack of stability in the first 3 years cannot be attributed to poor reliability of the measuring instrument" (*3*, p. 1174).

So much for the implication that we misrepresented Bayley. In fact, Bayley's conclusion is supported, as is ours, by a whole set of other reports and reviews (*4*). Indeed, Wilson's own work, as reported above, could be construed to support the general finding. His best across-age consistency (which happens to be from one age to the next) accounts for only 17 to 28 percent of the variance between scores—hardly a basis for much predictive strength.

Finally, let us quote from the most recent of the reports on the relationship of early intelligence test scores and later performance on IQ tests. After reviewing data obtained from the Fels longitudinal study, McCall, Hogarty, and Hurlburt conclude (*5*, p. 746):

> The overriding implication of this discussion is that a simple conception of a constant and pervasive g factor is probably not tenable as a model for "mental" development, especially for the infancy period. *The data are strong in their denial of simple continuity of general precocity at one age with general precocity at another age during the infancy period*, and emphatic in demonstrating marked qualitative shifts in behavioral dispositions [italics added].

We therefore must maintain our original position and argue for the failure of present "tests of infant intelligence" to have any predictive power. Whether this is due to the nature of intelligence (a view we hold) or a measurement problem has not been determined.

Given this overwhelming evidence, it is not unreasonable to question the uses to which the Bayley, or any other infant intelligence test, is put. In so questioning we hope to alert the community in general to the possible risks of using such instruments. It remains to be demonstrated that, while these tests have no predictive ability, they *may* reflect the infant's current mental capacity. No such proof has been offered, and until such evidence is produced we must continue to question the uses and misuses of infant tests of intelligence.

MICHAEL LEWIS
Infant Laboratory,
Division of Psychological Studies,
Educational Testing Service,
Princeton, New Jersey 08540

HARRY MCGURK
University of Surrey, Surrey, England

References and Notes

1. M. Lewis and H. McGurk, *Science* **178**, 1174 (1972).
2. N. Bayley, *Amer. Psychol.* **10**, 805 (1955).
3. ———, in *Carmichael's Manual of Child Psychology*, P. H. Mussen, Ed. (Wiley, New York, 1970), pp. 1163–1209.
4. L. H. Stott and R. S. Ball, "Infant and preschool mental tests: Review and evaluation," *Monogr. Soc. Res. Child Develop. No. 30* (1965), ser. 3; H. Thomas, *Merrill-Palmer Quart. Behav. Develop.* **16** (No. 2), 179 (1970); I. C. Uzgiris, *Merrill-Palmer Quart.* **19**, 181 (1973).
5. R. B. McCall, P. S. Hogarty, N. Hurlburt, *Amer. Psychol.* **27**, 728 (1972).

6. This work was supported by National Science Foundation grant 28105.

23 September 1973