

# Studies regarding the impact of micronutrient status on behavior in man: iron deficiency as a model<sup>1-3</sup>

Rudolph L. Leibel, M.D., Ernesto Pollitt, Ph.D., Insun Kim, M.S., and Fernando Viteri, M.D., D.Sc.

## Introduction

In studies conducted in both Cambridge, Massachusetts and Guatemala (1), we have obtained data confirming our hypothesis that iron nutrition affects cognitive function in children (2, 3). The design and conduct of these studies led us to realize the great complexities involved both in the definition of nutritional status in man and in the selection of relevant measures of cognitive function. Herein, we present some of our approaches to these problems.

The first section deals with problems surrounding the measurement of iron status; the second deals with issues relevant to the quantitation of cognitive function in children and to the cross-cultural comparability of such data.

## Selection of optimal measures of iron nutrition

The specific purpose to which a measure of nutritional status is applied, e.g., use as a public health screening tool as opposed to use as an independent variable in experimental studies, will greatly influence the selection and mode of application of available tests. Studies in which nutritional status is treated as an independent variable and related to such outcomes as frequency of organic disease, immune function, or behavior, obviously rely heavily on the ability to determine relevant nutritional status by reliable and accurate means. Such assessments are frequently difficult or impossible to obtain. In addition to straightforward problems associated with biochemical or biophysical assay procedures, one is frequently confronted with difficulty in assessing the true physiological significance of any given measure with

regard to a specific individual. Such difficulties arise from genetic and environmental influences on the nature of any individual's response to a given disturbance in nutritional status. In population studies these problems may be dealt with by appropriate statistical treatment of the measures obtained. In reaching decisions regarding individuals, however, more specific means of discrimination are required.

The assessment of an individual's plane of iron nutrition illustrates all of these problems. For, while a substantial number of direct and indirect measures of iron nutriture are now available, all are variously conditioned by genetic and environmental factors, and no single test or combination of tests has been proven superior in all contexts. As a micronutrient, iron is somewhat unusual in that response to its administration can be gauged not only by biochemical variables [serum Fe, transferrin saturation, free erythrocyte protoporphyrin (FEP)], but also, by direct quantitation (as Hb concentration) of red cell mass a tissue that is its primary physiologic "target." Thus, iron status can be defined post hoc by the individual's response to iron treatment, and the biochemical variables characterizing the pretreatment state can be related to the probability of a given degree of Hb response. In a recent study of the effects

<sup>1</sup> From the Laboratory of Human Behavior and Metabolism (R.L.L.), Rockefeller University, New York, NY; The Human Nutrition Center, School of Public Health (E.P., I.K.), University of Texas Health Science Center at Houston, Houston TX; and The Institute for Nutrition of Central America and Panama (INCAP), Guatemala City, (F.V.).

<sup>2</sup> Supported by NIH Grant R01-HD 12843.

<sup>3</sup> Address reprint requests to: Dr. Rudolph L. Leibel, Rockefeller University, 1230 York Avenue, New York, NY 10021.

of mild iron deficiency on cognitive function in children, we have used this technique to isolate these specific biochemical measures which are most sensitive to systemic iron status.

Although terminology in this area is not yet standardized, we use the term *indicator* to designate a test being used to discriminate between two diagnostic groups (e.g., the use of Hb as an indicator of iron deficiency). The *cutoff point* is the indicator value chosen as the discriminator between normal and abnormal, well, and sick, etc. The term *referent* is used to designate the outcome (dependent variable) being employed to assess a given indicator; the *referent*, then, represents the underlying reality used to discriminate between groups of subjects and can be based on clinical judgment or an operational definition (e.g., response to therapy, pathologic diagnosis, etc.). The ideal cutoff point for a given indicator is one which results in a complete separation of normals from abnormals within a population. In the usual sorts of clinical and epidemiological studies—where the distribution of normal and abnormal individuals are likely to overlap with regard to a given cutoff point—the selection of a cutoff point for any particular test will greatly affect that test's performance with regard to the separation of normal and abnormal individuals. The ability of an indicator to separate such overlapping distributions may be simply quantitated by preparing a binary table in which indicator performance at a certain cutoff point is compared to the presence or absence of disease (Fig. 1). The *sensitivity* of an indicator at a given cutoff point is the probability that the test will be abnormal in individuals with disease; the *specificity* of an indicator at a given cutoff point is the probability that the test will be negative in an individual without disease. The *predictive value* of an indicator—the probability that it will be positive or negative when the disease is present or absent—is highly dependent upon the prevalence (or physician estimate of disease likelihood) of the disease being studied. As prevalence rises, the predictive value of a positive test increases, while the predictive value of a negative test falls. The converse occurs in circumstances of lowered prevalence. Thus, the prevalence rate of the condition under

		Disease	
		Present	Absent
Test	Pos	a True Pos	b False Pos
	Neg	c False Neg	d True Neg
		$\frac{a}{a + c}$	$\frac{d}{b + d}$
		Sensitivity	Specificity

FIG. 1. Binary table for test data display. A *true positive* is an individual with a positive test who has the disease; a *false positive* is an individual with a positive test who does not have the disease; etc. *Sensitivity* is the probability that a given test will be positive in individuals with disease. *Specificity* is the probability that a given test will be negative in an individual without disease. The presence or absence of disease is defined by the *referent* (see text).

study is an important determinant of the optimal cutoff point for any indicator.

In practice, if a test of iron status is to be used as a clinical screening device for iron deficiency, one would select a cutoff point tending to maximize *sensitivity* without too much loss of *specificity* (see Fig. 1). Loss of *specificity* means that the number of false positives will be increased. Thus, some individuals without iron deficiency would be treated, but few truly deficient individuals would be missed in the screening process. In the case of iron deficiency, treatment of false positives is not much of a problem with regard either to cost or patient risk. If, on the other hand, one wanted to use this test to select a group of iron-deficient individuals for experimental studies, the *cutoff point* should be adjusted to enhance *specificity*. In actual practice, there are not separate issues. Because of the reciprocal relationship of *sensitivity* and *specificity*, one is compelled, for any given test, to trade one for the other. These points, as well as others relevant to conditional probability analysis, are well described in excellent recent reviews by Griner et al. (4) and Metz (5), and are further discussed by Habicht et al. (6) in this conference report.

In situations where multiple tests are available for the assessment of a given diagnostic entity—such as iron deficiency—one wants

not only to be able to select the optimal *cutoff point* for the application proposed for a given test, but also to be able to select the best test from among those available. Using methods derived from the theory of signal detection, an index of diagnostic accuracy for a given test may be evaluated by examining its relative (or receiver) operator characteristics (ROC) (7). The ROC is a plot of a test's true positive rate (sensitivity) versus false positive (1-specificity) rate as the cutoff value is systematically varied. In Figure 2 [adapted from Griner et al. (4)] a hypothetical ROC plot comparing two tests is presented. As the criterion (cutoff) for a positive test is made more stringent, the curves move left and down, reflecting increased *specificity* and reduced *sensitivity*. This is the cutoff location one would want in a test being used to confirm a diagnosis suspected on clinical grounds. Lessening of diagnostic stringency moves the curve up and right, increasing *sensitivity* and decreasing *specificity*. This is the cutoff position one would desire in most screening tests. Regardless of the test's context of application, it can be seen that test A is superior to test B. Thus, for example, test A achieves a sensitivity of 90% with a false positive rate of only 5%, while equivalent sensitivity is achievable by test B only at a cost of a 50% false positive rate. Conversely, the 90% level of specificity (false positive rate = 10%) occurs for test A

at 90% sensitivity but at only 75% sensitivity for test B.

The large range of normal values for Hb in iron replete individuals at any age presumably reflects the complex interaction of genetic and environmental forces in determining the level of Hb which is physiologically appropriate for a specific individual. One means of identifying iron deficient individuals for experimental purposes is to define them post hoc as having had a Hb increase (*referent*) of a given magnitude in response to treatment with iron. While this procedure may be ideal for defining iron status with regard to hematological function, one must realize that the response of Hb concentration may not reflect the status, before treatment, of organs where iron is used for biochemical processes other than Hb production. For example, the functional status of brain processes dependent upon iron for neurotransmitter synthesis/catabolism and oxidative metabolism may not be affected in parallel with the bone marrow. Any degree of systemic iron-lack may have qualitatively and quantitatively different effects on specific tissues (8). Nonetheless, a shift in Hb mass with treatment represents a firm and physiologically relevant operational definition of iron status within one readily assayed tissue, and can provide inferential insight into iron status elsewhere. Siimes et al. (9) have recently

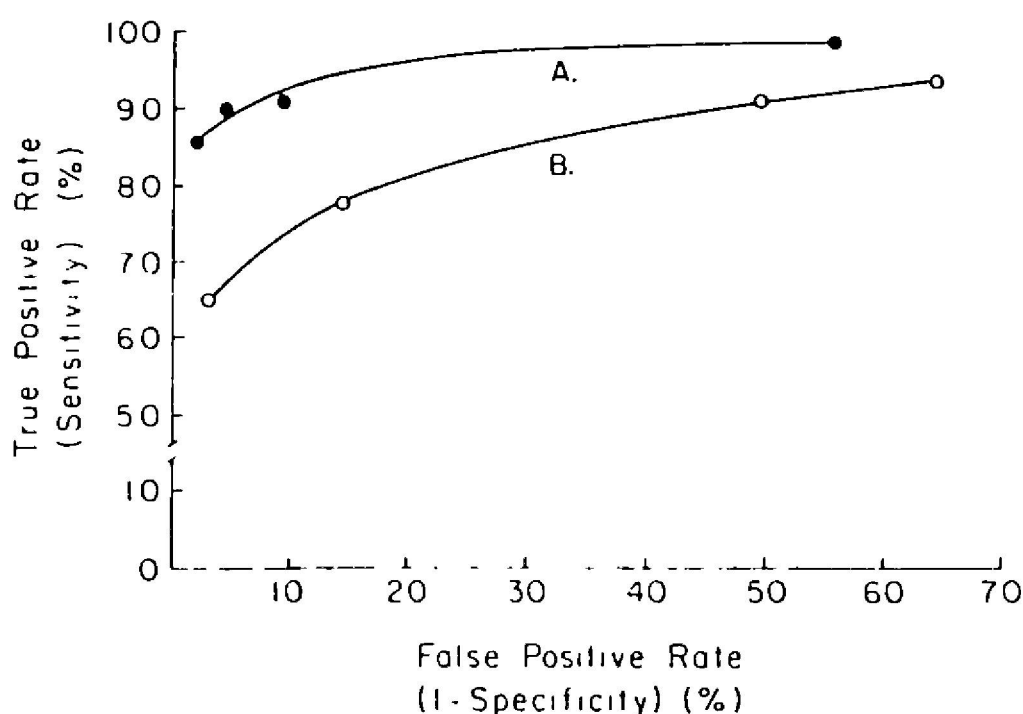


FIG. 2. Receiver operator curves for two indicators (A and B). These curves are prepared by systematically changing the cutoff point for each indicator and calculating sensitivity and specificity versus the referent at each cutoff point [adapted from Griner et al. (4)].

shown evidence for parallel changes in muscle cytochrome C, myoglobin, and Hb levels in rats fed varying amounts of iron. Thus, at least with regard to certain iron containing compounds in muscle, circulating Hb is a good "marker" for adequacy of iron availability.

We have applied this type of post hoc analysis to test selection in a study of the effects of systemic iron nutrition on cognitive function. Working through INCAP in Guatemala we screened 153 children between the ages of 3 and 6 yr for the presence of iron deficiency and anemia. Venous blood was obtained for Hb, iron, total iron binding capacity, FEP, and ferritin. After psychometric testing, 96 children were treated with oral iron (3 mg/kg/day) for a period of 3 months. All tests for iron status and psychometric performance were then repeated.

If iron deficiency is defined, post hoc, as a change of  $Hb \geq 2$  g/dl ( $n = 25$  children) with iron administration, then the various tests of iron status used in the Guatemala study can be examined with regard to their ability to detect iron deficiency (i.e., predict a Hb response  $\geq 2$  g/dl). A ROC for each test is prepared by plotting sensitivity versus 1-specificity as the cutoff value for the test is systematically changed. Figure 3 shows such a plot for all the tests used. The use of initial Hb concentration as an indicator in an ROC curve based on  $\Delta Hb$  as the referent can be questioned, since it may be assumed that both biological factors and regression to the mean effects will tend to produce an inverse relationship between initial Hb concentration and Hb with treatment. However, since multivariate analysis of initial Hb concentration as a "predictor" of  $\Delta Hb \geq 2$  g/dl yields only a relatively small (though significant)  $F$  value of 6.8, it may be concluded that in this population initial Hb concentration, per se, did not have an inordinately large impact on  $\Delta Hb$  after treatment. We have, therefore, elected to plot initial Hb as an indicator. One means of controlling for the possibility of an effect of initial value on response would be to subtract off a regression-to-the-mean factor obtained by measuring  $\Delta Hb$  in a matched but untreated group. Such a design was not possible in the study under discussion here.

It can be seen in Figure 3 that Hb and FEP are the best overall predictors of iron defi-

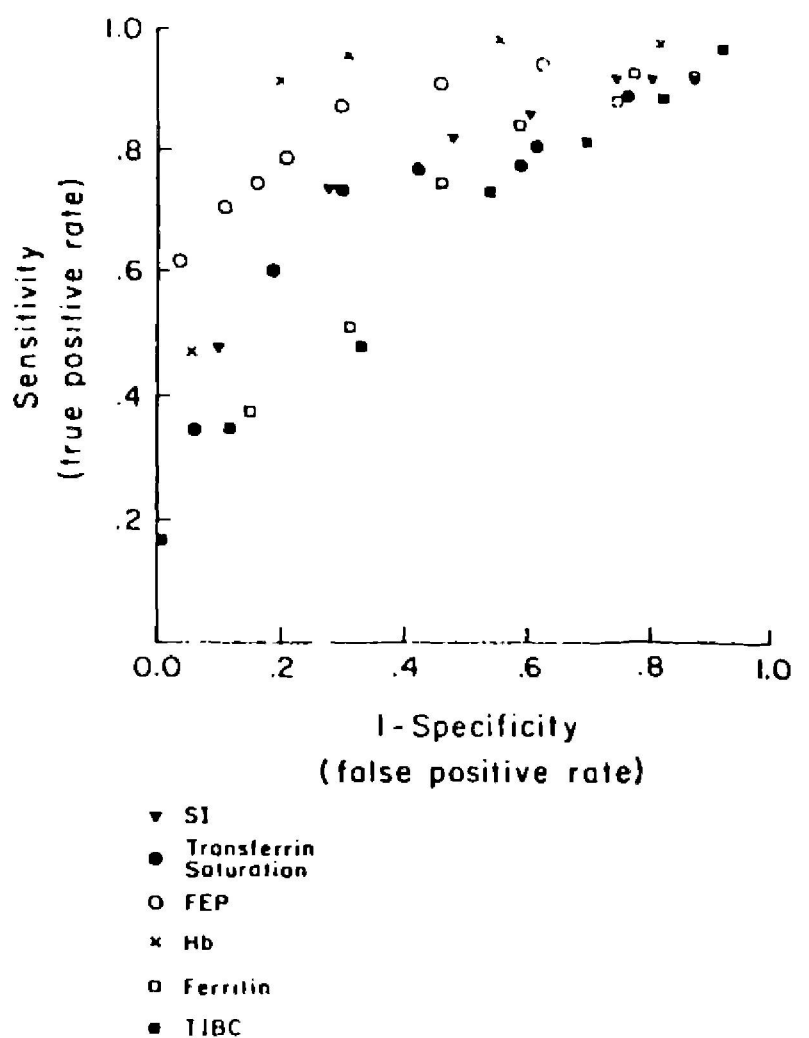


FIG. 3. Receiver operator curves for various indicators of iron nutrition. Iron deficiency is operationally defined as a  $\Delta Hb \geq 2$  g/dl in response to oral iron therapy.

ciency as defined. Serum iron or transferrin saturation are moderately efficient and, somewhat surprisingly, serum ferritin is a relatively poor test in these circumstances. The wide range of initial ferritin values indicates that this result is not an artifact due to the presence of extremely low ferritin levels in all children studied.

The significance of the difference between any pair of ROC's can be assessed in a number of ways. If the lines are parallel when plotted on double probit paper, one may compare two ROC's over the entire array of cutoff values used to generate the curves. This may be done by simply calculating the mean and standard deviation for initial test values of those individuals responding ( $\Delta Hb \geq 2$  g/dl) and not responding to the administration of iron. A simple  $t$  test of the means for a given test indicates the statistical reliability with which a test is able to distinguish responders from non-responders.

If the ROC's overlap at any point (or are otherwise nonparallel) this simple test cannot be used. However, if a specific cutoff point is selected for each test in question, e.g., by



designating a common specificity or sensitivity for each test, then one can examine the magnitude of the  $\Delta$  Hb for all individuals above and below the cutoff for each test. The "separating power" of the specific cutoff for each test is quantitated by calculating a  $t$  score for the difference between mean  $\Delta$ Hb for treated individuals above and below the cutoff. More detailed treatment of the issues and methods involved in evaluating *indicator* performance is provided elsewhere (6) in this supplement.

Having identified Hb and FEP as the best overall measures of initial iron status, the cutoff point most efficient in maximizing both *sensitivity* and *specificity* for certain purposes be determined by summing *sensitivity* and *specificity* for the range of available cutoff points and then identifying the cutoff point yielding the highest total for *sensitivity* and *specificity* ( $S_e + S_p$ ) (maximum = 2.0). A similar method was first proposed by Youden (10). An example of a plot of this series of calculations is given in Figure 4 for FEP and Hb as well as the other tests examined in this study. For each test (*indicator*) there is a

single cutoff point which maximizes  $S_e + S_p$ . In the Guatemala study,  $(S_e + S_p)_{\text{Max}}$  Hb *sensitivity* = 0.92 and *specificity* = 0.80; FEP *sensitivity* = 0.71 and *specificity* = 0.89. Thus, if one wanted a screening test which would identify virtually all affected individuals while including a minimal number of false positives (individuals who would not benefit from treatment) use of Hb as indicator with a cutoff point of 10 g/dl would be optimal for the population. This is the situation which might pertain in a public health program aimed at eradicating iron deficiency by medical treatment. Limiting false positives in this instance is desirable but not crucial, since iron treatment of a nondeficient patient is not likely to cause harm. If, on the other hand, one wanted to identify a group of individuals, all of whom were almost certainly iron deficient, one would want a test in which *specificity* was maximal at reasonable levels of *sensitivity*. If *sensitivity* is allowed to go too low, very few cases will be identified unless the prevalence rate for iron deficiency is extremely high. This is the circumstance which arises in studies of nutrition-behavior inter-

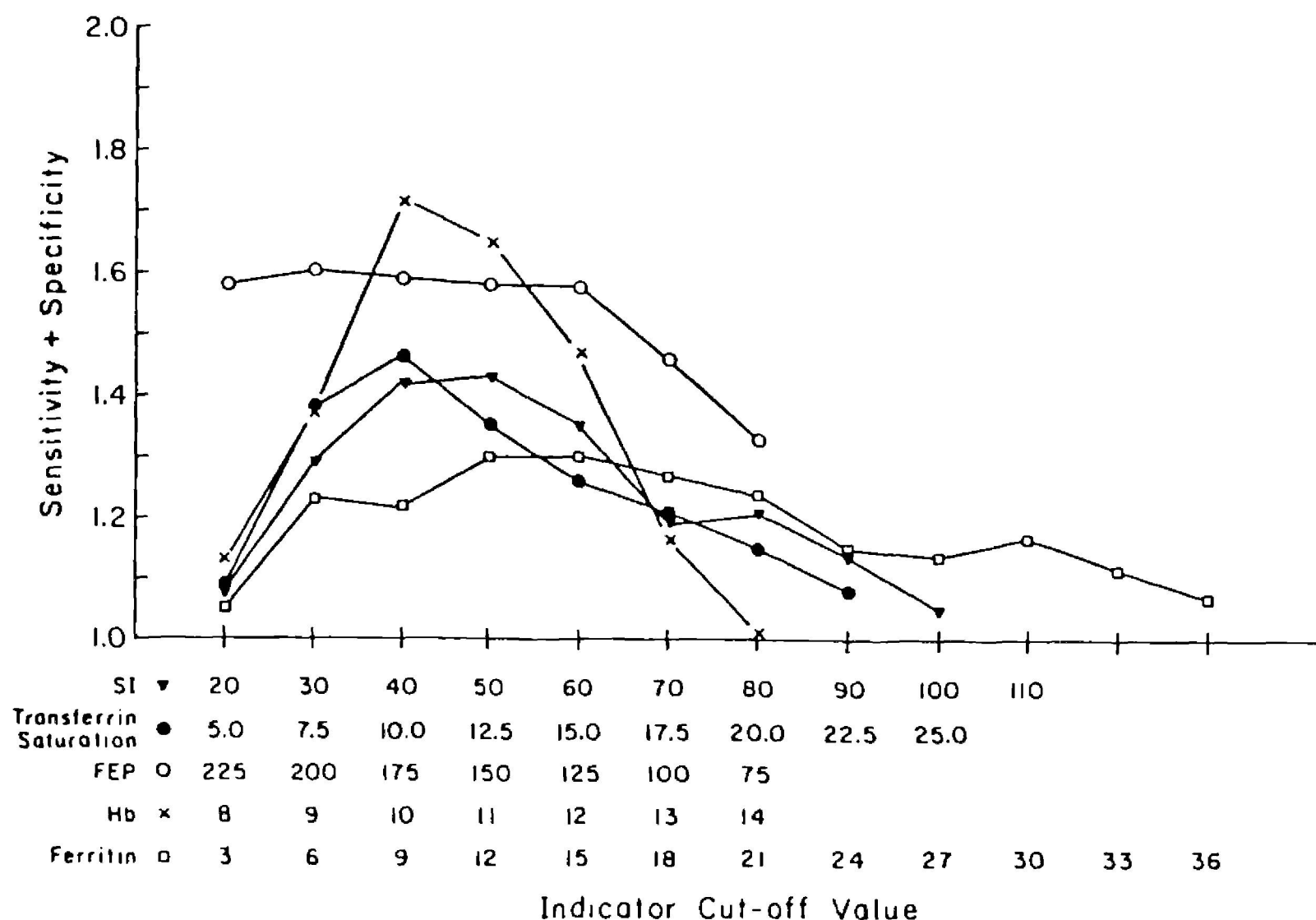


FIG. 4. Indicator selection plot using  $\Delta$ Hb  $\geq 2$  g/dl as referent. SI, serum iron ( $\mu$ g/dl); transferrin saturation %; FEP, free erythrocyte protoporphyrin ( $\mu$ g/dl); ferritin (ng/ml).

action where one wants to maximize the number of genuinely deficient subjects while at the same time to strictly minimize the number of false positives misidentified as deficient subjects. The location of *cutoff point* for any *indicator* in this circumstance will depend greatly on the prevalence of the disorder being tested for. Low prevalence rates, by reducing the predictive value of a positive test, will necessitate setting the *cutoff point* to achieve extremely high *specificity*. With regard to the performance of FEP at its  $(S_c + S_p)_{\max}$ , for example: if the prevalence of iron deficiency (Hb 2 g/dl with iron treatment) in the population is assumed to be  $25/153 = 16\%$ , then the true positive rate = (prevalence) (sensitivity) =  $(0.16) (0.71) = 0.11$ ; false negative rate =  $0.16 - 0.11 = 0.05$ ; true negative rate =  $(0.84) (\text{specificity}) = (0.84) (0.89) = 0.75$ ; and false positive rate =  $0.84 - 0.75 = 0.09$ . Thus, while only 9% of all subjects will be mislabelled as iron deficient,  $(9/11 + 9) 100 = 45\%$  of those with a positive test will, in fact, be nondeficient. On the other hand, only  $11/11 + 75) 100 = 13\%$  of individuals identified as nondeficient by the test will, in fact, be iron deficient. Thus, use of FEP at a cutoff of  $175 \mu\text{g/dl}$  as the sole *indicator* of iron deficiency in this population would generate an experimental group highly "contaminated" with well subjects. Use of FEP alone for selection of an "uncontaminated" experimental group would require use of an *indicator cutoff point* assuring a *specificity* very close or equal to 1.00. The attendant decline in *sensitivity* would necessitate use of a large population sample in order to obtain a sufficiently numerous experimental group. *Specificity* may also be increased by appropriate combinations of indicators (see below), but this also reduces *sensitivity*. Clearly, in circumstances of low disease prevalence, single or combined indicators of relatively high *specificity* may not be adequate for selection of a "pure" subpopulation of affected individuals. A converse set of considerations pertains to selection of a well or control sample. The use of an operational definition of nutritional status (e.g., response of Hb to iron administration) can circumvent many of these problems (11).

The usefulness of a given indicator as a screening test for affected individuals behaves, with regard to the influence of preva-

lence rate, in a fashion reciprocal to that for indicators selected to minimize false positives. Thus, use of Hb at a *cutoff point* of 10 g/dl as an epidemiologic screening test in our Guatemalan population (prevalence rate of iron deficiency of 16%) would result in only 1.5% of all negative tests being false negatives while 53% of all positive tests would be false positives. For reasons discussed above, such performance might be acceptable in this test used as a screening device since it identifies virtually all cases in the population and treatment of false positives is inexpensive and not hazardous. However, if the prevalence rate of iron deficiency in this population were e.g., 80%, use of the indicator Hb with a *cutoff point* of 10 g/dl would result in 27% of all negative tests being false negatives while only 5% of all positive tests would be false positives. Thus, at high prevalence rates, one requires a screening test of *sensitivity* close or equal to 1.00 in order to avoid missing a substantial portion of affected individuals. Adjusting the cutoff point to achieve such *sensitivity* will, of course, increase the false positive rate.

Allowance can be made for cost or risk factors that would, in many instances, influence one's relative weighting of preference for *sensitivity* or *specificity* in a given context. Such an adjustment can be readily made, if desired, by assigning relative numerical values (e.g., dollar cost) to false positive (FP) and false negative (FN) tests within the context examined. Because  $S_c = 1 - \text{FN rate}$  in patients with disease and  $S_p = 1 - \text{FP rate}$  in patients without disease, then  $S_c + S_p = 2 - (\text{FN rate} + \text{FP rate})$ . Thus, as  $(S_c + S_p)$  approaches its theoretical maximum value of 2.0 (FN rate + FP rate) approaches zero. If cost or other relative values are assigned to FN and FP test,  $S_c + S_p$  for any *cutoff point* can be calculated taking these values into consideration. In this fashion a "weighted"  $(S_c + S_p)$  value may be calculated for each indicator for specific cutoff points within a test. A similar result may be obtained by writing a cost/benefit equation for a given indicator, setting its first differential equal to zero and solving for the ROC curve slope at the optimal *cutoff point* (5).

An alternative means of manipulating the sensitivity and specificity with which a given group of individuals is categorized with re-

gard to a diagnostic entity is to use two or more tests to achieve the separation. If, for example, one applies the cutoff values for Hb and FEP described above to select all individuals who are deficient by *one or the other* of these tests, the *sensitivity* of the selection process will increase while *specificity* falls:

1) Combined sensitivity using Hb *or* FEP as indicator:  $\text{sensitivity}_{\text{Hb}} + (100 - \text{sensitivity}_{\text{Hb}}) \times (\text{sensitivity}_{\text{FEP}}) = 0.92 + 0.06 = 0.98$ .

2) Combined specificity using Hb *or* FEP as indicator:  $\text{specificity}_{\text{Hb}} \times \text{specificity}_{\text{FEP}} = 0.89 \times 0.80 = 0.71$ .

If these cutoff values are applied so that deficiency is defined by the presence of a below cutoff value *for both measures*, *specificity* is enhanced while *sensitivity* falls:

3) Combined sensitivity using Hb *and* FEP as indicator:  $\text{sensitivity}_{\text{Hb}} \times \text{sensitivity}_{\text{FEP}} = 0.65$ .

4) Combined specificity using Hb *and* FEP as indicator:  $\text{specificity}_{\text{Hb}} + [(100 - \text{specificity}_{\text{Hb}}) \times \text{specificity}_{\text{FEP}}] = 0.80 + (0.20 \times 0.89) = 0.98$ .

Appropriate combinations of these two measures come close to achieving optimal *sensitivity* and *specificity*. Addition of a third test could, therefore, yield only minimal improvement in these optima while further diminishing either *specificity* or *sensitivity* depending on the nature of the combination (*any* abnormal test versus *all* abnormal tests).

The type of conditional analysis described here clearly has several important applications. On the basis of relatively small pilot studies, it allows the characterization of a community with regard to optimal tests for certain nutrient-deficiency states. It assists in the selection and application of tests of nutritional status in a variety of epidemiological and experimental contexts, while taking into account a myriad of population-specific genetic and environmental imponderables which can influence biochemical response to a nutrient deficiency.

In summary, the definition of nutritional status is dependent on the purpose for which the definition is conceived. The choice of definition, in turn, dictates the location of the *cutoff point* of any test (*indicator*) subsequently used to categorize individuals by that definition. For most clinical and experimental purposes, the meaningful definitions and available biochemical measures are such that

no 100% efficient cutoff exists for any indicator; the reciprocal relationship between *sensitivity* and *specificity* forces "tradeoffs" in selectivity whose qualitative nature is determined by the investigator based upon the purpose of the study. The selection of *referent* to which various available *indicators* are compared is also critical, since the referent defines the biological outcome for which a predictive indicator is being sought. This interaction of definition of normality, selection of referent and indicator cutoff point, prevalence rates of "abnormality," and efficiency of subject categorization is frequently neglected in the design and interpretation of studies relating aspects of nutritional status to organismic function.

### Selection of tests of cognitive function

In the selection of tests to assess the cognitive function of iron deficient children we have been guided by an hypothesis derived from a critical review of the literature (2) regarding the functional consequences of sideropenia. We postulated that iron deficiency would affect the individual's capacity to attend to information in the immediate environment. Therefore, we assumed that because of their failure to attend to particular functional cues, iron-deficient children would be less successful than controls in problem solving situations.

To test this hypothesis we selected a battery of tests designed to assess aspects of attention and learning. These tests, described below, can be tailored to the specific cultural characteristics of the subjects under study without altering their basic structure. The battery includes discrimination learning, oddity learning and retention-memory tasks.

#### *Discrimination learning tasks*

Two-choice discrimination learning with three dimensional "junk" objects: the child is presented two three-dimensional objects (toy car, toy whistle, for example) mounted on 3" × 3" wooden bases. A yellow happy face is pasted on the bottom of only one of the bases. The child's task is to discover which stimulus "hides" the "happy face" underneath it. After each trial, the stimuli are rearranged (for possible left-right position alteration) out of the child's view and the procedure continues



until a criterion of seven correct responses in a row is met. The reverse problem (the previous incorrect stimulus is now correct) is then administered to the same criterion.

Two-choice discrimination learning with two dimensional "junk" pictures: the procedure is identical to the three dimensional problem except that the two stimuli are two-dimensional pictures cut from children's books, pasted on black posterboard with a "happy face" attached to the back of the appropriate stimulus.

Two-choice discrimination learning with two-dimensional color-form pictures: the procedure is identical to the two-dimensional problem above except that the two stimuli are two-dimensional colored forms. The same two colors and two forms are randomly paired on each trial (trial 1: blue X, red O; trial 2: red X, blue O) but only the form (always X or always O) is consistently correct.

#### *Oddity learning*

Three stimuli ("junk" pictures), two of which are identical, are presented simultaneously on a 7" × 18" black posterboard. The child must point to the correct picture, the one different from the other two. The only instructions given are "to find the winner." In the first series of problems, new stimuli are used on every trial. In the remaining series, the same stimuli are repeated every trial in an AAB, ABB manner. In this task, the specific makeup of a stimulus does not determine its correctness, but rather its relationship to other stimuli in the array.

#### *Memory*

A large number of two-choice visual discrimination learning problems consisting of "junk" pictures are presented concurrently for a total of four trials each. Trials 1 and 2 are massed (consecutive). Trials 2 and 3 have either 0, 4, or 8 interpolated items separating them. A "happy face" is pasted on the back of the correct stimulus.

We have been concerned with two issues related to the construct validity of these tests because we used them with children in two very different cultural settings: United States urban and Guatemalan rural areas. One concern deals with the structure of the test battery, and relates to whether the tests and test

items have the same comparative level of difficulty in both populations. Is the relative complexity of the tests—as indicated by interest comparisons of levels of performance in the children within cultures—the same in both cultures? If the comparative difficulty of test items within the battery did not yield similar trends in both cultures, then the structure and construct validity of the test would be suspect. Either we are not tapping the same psychological constructs in both places, or the children use the same psychological constructs differently in the same problem solving situations.

The second concern relates to the nature of the tests' intercorrelations. Here again, if the nature of the test associations differs between cultures, then it may be inferred that the tests are not assessing the same psychological constructs in both cultures. Although similarities in these associations do not, by themselves, confirm the construct validity of the tests, they would certainly point in that direction.

To compare the results of the test battery in both cultures, we have focused on 110 children in each of the two studies who have scores available for most of the tests used. In Cambridge, there were 59 girls and 51 boys; in Guatemala there are 56 girls and 54 boys. In both locations the ages ranged from 31 to 66 months. The comparisons are made when both groups of children were first enrolled in the studies and thus include both iron-replete and iron-deficient subjects.

Figures 5 to 7 present the levels of performance of the children in the three sets of tests. It can be seen that except for two tasks in discrimination learning where the children in both settings had the same level of performance, in all other instances, the United States children learned faster or made less errors than the Guatemalan children. Yet changes in the levels of performance from one test to the next consecutive one are nearly identical in both groups of children. In fact, there are no difference-test-scores between any two consecutive tasks, within the discrimination, oddity, and memory sets that are statistically different in one cultural group and not in the other. Accordingly, it can be concluded that the structure of the test battery as indicated by task complexity remained the same in both cultures.



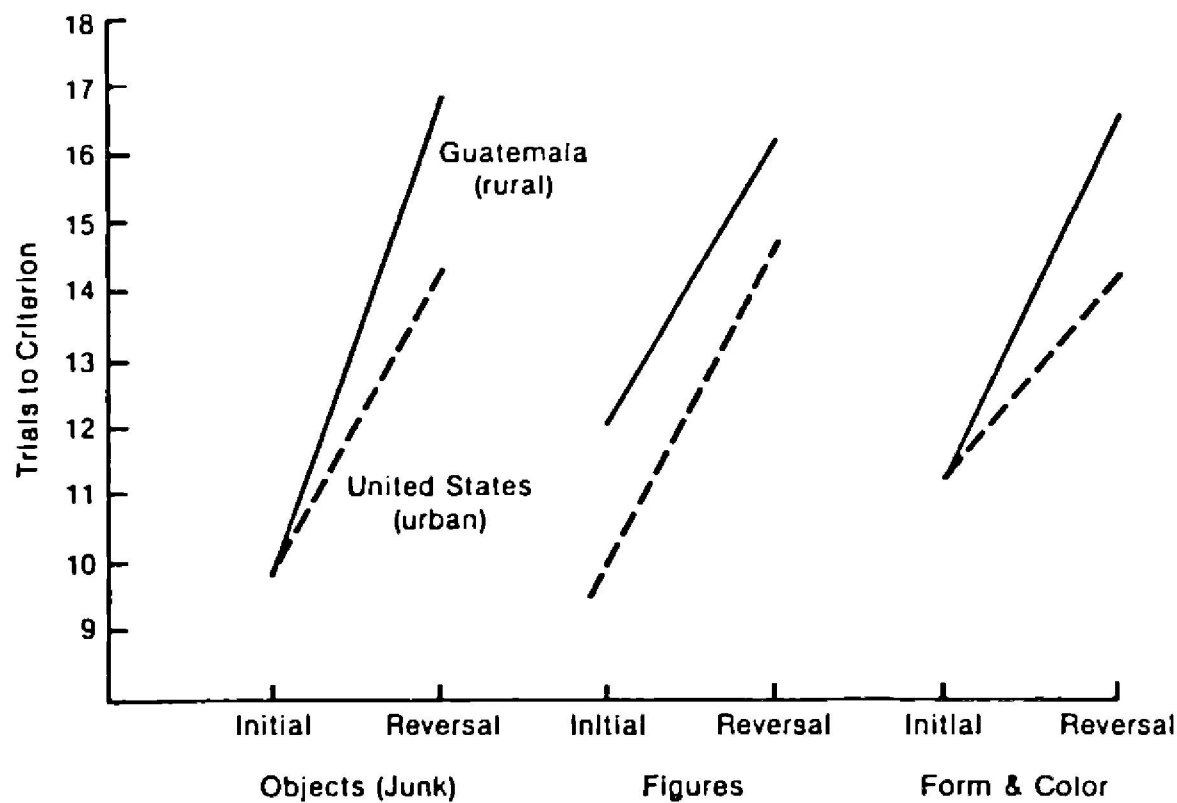


FIG. 5. Discrimination learning tasks. Estimate of difficulty of tasks for preschool children in two cultures: United States (urban) and Guatemala (rural).

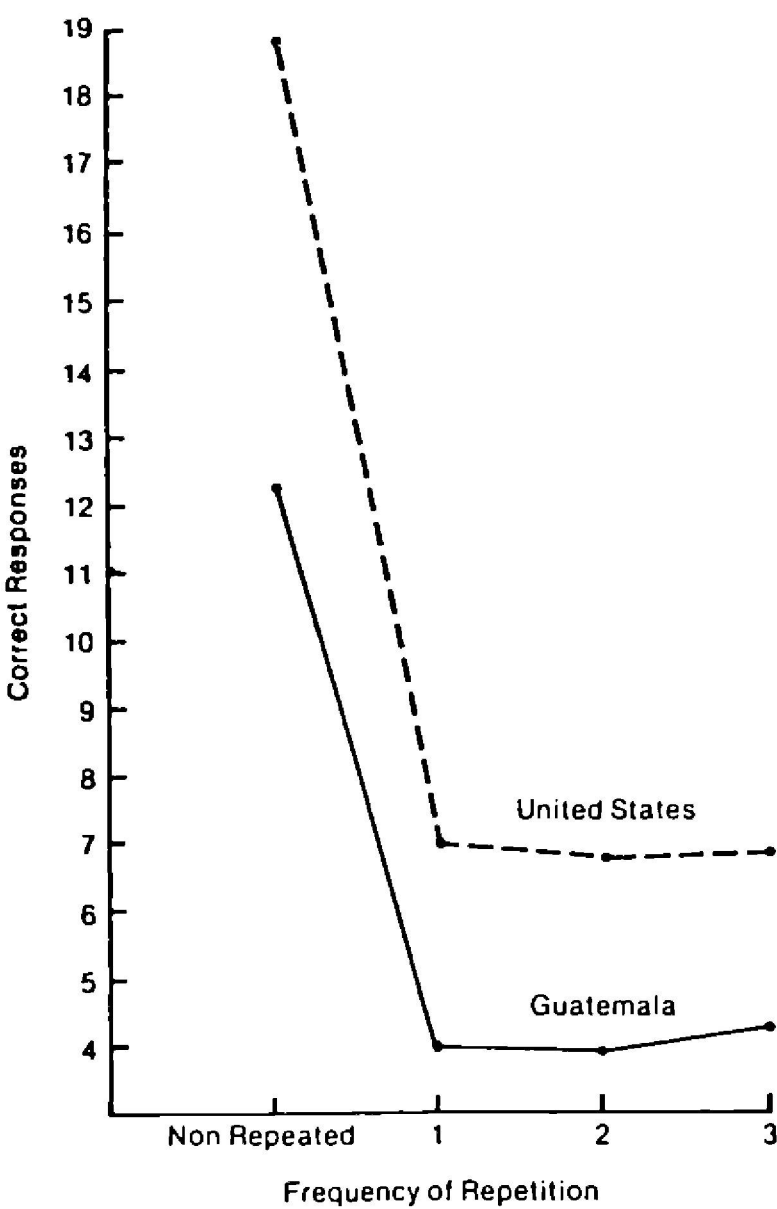


FIG. 6. Oddity learning. Estimate of degree of difficulty for preschool children in two cultures: United States (urban) and Guatemala (rural).

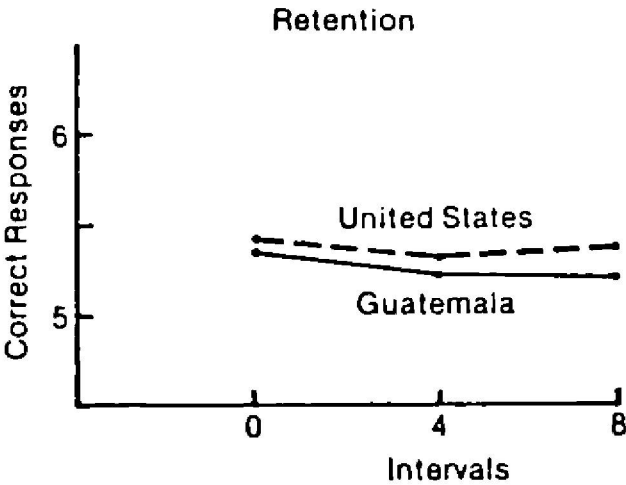


FIG. 7. Short-term memory tasks. Estimate of degree of difficulty for preschool children in two cultures: United States (urban) and Guatemala (rural).

The test scores were also factor analyzed (Table 1) to assess the nature of the test associations in the United States and Guatemalan samples. For this purpose, all scores were first transformed into Z scores to facilitate the comparison of results, both within and between cultures. Two issues are of interest here. One is whether the factors formed in the separate factor analyses of the United States and Guatemala test scores clustered the same test items. The other is whether the same factors in both data sets accounted for a similar amount of the respective test variance. The first factor formed in both sets of data includes the oddity learning tasks. In

TABLE 1  
Factor analysis\* (with varimax rotation) of scores on discrimination learning, memory, and oddity learning tasks for preschool children in Cambridge, MA (urban) and Guatemala (rural)

	Factor 1		Factor 2		Factor 3		Factor 4	
	Cambridge	Guatemala	Cambridge	Guatemala	Cambridge	Guatemala	Cambridge	Guatemala
Discrimination learning								
Object initial			0.742					0.456
Reversal			0.624	0.548				
Picture initial			0.486					0.604
Reversal			0.683	0.779				0.475
Color-form initial					0.964			
Reversal				0.602	0.430			
Memory 0						0.598	0.566	
Memory 4						0.528	0.633	
Memory 8						0.661	0.493	
Oddity learning								
Nonrepeated	0.843							
Repeated 1	0.889	0.566						
Repeated 2	0.876	0.872						
Repeated 3	0.849	0.702						
Eigen values	3.84	2.55	1.75	1.52	1.19	0.98	0.77	0.52
(Percent of variance)	(51.1%)	(45.7%)	(23.0%)	(27.2%)	(15.7%)	(17.7%)	(10.1%)	(9.4%)

\* Only those factor scores  $\geq 0.40$  are included.

Guatemala, one task (oddity nonrepeated set) was excluded from this factor; in the United States, all four oddity learning tasks were clustered together. In both analyses, this factor accounted for the highest portion of the total test variance (having the highest Eigen value). In the United States, oddity learning accounted for 51.1% of the variance, while in Guatemala, it accounted for 45.7% of test variance. This finding is of particular significance to us in connection with the use of these tests to assess the validity of our hypotheses regarding the effects of sideropenia on attention because oddity learning tasks are primarily directed toward the assessment of attention.


The results of the factor analyses of the discrimination learning tasks are particularly interesting in connection with the role of the underlying constructs in both cultures. The object and picture learning tasks tended to cluster together in both groups of children. However, while in the United States the factor made of these two tasks accounted for 23% of the total test variance, in Guatemala its explanatory power was limited to 9.4% of the variance. Thus, although the constructs behind these tasks may be similar in both groups, they seem to have different roles in relation to other cognitive processes in these two cultures. At this point, the results of the

memory test are enlightening because to an extent, they are almost the mirror image of those in the discrimination learning tests. The composition of the memory factor is identical in both places. However, in Guatemala, memory accounted for 17.7% of variance while in Cambridge it only explained 10.1% of variance.

An interpretation of the differences in the explanatory power of the discrimination and memory factors in the two cultures is that attention to salient cues in problem solving may play a more dominant role in the United States urban children as compared to the Guatemalan rural children. Conversely, memory may have a more central role in cognitive function among the Guatemalan children. These hypothetical cognitive differences may account for the differences in the amount of variability in the total test variance accounted by these two sets of tests.

An objection to the above explanation may be that within the Guatemalan data the three reversal discrimination learning tasks were clustered together explaining as much as 27% of total test variance. Presumably, the reversal component of these discrimination tasks also relies on attention to cues, i.e., change in the location of the correct response. However, it is also true that learning in this reversal shift depends on, or is mediated by, learning the

initial part of the test. If the child does not learn the first part, then the reversal shift is not given. On the other hand, learning in this initial part does not depend on any previous experience; it basically depends on the attention given to the cues given in the test. Therefore, the construct behind the three reversal tasks which make the second factor in the Guatemala data, cannot be considered equivalent to that behind this second factor in the Cambridge data. In this latter case, the factor includes the initial and the reversal shift of the object and picture determination tasks. For this reason, it still seems possible that the role of attention and memory may differ in the children tested in these two cultures.

In conclusion, both the assessment of the comparative difficulty of the tests and the results of the factor analysis, are in keeping with the assumption that the oddity, discrimination learning and retention memory tests are tapping the same constructs in both cultures. However, the differential ordering and explanatory power of the factors derived from the two independent factor analyses suggests that the same constructs may not have the same roles in problem solving in these two groups of children. The manner in which these findings qualify the results of the analysis of the effects of iron deficiency on cognition in children in both Guatemala and the United States needs to be determined. 

## References

1. Pollitt E, Viteri F, Saco-Pollitt C, Leibel RL. Functional aspects of iron deficiency. In: Pollitt E, Leibel RL, eds. Iron deficiency: brain biochemistry and behavior. New York: Raven Press (in press).
2. Pollitt E, Leibel RL. Iron deficiency and behavior. *J Pediatr* 1976;88:372-81.
3. Leibel RL, Greenfield DB, Pollitt E. Iron deficiency: Behavior Brain Biochemistry. In: Winick M, ed. Nutrition pre- and postnatal development. New York: Plenum, 1979:383-439.
4. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Ann Int Med* 1981;94:553-600.
5. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;283-98.
6. Halicht J-P, Meyers LD, Brownie C. Indicators for identifying and counting the improperly nourished. *Am J Clin Nutr* (in press).
7. Swets JA, Pickett RM, Whitehead SF, et al. Assessment of diagnostic technologies. *Science* 1979;205:753-9.
8. Dallman PR. Tissue effects of iron deficiency. In: Jacobs A, Worwood M, eds. Iron in biochemistry and medicine. London: Academic Press, 1974:437-75.
9. Siimes MA, Refino C, Dallman PR. Manifestations of iron deficiency at various levels of dietary iron intake. *Am J Clin Nutr* 1980;33:570-4.
10. Youden WJ. Index for rating diagnostic test. *Cancer* 1950;3:32-5.
11. Dallmon PR. Biochemical and hematologic indices of iron deficiency In: Pallit E, Liebel RL, eds. Iron deficiency: brain biochemistry and behavior. New York: Raven Press (in press).