V. DISCRIMINATING AMONG NORMAL INFANTS BY MULTIVARIATE ANALYSIS OF BRAZELTON SCORES: LUMPING AND SMOOTHING

KENNETH KAYE

University of Chicago

One of the aims of the Brazelton Neonatal Assessment procedure (NBAS) is to elicit the same kind of behavior in an infant that is likely to affect the parents as they handle the infant during the first few months of life. The procedure and ratings which Brazelton included in the exam were chosen less for correspondence to personality traits in later life than for their salience in the social environment of early infancy. Thus, NBAS scores provide an attractive possibility for assessing the infant's initial contribution to the mother-infant dyad.

This particular task can only be successful, however, if the rating scales indeed distinguish reliably among normal infants. While we agree whole-heartedly with Horowitz, Sullivan, and Linn (Chap. III) about the folly of assuming behavioral stability in the newborn, the NBAS can only assess the infant's contribution to interaction with others if it is responsive to something "true" of the infant beyond the one examination session.

A first step in our task, therefore, is to reduce the several dozen items of the NBAS to a set of dimensions showing a reasonable test-retest reliability, that is, some characterization of individual differences stable in newborn infants at least over the first few days or weeks of life. I shall show that such dimensions exist and that they can be measured by averaging normalized scores over related items and then averaging over several independent examinations of the infant in each age period.

GROUP COMPARISONS VERSUS INDIVIDUAL DIFFERENCES

Scores on the NBAS have been shown to distinguish between healthy normal control samples of American infants and such groups as prenatally malnourished newborns in Zambia (Brazelton, Koslowski, & Tronick 1976a) and Guatemala (Brazelton, Tronick, Lechtig, Lasky, & Klein 1977), "small-for-dates" infants (Als, Tronick, Adamson, & Brazelton 1976), and methadone-addicted newborns (Strauss 1976), and to predict which infants in a "high-risk" sample will be neurologically abnormal by age 7 years (Tronick & Brazelton 1975).

Without reviewing this literature, three issues should be kept in mind when we turn to the problem of detecting reliable individual differences in normal infants. First, how should we interpret studies which do not find differences between target groups? For example, in a well-screened sample of healthy delivering mothers in Boston Lying-In Hospital, no consistent effects of regional obstetric anesthesia could be found (Tronick, Wise, Als, Adamson, Scanlon, & Brazelton 1976). A problem is that investigators do not yet have enough confidence in the power and theoretical significance of the assessment to decide whether these studies should be regarded as failures to detect differences or as successful demonstrations of the absence of differences.

Second, from one group-comparison study to the next there is little consistency in the items, clusters, or factors which distinguish each target sample from each other or from controls. In principle there is nothing wrong with this state of affairs—it is quite reasonable that malnourished newborns should be irritable, hypotonic, and unreactive, while methadone addicts are irritable, tremulous, and jerky in their movements. But the danger is that there are so many possible combinations of scores one can easily find differences of some kind in any one study and interpret them sensibly. Despite the between-group differences found in the studies above, all of them found more items on which the groups were *not* different than items on which they were different. Until recently (e.g., Strauss & Rourke, Chap. VI), we have not had empirical analysis of patterns of scores across many studies.

Third, what distinguishes one group of infants (e.g., prematures, abnormals, addicts, Chinese, Navajo, etc.) from another group need not be the same dimension which distinguishes individual infants from one another. The important aspects of newborn infants considered from the point of view of neurological abnormality, cross-cultural differences, or treatment effects may not be the important aspects from our own point of view: the development of communication skills in normal mother-infant dyads.

The question of primary concern in this paper is one which is relatively unimportant to investigators looking at group differences. In those studies each infant within a sample is used essentially as a replicate of the treatment

or group effect; so long as systematic biases are kept out of the testing procedure, it is of little importance how representative any one infant's performance is of his or her performance on other occasions. When individual differences are the object of study, on the other hand, within-sample variance is the whole point, and the extent to which an individual's NBAS scores are stable over time relative to other individuals is very important.

RELIABILITY AND STABILITY OF SCORES

We were concerned with test-retest reliability over two or more different examinations, preferably on different days. Reliability of the examiner at any one testing can readily be established by training, practice, and periodic checking side by side with specially trained "trainers" in Boston, Kansas, or Seattle. The agreement of two examiners, however, that Baby X is a 6 on pull-to-sit at a particular time tells us nothing about the adequacy of the test itself as an instrument for assessing pull-to-sit performance as a reliable characteristic of individual babies.

Few studies had attempted to ascertain test-retest reliability, and no studies of which we were aware tested the number and degree of item score agreements against the number and degree that would be expected by chance (e.g., the pioneering study by Horowitz, Self, Paden, Culp, Faub, Boyd, & Mann [1971] did not include this control). In general, as Horowitz et al. point out in Chapter III, item-by-item correlations, even when statistically significant, tend to be low.

Nor had test-retest reliability been shown for factor scores. (However, see Sameroff, Krafchuk, & Bakow, Chap. IV.) Factor analysis has been the most common multivariate treatment of NBAS item scores, and, since Bakow, Sameroff, Kelly, and Zax (1973), a comparable set of factors has emerged in study after study (see Strauss & Rourke, Chap. VI): alertness, irritability, consolability, sometimes habituation. Items not loading consistently on these factors, such as tonus, defensive movements, activity, cuddliness, and others, usually cluster in two or three additional factors in any one sample, but there is less consistency across studies than is the case with alertness and irritability.

The similarity of factor solutions from one study to the next tells us only that trained testers rate on the basis of similar factors in Boston, Chicago, Rochester, Seattle, etc. Because the items are not independently rated, we do not know whether the factor solutions tell us about underlying dimensions in infants or about underlying global reactions on the part of the tester, which cause him or her to rate sets of items in a correlated way.

In view of the paucity of evidence for the validity of factor scores, it may be wondered why we persist in this kind of analysis (i.e., the attempt to reduce 27 scores to a smaller set of dimensions). The answer (and here I

presume to answer for all the authors of this *Monograph*) is that the factors correspond better to our sense of what distinguishes one newborn infant from another than do any of the single items (Brazelton 1973). The constructs—alertness, irritability, consolability, maturity, and state control—have face validity to anyone who has examined newborns, who has observed infants interacting with adults, and who has talked with mothers and fathers about the impressions their babies make upon them.

In fact, it is just as likely that the problem lies with factor analysis and that more appropriate data-reduction techniques would yield more reliable individual differences. Items not rated independently will inevitably yield factors. However, far from being a flaw in the NBAS, this is one of its deliberate features. The assessment is expressly designed to consist of a full examination, followed some minutes later by ratings in which the tester reflects back over the session as a whole. This in turn arises directly out of the thesis that an infant's performance fluctuates over time (both between and within neurophysiological states), as well as the attempt to make the tester's experience more comparable to that of a sensitive caretaker than to that of a laboratory technician. The items which load together turn out to be those which the designers of the assessment expected to correlate with one another. Accordingly it may be more appropriate to regard each group of items—for example, irritability, peak of excitement, and rapidity of build-up—as a set of corroborating measures of something one knew a priori one was interested in, and ignore the loadings of the 24 other items on that factor. One would then abandon the factor score based on all 27 loadings in favor of a simpler concatenation of the three deliberately redundant measures.

For similar reasons Adamson, Als, Tronick, and Brazelton (1975; Als, Chap. II) designed a method of reducing item scores to ratings on four dimensions which were based upon postulated neurophysiological and psychological processes in the newborn. The typologies were particularly designed to isolate individuals with markedly superior or markedly inferior performance and therefore did not reveal much variance among our "healthy normals." (The value of this approach in the identification of high-risk infants is shown by Sepkoski [1977].)

The purpose of our study was, initially, to find a small set of dimension scores which could be said to characterize the infants in a Chicago longitudinal sample during the first two weeks of their lives. To help us with this task Brazelton, Tronick, and their colleagues provided a much richer set of data from Boston Lying-In Hospital. We then expanded the purpose of the study: to find out how much improvement could be obtained in the stability of various dimension scores by smoothing them over several independent examinations. We began with what we took to be the underlying notions of the Adamson/Als scheme as well as a distillation of the results of factor analysis. Our two requirements were that the data reduction be applicable

to NBAS scores obtained with the existing procedure (Brazelton 1973) and that it preserve variance along continuous dimensions.

METHOD

Two co-workers and I administered the NBAS "blind" to 50 full-term Caucasian neonates on the second day of life. Examiners were ignorant of the sex, name, ethnic origin, and prenatal and perinatal history of each infant. Mothers had been screened for pregnancy complications requiring prolonged bed rest and for excessive use of drugs or alcohol, and were limited to native English speakers born in the United States. In fact, they were a cross-section of the working-class white Catholic and Protestant families of Chicago, with a few welfare mothers and a few upper-middle-class mothers. All of the babies were delivered in Columbus Hospital, a community hospital serving the North Side of Chicago.

Only infants to be discharged within 3–6 days after birth were included in the study (actually, all infants were discharged by the fourth day except two who were detained for phototherapy of mild hyperbilirubinemia and two whose mothers were detained because they had been delivered by Caesarean section). At age 2 weeks the exam was repeated in the infants' homes.

Three major differences were unavoidable between the two testings: the 2-week exam at home could not of course be done blind; the mother, often the father, and sometimes up to seven other family members, relatives, and neighbors were present; and the time of day varied (in fact these testings were sometimes interrupted for a feeding and resumed afterward if the infant did not fall asleep). The examinations in the hospital all took place at about 5:00 o'clock in the afternoon of the infant's second day of life, 1 hour prior to feeding time (i.e., an infant born Sunday morning was examined Monday afternoon, and one born Sunday night was examined Tuesday afternoon). The "2-week" testings were scheduled more loosely, ranging between day 13 and day 19 (90% by day 17).

Between-tester agreement was achieved in the following way. I trained my two collaborators (Marilyn deBoer and Alan Fogel), both of whom had also been trained by Daniel and Nina Freedman, so that agreement between any two of us reached the criterion of "100% within two levels and 85% within one level." The two collaborators then went to Boston and reached the same criterion with one of Brazelton's assistants. On separate occasions early in our data-gathering period, Brazelton and a different assistant visited Chicago and again checked our agreement. Toward the end of the data-gathering period the Chicago testers confirmed that our agreement among ourselves was at the same criterion level. (The division of labor was not equal after the first dozen infants: deBoer became our main Brazelton examiner, Kaye and Fogel serving mainly as stand-ins.)

We were able to test only 50 of 51 infants at 2 days and 43 at 2 weeks. Thus for longitudinal analysis our N was 42.

Since the major purpose of the assessments was to extract variables characterizing the infants' initial contributions to the developing infant-mother dyad over a 6-month period, we attempted to reduce the item scores to a few factors or dimensions. Four different approaches to this problem will be discussed: factor analysis, canonical correlation, lumping and smoothing, and multiple regression.

In view of the limitations of our Chicago data, we obtained scores from seven NBAS exams (days 1, 2, 3, 4, 5, 7, and 10) on each of 54 infants examined in Boston Lying-In Hospital (Tronick et al. 1976). This sample was comparable to our own but more rigorously selected for optimal pregnancies and deliveries. The same examiner never performed more than two successive exams nor more than five of the seven exams on each infant. The data from Boston will be referred to as the BLI data, and our Chicago sample will be referred to as the Columbus data.

RESULTS

We begin with the Columbus data.

An important nonresult of our study—important only because we know of no other study in which examiners were kept ignorant of the infants' sex—was our failure to find significant sex differences on any item or any factor score, or any of the dimensions derived later by lumping and smoothing, either at 2 days when our examinations were "blind" or at 2 weeks when they were not. Our guesses as to the infants' sexes, after completing both the exam and the scoring, were no better than chance: 61% of the male and 60% of the female infants were guessed as male. Other investigators report that they are able to distinguish male from female infants (Brazelton, personal communications, 1967–1977; Strauss 1976), but they present no data. We (Kaye, deBoer, and Fogel) also thought we had done better than chance, until we looked at the data.

More disturbing, we found not a single significant difference on any item as a function of the sex we had *guessed*. This leaves considerable doubt as to what basis we used in making the guesses: perhaps a physical characteristic which did not affect our scoring, or perhaps no consistent feature at all.

Factor Analysis

Table 1 gives loadings for a five-factor solution rotated so as to maximize total variance accounted for (Varimax). Habituation items are excluded

TABLE 2

Pearson Correlation Coefficients: Factor Scores from Age 2 Days with Factor Scores from Age 15 Days (Columbus Infants)

15 Days	2 Days				
	Factor I	Factor II	Factor III	Factor IV	Factor V
Factor I...............	−.100	−.120	.109	−.161	−.123
Factor II..............	.079	.253	−.230	.024	.266
Factor III............	−.172	−.320*	.142	.030	.121
Factor IV............	.299	.195	.089	−.154	−.102
Factor V.............	−.010	.163	−.161	.049	.120

* $p < .05$.

present purposes was the fact that we could detect no significant degree of correlation (taking account of the number of correlations run) between any factor at any age and the "same" factor at any other age in the BLI data.

Canonical Correlation

Canonical correlation is technically questionable with such data for most of the reasons factor analysis is questionable: systematic nonindependence of the rating scales, the presence of discontinuous and nonlinear scales, the variety of shapes found in the distributions of scores for different items, etc. It does have the advantage, however, of looking for possible relationships between any and all items over time. With factor analysis we tried first to find items which were related to one another within a testing, and then to see if each cluster of items so related, or "factor," might correlate with similarly derived factors on other days. Canonical correlation takes essentially the opposite approach: deriving just those combinations of items on different days which best correlate with one another over time.

Using the same items listed in table 1, we were unable to extract a statistically significant set of canonical variates. Using days 1 and 10 of the BLI data as well as days 5 and 7 and days 7 and 10, we again failed to produce canonical solutions whose level of significance exceeded chance.

We then returned to the Columbus data and related smaller sets of items which had consistently been found to load on common factors (e.g., items 5–9 from 2 days canonically correlated with items 5–9 from 2 weeks), but without success. These items clustered together within any given test session, but the factor did not predict from one test to another 2 weeks later, whether we used factor scores (table 2) or canonical correlation.

Lumping and Smoothing

Factor analysis and examination of the distributions of scores in the Columbus data, the BLI data, and the other studies cited above led to the following reasoning:

Items 5–10 (tracking, orienting, alertness) are somewhat redundant and to some extent are interdependently remembered and scored by the examiner. They are comparable to separate questions on a subtest of the Stanford-Binet. In this respect it makes sense to average each infant's performance over these items. This is different from simply using Factor Score I in that all of the other items also contribute their loadings to that factor, probably in a spurious and inconsistent fashion.

Items 17–19 (peak of excitement, rapidity of buildup, irritability) form another such set.

Items 16, 25, and 26 (consolability, self-quieting, and hand-to-mouth) generally load together on a common factor, and, while they clearly tap different behavior, we have no empirical basis for separating them. They all reflect an infant's ability to enter or maintain a state of calm (i.e., state 4), though with different degrees of intervention on the part of the examiner.

Items 21 and 22 load together in other studies if not ours (e.g., Strauss 1976). Tremulousness and startles might make a similar impression upon caretakers. In the absence of a theoretical distinction, it makes sense to regard these two items together as did Adamson et al. (1975; see Als, Chap. II) in their Dimension IV.

Items 11–13 (tonus, motor maturity, pull-to-sit) are very inconsistent in their loadings in different studies as well as in the BLI data for different days. However, all three are sometimes intercorrelated and all pairs of the three are often intercorrelated.

Skin-color lability (23) and state lability (24) were found together in five of our seven factor analyses of the BLI data.

Cuddliness and activity (items 14 and 20) are so inconsistent in their factor loadings that we cannot justify lumping either of them with any other item; finally, item 15 (defensive movement), though clear and easy to score, has none of the empirical, theoretical, or intuitive associations on which our reasoning was based, so we chose to ignore it entirely.

We thus created six sets of related items, plus cuddliness and activity. The grouping of items is shown in table 3. (Because these are not factors, I have avoided the traditional names for NBAS factors such as irritability, alertness, etc.) The BLI data were used for this analysis. Before the scores within each set were averaged together, we separately standardized the item scores for each day. A subject's score for tracking and orienting (TO) on day 1 then became the mean of six z scores (or fewer if any of the items were missed on that day). I shall refer to these means over related items as dimension scores, using TO for tracking and orienting, etc., as shown in table 3.

The next step was to smooth the dimension scores by averaging them over successive days. Selecting the first 4 days as representing week 1, we produced for each of the eight dimensions a set of 10 variables: the 4 single

TABLE 3

GROUPING OF ITEMS

Tracking, orienting (TO)	Strength, tone (ST)
5. Tracking ball	11. Tonus
6. Orienting rattle	12. Motor maturity
7. Tracking face	13. Pull-to-sit
8. Orienting voice	
9. Face and voice	Twitchiness (TW)
10. Alertness	21. Tremulousness
	22. Startles
Fussiness (FS)	
17. Peak of excitement	Lability (LA)
18. Rapidity of buildup	23. Skin-color lability
19. Irritability	24. State lability
Calming (CM)	Cuddliness (CD)
16. Consolability	
25. Self-quieting	Activity (AC)
26. Hand-to-mouth	

days, three means of pairs of successive days, two means over 3 successive days (1-2-3 and 2-3-4), and one over all 4 days. Each of these 10 predictor variables was correlated with two outcome variables: the dimension score for day 10 alone and the mean of days 7 and 10. (Day 5 was ignored in this analysis in order to leave a gap in time between predictor and predicted variables.)

Although there were 10 × 2 × 8 correlation coefficients, our interest was in comparing the prediction obtained by simply reducing item scores from a single exam with that obtained by further smoothing over two or more separate assessments on different days. This is shown in figure 1 for the TO dimension. The lower curve represents prediction to day 10, the upper curve prediction to the mean of days 7 and 10. Note that .18 is the mean of four correlation coefficients: day 1 with 10, 2 with 10, 3 with 10, and 4 with 10. In this and the following figures, a coefficient better than .26 must be obtained in order to go beyond the .05 level of significance. We can see that this dimension predicts from week 1 to week 2 only if we smooth over two different tests in each week, or over three tests in week 1.

Figures 2–6 present comparable data for the dimensions twitchiness, fussiness, calming, strength and tone, and cuddliness. There are no curves for lability or activity because no amount of smoothing over days produced coefficients significantly different from zero.

Twitchiness (fig. 2) proved the most stable of these dimensions and the only one predicting significantly from only a single exam in days 1–4 to a single exam at day 10.

Fussiness was a relatively poor dimension, from the point of view of ...reliability, but figure 3 shows that by smoothing an infant's

performance over three or four examinations one begins to be able to predict something about this dimension at least to the next week.

Strength and tone and cuddliness (figs. 4 and 5) showed the most marked improvement in predictability as a result of smoothing. Calming (fig. 6) was also greatly improved, but the coefficients never became as high as those for some of the other dimensions—probably because calming is partly dependent upon fussiness.

Multiple Regression

To measure how much additional variance in the predicted measures was accounted for by each additional predictor day, we performed multiple linear regression with each of the Ys in figures 1–6 as dependent variables and the single days 1–4 as independent variables (e.g., TW10 regressed on TW1, TW2, TW3, and TW4; TW7–10 also regressed on TW1, TW2, TW3,
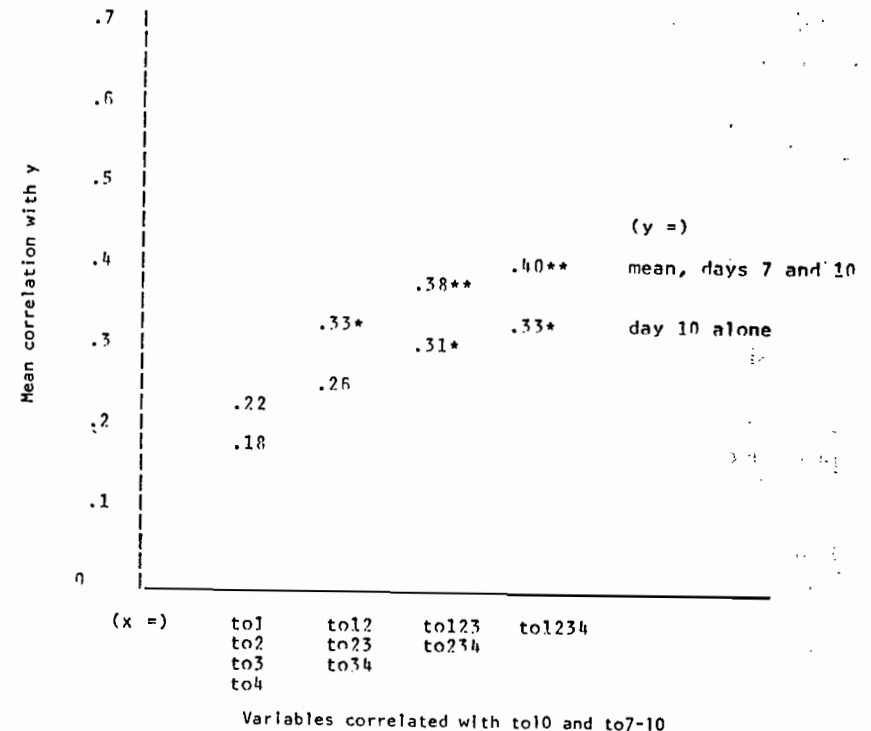

FIGURE 1.—Tracking, orienting dimension. Correlation between week 1 and week 2 scores as a function of number of independent examinations in each week. Each point is the mean of the correlation coefficients of the variable listed under Y with each of the variables listed under X. These variables themselves have been smoothed over several days where indicated (e.g., TO234 is the mean for this group of items over days 2, 3, and 4). *$p < .05$; **$p < .01$.

and TW4; etc.). Surprisingly, with a few exceptions the first independent variable entered (using a stepwise method) accounted for most of the multiple r^2, and additional days' data did not significantly increase the variance accounted for. This was the case because the independent variable with the highest correlation to the dependent variable was always entered first—sometimes day 3, sometimes day 2, etc.

More Canonical Correlation

In view of the improvement in linear correlation achieved by smoothing over successive days, we again attempted canonical correlations. This time we used as sets of variables the first five dimensions listed in table 3; lability and activity were omitted because of their near-zero correlation over time, and cuddliness because it consisted of only a single item. Canonical correlation was done in three different ways, all more successful than our previous attempts.

First, the full set of five dimension scores, each smoothed over days 1–4, was correlated with the full set of five dimensions smoothed over days 7 and 10
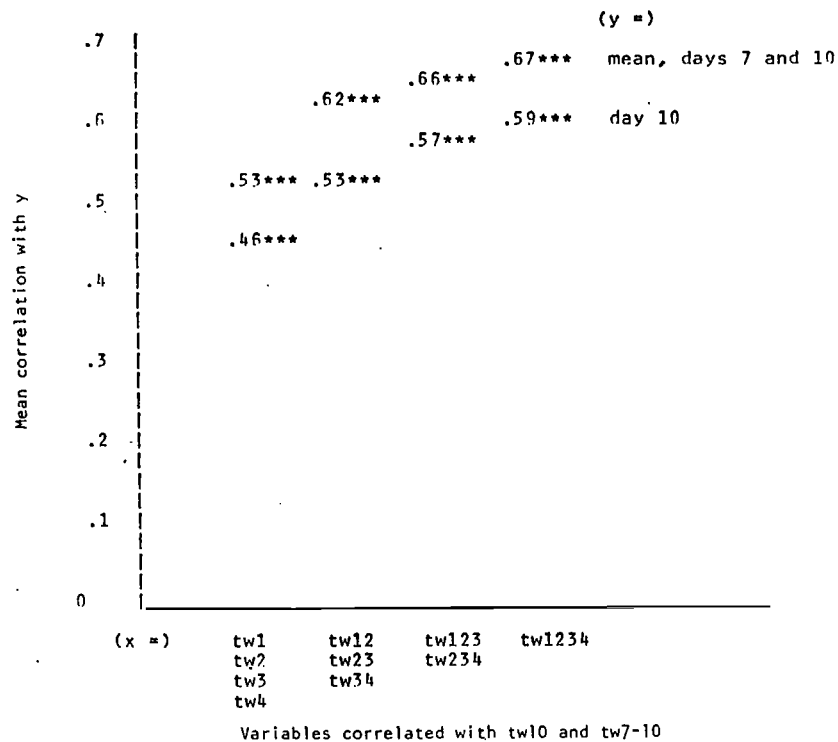
(i.e., TO1234, TW1234, FS1234, CM1234, ST1234 with TO7–10, TW7–10, ... ST7–10). Three canonical variates accounted for 75% of the variance in this set of five scores "lumped and smoothed" from the original item scores (Bartlett's $\chi^2 = 74.0$, 39.5, 20.3; 25, 16, and 9 df, respectively).

Second, for each dimension we correlated the four scores representing the separate days in the first week with the two scores representing the second week (e.g., TO1, TO2, TO3, and TO4 with TO7 and TO10). This resulted in substantial correlations for only two of the five dimensions, as shown in table 4, part A. Note that twitchiness and strength-and-tone were the two dimensions obtaining the highest linear correlation when we simply averaged over days. The eigenvalues in table 4, pt. A (% variance accounted for) are approximately equal to or slightly larger than the multiple r^2 obtained for each set using multiple regression (see above).

Finally, we performed corresponding canonical correlations, this time keeping the items separate within each dimension and smoothing them individually over days (table 4, pt. B). Using this method, all dimensions except fussiness were found to be correlated strongly over time.
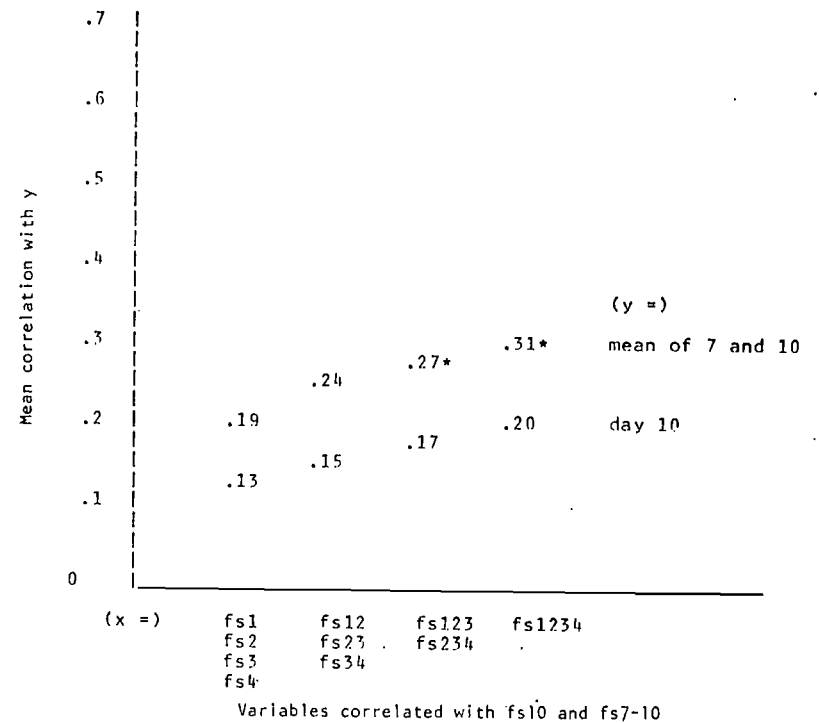

FIGURE 2.—Twitchiness dimension. ***$p < .001$.

FIGURE 3.—Fussiness dimension. *$p < .05$.

TABLE 4

CANONICAL CORRELATION, WEEK 1–WEEK 2[a]

First Set	Second Set	Canonical Correlation, Eigenvalue	x^2	p
A. Smoothing over Related Items, Set of Separate Days				
Tracking, orienting..... TO1 TO2 TO3 TO4	TO7 TO10	CC = .42 EV = .18	11.9(8)	N.S.
Twitchiness.......... TW1 TW2 TW3 TW4	TW7 TW10	CC = .68 EV = .46	32.1(8)	<.001
Fussiness......... FS1 FS2 FS3 FS4	FS7 FS10	CC = .34 EV = .12	6.7(8)	N.S.
Calming......... CM1 CM2 CM3 CM4	CM7 CM10	CC = .49 EV = .24	13.2(8)	N.S.
Strength, tone......... ST1 ST2 ST3 ST4	ST7 ST10	CC = .65 EV = .42	32.0(8)	<.001

Note.—Degrees of freedom are given in parentheses.
[a] First canonical variate only.

TABLE 4—Continued

First Set	Second Set	Canonical Correlation, Eigenvalue	x^2	p
B. Smoothing over Days, Set of Separate Items				
Tracking, orienting...... Ball wk 1, Rattle wk 1, Face wk 1, Voice wk 1, Face-Voice wk 1, Alert wk 1	Ball wk 2, Rattle wk 2, Face wk 2, Voice wk 2, Face-Voice wk 2, Alert wk 2	CC = .69 EV = .47	56.6(36)	<.02
Twitchiness........... Trem wk 1, Startle wk 1, Peak wk 1, Rapidity wk 1, Irrit wk 1	Trem wk 2, Startle wk 2, Peak wk 2, Rapidity wk 2, Irrit wk 2	CC = .67 EV = .45	30.3(4)	<.001
Fussiness...........		CC = .33 EV = .11	9.9(9)	N.S.
Calming........... Consol wk 1, Self-Quiet wk 1, Hand-Mouth wk 1	Consol wk 2, Self-Quiet wk 2, Hand-Mouth wk 2	CC = .56 EV = .32	27.5(9)	<.001
Strength, tone........... Tonus wk 1, Matur wk 1, Pull wk 1	Tonus wk 2, Matur wk 2, Pull wk 2	CC = .58 EV = .34	37.4(9)	<.001

CONCLUSIONS

Our work has by no means resulted in the final word on reliable individual differences among normal newborns. The NBAS itself omits several important dimensions along which individual differences are likely to be found, including learning and feeding. Furthermore, we have omitted the habituation items from this analysis. Even within the set of items analyzed here, many reasonable combinations have not been tried. We do not feel prepared to conclude that any of our dimensions is more reliable than any other, or than others which we have not tried. Instead, my conclusions will be confined to some observations on the limitations of multivariate methods applied to this problem and some comments on particular items as presently scored.

Given the variance among normal infants, the variance in performance of individual infants from one examination session to the next, the inevitability of some missing data, and the modest number of subjects possible in any one homogeneously selected and uniformly tested sample, I believe that factor analysis, canonical correlation, and multiple regression of item scores are inappropriate. These techniques are in a sense too powerful in that they make maximum use of fortuitous correlation among items within any one sample, encouraging an investigator to "see what goes with what" rather than to test an hypothesized dimension.

Thus our factor scores failed to predict significantly over days, I believe, because each factor was rotated so as to maximize the variance accounted for by the full set of items, and so as to be orthogonal to all other factors. In fact our sense of what the newborn human is like directly opposes such an approach: the Brazelton examination was designed to tap sets of responses and aspects of behavior which fall into relatively discrete, logically, and phenomenologically distinct groups but are not expected to be orthogonal in any sense. Alertness or tracking-and-orienting happened to correlate about .25 with calming in both our samples, yet they are clearly different dimensions. Rather than cope with this by oblique rotation of factors and/or by fixed rotation through designated items, the nature of NBAS data makes it more appropriate simply to average over particular sets of normalized item scores, equally weighted.

Canonical correlation was used to find factors at each of two ages which would have maximal correlation with one another. The first attempt involved sets of items from one day to another and accounted for no more
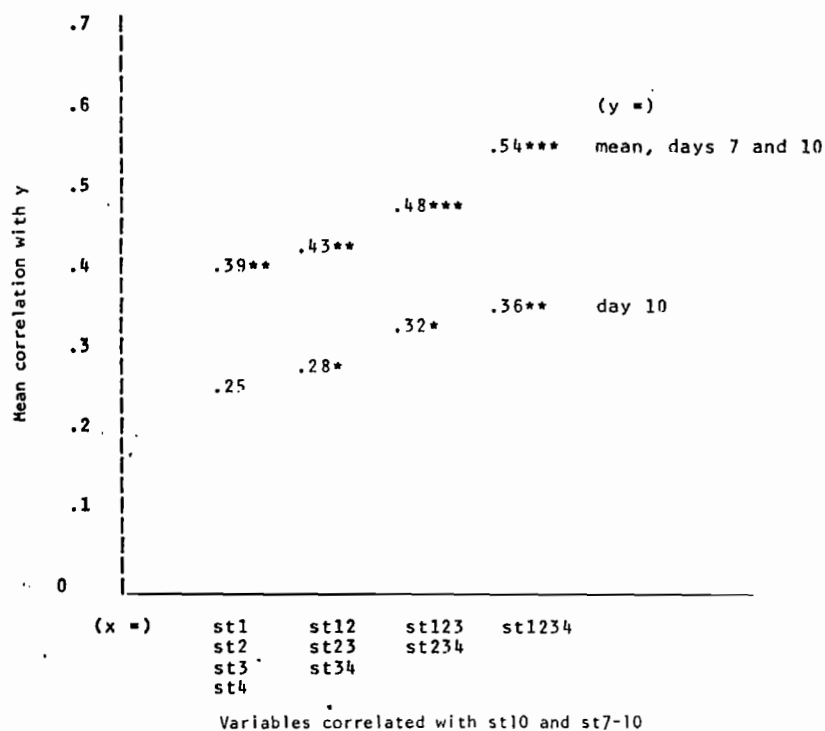

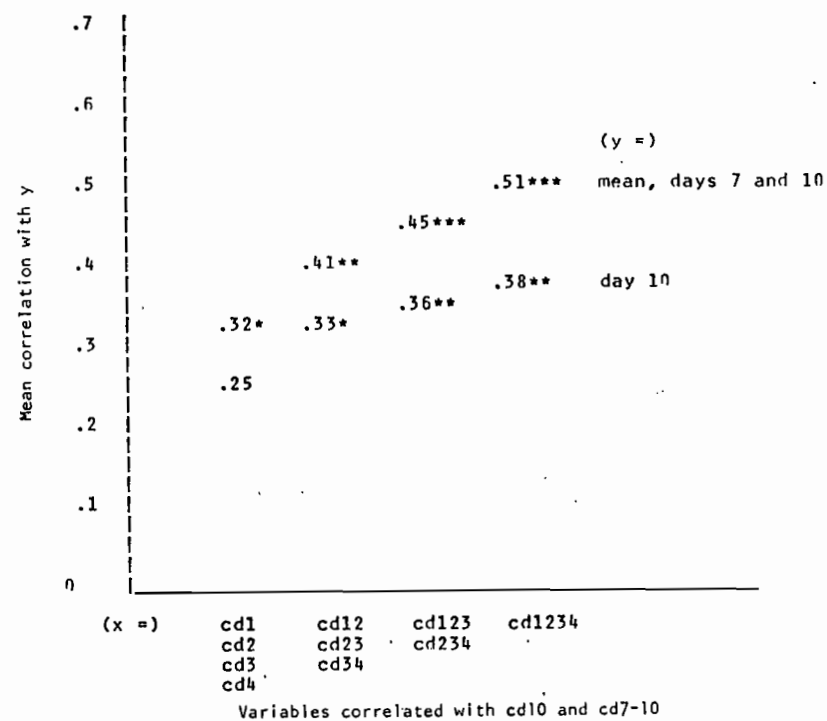FIGURE 4.—Strength, tone dimension. *$p < .05$; **$p < .01$; ***$p < .001$.

FIGURE 5.—Cuddliness (item 14 alone). *$p < .05$; **$p < .01$; ***$p < .001$.

variance than could have been achieved by chance, given the size of the matrix. The second attempt was more successful, based on smoothed scores over several days. One cannot, however, predict that the particular item weights on each canonical variate will bear any resemblance from one sample of infants to another: only that in general a canonical variate exists. So there is no evidence that the canonical solution is more generalizable than the simpler expedient of averaging equally weighted items.

A good method seems to be the following: (1) score the exam in the conventional manner on as many separate occasions as possible; (2) then normalize the scores for each occasion across the whole sample of infants; (3) then average over exams for each separate item (the method used in table 4, pt. B, rather than figures 1–6 or table 4, pt. A); (4) then average together the related items forming each dimension. Whether one then produces scores with predictive validity for mother-infant interaction or for any other criterion remains an empirical question, but at least one proceeds with greater confidence that these scores are representative of individual differences in the first 2 weeks of life. (Probably the best predictor would be
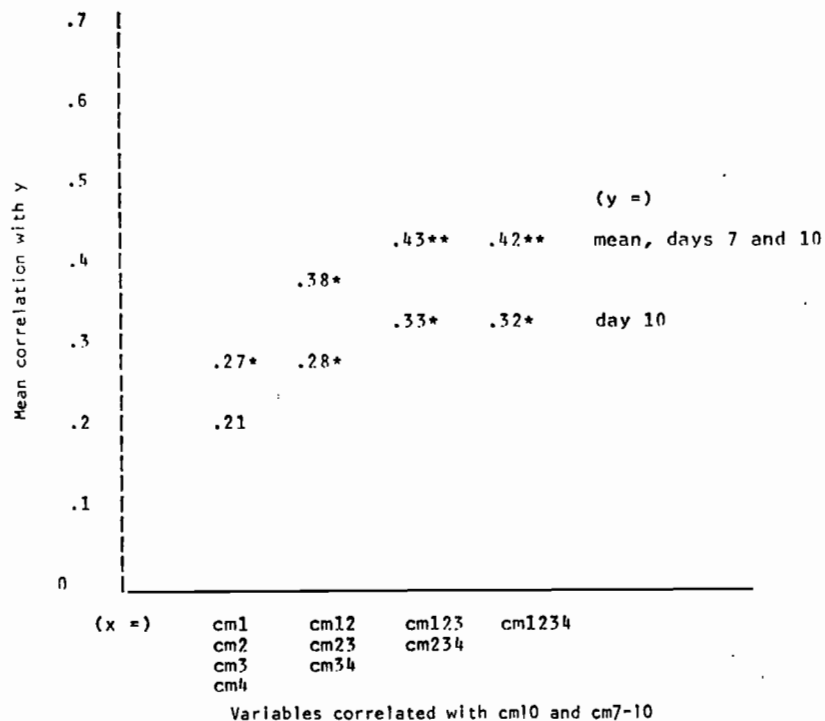
the canonical variates rather than the simple means, but this has the disadvantage of using different formulas for every sample.)

Multiple regression too proved a more powerful technique than was needed or wanted from the point of view of generalizability. We saw that averaging over successive days accounted for additional variance and, in the case of our dimensions, elevated the prediction above the magic threshold associated with a .05 level of significance. Yet, in most cases, multiple regression indicated that entering a second day into the equation accounted for no significant variance. The discrepancy in these findings is accounted for by the fact that stepwise regression entered first the score from whichever day (1–4) had the highest correlation with the dependent variable (day 10, or the mean of 7 and 10). The point is that each investigator will not know, for his or her new sample of infants and for any given dimension, which precise day is going to be the best predictor—our multiple regression, like the other multivariate techniques discussed, was capitalizing on fortuitous correlation. In actuality, the best method is to examine infants as independently as possible on two or three occasions and smooth the data. Exactly the same point can be made with respect to averaging over related items to arrive at a dimension score: after the fact, it may turn out that one item accounts for all the significant variance; a priori, however, a set of mutually redundant but somewhat distinct scores maximizes the chance of predicting significantly.

The dimension "fussiness" is a good illustration of the points argued in this paper. In virtually every sample of NBAS scores, a factor of irritability or fussiness emerges. Sometimes there is a separate consolability factor, sometimes inconsolability appears on the same factor with irritability. In any case, the latter is nearly always the strongest or second strongest factor. Yet, whether we use the factor score or simply average over the three items which typically load on that factor, we cannot predict better than chance, from one testing to another, which infants will be irritable. This means it is a matter of predominant states in which the infant happens to be at the time of the exam rather than an aspect of the infant as an individual.

However, when we average the fussiness scores over a few exams, we begin to detect a stable dimension as shown in figure 3. Thomas, Chess, and Birch (1968) found that infants' temperament may be one of the earliest stimuli to the kind of negative-affect interaction cycles which breed later disorders. It is clear that temperament cannot be assessed by one examination session, and probably not with much predictive validity even by two examinations.

The general point, that NBAS items are intentionally redundant and can be lumped into a priori dimensions, applies to research involving group comparisons as well as individual differences. The discriminant analysis

FIGURE 6.—Calming dimension. *$p < .05$; **$p < .01$.

used by Lester and his students (Coll 1977; Sepkoski 1977) is an approach similar to ours, suited especially to compare target groups.

Future attempts to arrive at the most stable dimensions of individual difference, and future studies of particular groups as well, should reexamine some of the interpretations I made of individual items and improve upon them either in the scoring or in the lumping into dimensions. (E.g., I cannot really defend placing item 12, motor maturity, with items 11 and 13; or explain the relation of skin-color lability to state lability; etc.) Unfortunately this recommendation creates problems in generalizability across studies, but I believe that ossification of the exam in its present state would be even more unfortunate. Having proved its clinical utility (Brazelton, Parker, & Zuckerman 1976b; Tronick & Brazelton 1975), it is nonetheless open to further improvement as a research tool. Linn (1978), for example, has demonstrated the correlation of the orientation items—especially as scored in the NBAS-K—with mothers' proportion of feeding as opposed to "grooming" activity during a feed, and with the infant's being found in deep sleep between feedings. The picture this suggests, that an infant who is alert during the exam is one who sleeps well and eats well (as is true throughout life!) contributes to an understanding of newborn behavior and justifies the Kansas group's attempts to improve upon the NBAS.

Finally, as the other authors of this *Monograph* show, stability of neonatal characteristics is not the whole story. Brazelton from the outset suggested that change over the first week or 2 postpartum would be more likely to tell us something important about infants—that is, would be more sensitive to prenatal and perinatal stresses and would have a greater effect upon caretakers—than the scores on any particular day. In particular, the scores on days 1 and 2 would reflect immediate response to the birth trauma; infusion of hormones from the mother, and obstetric medication (Tronick et al. 1976). The way an infant recovers from these shocks, Brazelton and his group argue, reveals something about his neurological as well as his metabolic coping abilities and may make a more lasting impression on his mother than do his characteristics on any one day.

Clearly what is still lacking is an integration of the vast body of assessment and prediction data with a coherent theoretical conception of human infancy.