INTELLIGENCE 12, 183–197 (1988)

Continuity in Intellectual Growth from 12 Months to 9 Years

LLOYD G. HUMPHREYS
TIMOTHY C. DAVEY
University of Illinois
Urbana–Champaign

The traditional view that infant tests of development do not measure intelligence because they cannot include appropriate abstract, symbolic content is discussed and an alternative interpretation is proposed. Intercorrelations of measures of intelligence from 9 months to 9 years are used as the principal empirical basis for the alternative interpretation. Although correlations over time decrease at a more rapid rate in infancy, change appears to be smooth and continuous starting at 12 months. Nine months is questionable and earlier ages can be rejected. Estimated true score stabilities obtained by fitting the simplex model to the observed correlations increase rapidly during infancy and the preschool period and appear to level off at about .97 during the school years. This degree of stability is certainly high but, also, clearly imperfect. It allows for a great deal of change in relative intelligence during the school years. The hypothesis that infant tests, starting at 12 months, measure the construct of general intelligence on which children's relative scores change rapidly early in development cannot be rejected in these data.

INTRODUCTION

It has been known for many years that individual differences measured by infant tests of development are not highly correlated with individual differences on standard intelligence tests of school-age children. Some of the reported correlations are very close to zero (Bayley, 1949), but most are small positive (McCall, Hogarty, & Hurlburt, 1972). The widely accepted explanation for these small correlations is that it is not possible to measure in infancy the same factor or factors measured later in development. As the content of these tests changes, shifting from perceptual-motor to verbal items, the tests become better measures of intelligence. A related view is that the instability of preschool IQs gives way to stability once appropriate items are included in the test.

This research was supported by a grant from the National Institute of Mental Health MH2361-06 and by the Research Board, Urbana–Champaign Campus, University of Illinois.

Correspondence and requests for reprints should be addressed to Lloyd G. Humphreys, Department of Psychology, 603 East Daniel, Champaign, IL 61820.

Anderson's Alternative

Anderson (1939, 1940) noted that correlations tended to be high between adjacent occasions even in the preschool period and became progressively smaller with the passage of time between test and retest. He developed an explanation for the preschool instability, however, that required some degree of continuing instability during development. A child does not start a test from scratch on each occasion of measurement. Items or item types that were passed the year before are passed again during the current year. In addition, a child attempts additional items. This provides, in effect, part–whole correlations between successive occasions of measurement. Anderson concluded that the various test–retest correlations could be explained if raw scores or mental-age gains from year to year were independent of the raw score or mental-age base at the beginning of the year. The longer the time interval, the larger and more variable the gain, thus producing the pattern of high correlations between adjacent occasions and low correlation between remote. Anderson's analyses were shortly corroborated by Roff (1941) who published distributions of correlations between annual gains and initial bases. These were distributed with means very close to zero.

The Simplex Model

Several years later, Guttman (1954) described a model for intercorrelations taking the form of those analyzed by Anderson and Roff. This was the simplex. Initially, however, the model was applied to tests administered at the same point in time that were inferred to be measuring the same content at varying levels of complexity. For example, tests with numerical content can be arranged from simple clerical checking of numbers, through increasingly complex numerical operations, to arithmetic reasoning. Adjacent tests are highly correlated, and remote tests have lower correlations. Guttman assumed that the elements responsible for each addition to complexity were independent of the elements at the simpler stage. This explanation is basically similar to the one suggested by Anderson for successive performances on intelligence tests, but Guttman did not discuss this possible relationship.

The simplex model did clearly apply to the product–moment correlations among binary items in a perfect Guttman scale. Such items measure a single common factor, as the numerical tests measure the same content, but differ from each other along a continuum of difficulty or popularity. It is important to realize that a simplex matrix will necessarily define multiple factors when analyzed by the standard factor methods but can be explained by a single factor moderated by a process such as increasing complexity or difficulty. When test items are arranged in order of difficulty or popularity, correlations can change gradually from high to low from one item to another without changing the factor content measured by each one.

Humphreys (1960) applied the simplex model explicitly to the correlations between successive trials or occasions of learning data. It is not known that such

correlations approximate the simplex form with a high degree of generality. When measurement error is held constant, the highest correlations in a matrix are between adjacent trials or occasions, and the smallest correlation in the matrix is the one between the pair of trials or occasions that are most remote from each other. It is not essential that gains be independent of the base, as Anderson suggested for growth in intelligence and as Guttman required of true scores in the simplex model, but only that true score gains be less than perfectly correlated with the true score base. This condition seems inevitable as people learn and forget. As means increase or decrease, the relative standing of persons in the group changes.

It is useful at this point to define terms used to describe such matrices more precisely. A true simplex requires error-free measurement and, under that condition, zero partial correlations between all nonadjacent elements when an intermediate element is held constant. It is possible to fit this model to observed correlations, estimating the reliabilities required to correct for attentuation in the process. Inability to reject the model does not, of course, prove that gains are independent of the base. Other true-score models can, as described in the preceding paragraph, produce observed correlations that are similar to the ones required by the simplex model. As a generally applicable term, use of quasi simplex is desirable for observed matrices that have the descriptive characteristics of large values adjacent to the principal diagonal with a reduction in size from there to the end of the column or row.

Explanations for a Quasi-Simplex

In a 1960 article, Humphreys described a dilemma in the interpretation of quasi-simplex matrices. This was described somewhat simplistically as "change in people," relative to the performance of others in the sample, versus "change in task." Insight to this problem can be gained by consideration of physical development. Measures of height during development produce intercorrelations that form a quasi-simplex matrix as a function of change in people. Examples can be found in Humphreys, Davey, and Park (1985). Height at 8 years has correlations of .80 and .84 in girls and boys with height at 17 years; height at 2 years can be inferred to have still lower correlations with height at 17 years. In one sense, the task does not change because a measure of length is still a measure of length at any age. On the other hand, the task does change because height at 8 years and other preadolescent ages, especially for girls, has larger correlations with measures of intelligence than has height at 17 years. There are undoubtedly other functional differences between early and late height, even though there is no doubt that the same physical dimension is being measured throughout.

Fleishman's research (Fleishman, 1960; Fleishman & Hempel, 1954, 1955; Fleishman & Rich, 1963) on the acquisition of motor skills is also relevant to this issue. The intercorrelations of trials or blocks of trials form quasi-simplex matrices. The metric used to measure the acquisition of a motor skill is constant

from trial to trial, as is the metric for length, but the correlates of the task change as learning progresses. For example, early trials in a discrimination reaction-time task are more highly correlated with a variety of cognitive measures than are later trials, whereas the latter are more highly correlated with simple reaction time. Both the height and motor skills findings are the parallel of those obtained for Guttman's simplex matrices in tests administered on the same occasion. Complex verbal tests are more highly correlated with complex than with simple numerical tests, and simple verbal checking tests are more highly correlated with simple numerical checking than with complex arithmetic reasoning. The tests on each simplex differ functionally from each other. On the other hand, all of the tests in a given simplex are measuring the same factor in terms of content, either verbal or numerical. These examples indicate why Humphreys's contrasting explanations were simplistic. They cannot be disentangled simply by observing different correlates of early and late trials or ages.

Here we look at the problem of change in relative intelligence test performance in infancy, preschool, and the early school years. Our starting point is a table of previously published (Wilson, 1983) intercorrelations of observed scores in this age range. Next, we fit the simplex model to the obtained correlations. The purpose of fitting the model is to determine whether or not the data are congruent with a continuous process of change of the sort envisioned by Anderson, Roff, and Guttman. This is followed by a theoretical formulation that allows the construct of general intelligence to be measured by the infant tests. Finally, we suggest research that will further clarify the nature of the changes taking place as we measure intellectual development.

LOUISVILLE TWIN PROJECT DATA

The Observed Intercorrelations

The tests in the Louisville Twin Project of Wilson (1983) were administered at ages varying from 3 months to 15 years at varying time intervals of 3 months to a 6-year gap between 9 and 15. Tests used included the Bayley, Stanford–Binet, McCarthy, and Wechsler scales. The data are longitudinal in nature, but sample sizes vary widely among the correlations reported. As in all longitudinal studies, there are missing cases at particular ages, but larger gaps were the result of the vagaries of funding. Use of the largest possible sample size for each correlation minimizes sampling errors, but the somewhat greater independence of sampling errors produces more reversals for the model.

Table 1 contains the correlations published by Wilson along with the sample sizes obtained from the author. The steps taken to obtain a sample of twins ensured a wide range of talent. Samples are larger than most of those that have been used in longitudinal studies, but both members of the twin pair contributed scores to these correlations when scores for both were available. This means that correlations are not as stable in the sampling sense as the *n*s suggest. On the other

TABLE 1
Intercorrelations and Sample Sizes of Mental Test Scores Louisville Data, From 3 Months to 15 Years

Ages	Months								Years						
	3	6	9	12	18	24	30	36	4	5	6	7	8	9	15
3[a]		54	48	44	36	26	25	16	13	22	20	24	18	30	
6[a]	439		56	49	41	29	34	29	19	31	32	32	25	31	
9[a]	380	465		60	46	36	40	36	27	32	29	22	21	20	
12[a]	404	484	455		56	48	42	40	31	38	36	32	32	33	
18[a]	387	447	422	470		67	62	60	54	54	49	47	47	48	31
24[a]	379	448	419	469	522		70	74	68	63	61	54	58	56	47
30[b]	274	326	304	335	388	400		78	71	66	66	60	59	56	55
36 months[b]	346	424	391	432	491	505	414		76	72	73	68	67	65	58
4 years[c]	297	352	314	365	410	427	326	509		80	79	72	72	71	60
5[d]	192	343	306	342	402	410	427	326	509		87	81	79	79	67
6[d]	281	318	282	316	381	385	324	468	477	590		86	84	84	69
7[e]	162	293	259	280	315	317	246	328	326	414	487		87	87	69
8[e]	221	250	214	238	301	300	248	370	379	471	529	497		90	78
9[e]	179	202	173	189	249	242	195	308	278	341	336	250	402		80
15 years[e]					54	49	70	114	161	201	246	183	313	163	

[a]Bayley
[b]Stanford–Binet
[c]McCarthy or WPPSI
[d]WPPSI
[e]WISC

hand, the correlations are more stable than the number of pairs suggests. The second member of the pair furnishes information that is partially but not completely redundant with that of the first. The sampling problem, however, has little effect on the size of the correlations.

Correlations at 3 and 6 months depart quite radically from the quasi-simplex pattern shown by the remaining occasions of measurement. Starting with 9 months, there are relatively few small reversals from the progression in size expected. When one allows for the differences in elapsed time between occasions, the occasion-to-occasion correlations adjacent to the principal diagonal indicate gradually increasing stability during early development. There is no age in these data when intelligence stabilizes, and there is no difference in pattern of correlations between infant tests other than those administered on the first two occasions and later tests. The differences are in the amount of change as a function of chronological age.

Methodological Considerations in the Model Fitting

Reliabilities are required for fitting a simplex. Because they are a function of the range of talent in the population sampled and of the conditions of test administration, we decided to estimate them from the intercorrelations rather than use

values from test manuals. In order to obtain a solution, it is necessary to fix a minimum number of values for one of the parameters. We chose to fix the reliabilities of the first and last occasions to be equal to their respective neighbors. This has the effect of making the true-score stability coefficients between the first and last pairs of occasions arbitrary, as well.

In addition to being a measure of goodness of fit, chi square is also a direct function of sample size. There seems to be fairly general agreement that the common use of alpha equal to .05 for a difference in means should not be generalized automatically to use of LISREL (Joreskog & Sorbom, 1978) for model fitting. The problem of interpreting chi squares is made more acute in the present case by our decision to use the maximum n for each correlation. Dependencies among the sampling errors among intercorrelations are consistent and predictable when a constant size sample is used, whereas the introduction of greater independence makes it more difficult to find a good fit for a model. Finally, there is the problem of partial dependency from one member of the twin pair to the other.

It is possible to play a research game with sample size that is unproductive scientifically. If we had been able to recompute Wilson's correlations on a sample of single births, we would have done so. We would not have eliminated cases that lacked complete data because our expectation would have been smaller chi squares and larger p-values, obtained at the price of larger residuals. Because of our interest in the size of residuals, we also used a program that fits a simplex to observations using a least-squares criterion.

Goodness of Fit of the Model

Three analyses are presented for the Louisville data for which we report approximations of the chi square statistic. One includes all ages from 3 months to 7 years. Ages 8 and 9 were deleted because inspection alone indicated an inadequate fit. A second analysis includes 9 months through 9 years, and the third is based on 18 months through 15 years. Summaries of the goodness-of-fit information for the three analyses are contained in Table 2. In place of a single chi square for each analysis, which would have required the section of an arbirtrary sample size, we elected to use several levels of n. The largest of these is well above the median (290) of the tabled ns. The second is below the median and the third, well below, but these and other smaller values might be justified by the use of both members of the twin pairs in the correlations.

The values of the root mean-square (RMS) residual at the bottom of Table 2 are based on the least-squares criterion. Because we were also interested at this point in the stabilities of our parameters as the composition of ages changed, we reinserted ages 8 and 9 in the group that included 3 and 6 months.

The fit of the model for the youngest group in the Louisville data is probably acceptable, but the omission of ages 8 and 9 brought this about. It was clear at the outset that their correlations with 3 and 6 months represented a departure

TABLE 2
Several Indices of Goodness of Fit of the Model in Three Age Ranges

	Maximum-likelihood Criterion		
	3 months to 7 years (45 *df*)	**9 months to 9 years (45 *df*)**	**18 months to 15 years (36 *df*)**
χ^2, n = 325	95.30	63.49	92.18
χ^2, n = 275	80.62	53.71	77.99
χ^2, n = 225	65.96	43.41	63.81
p = .05	61.63	61.63	50.96
p = .01	69.92	69.92	58.57
	Least-squares Criterion		
	3 months to 9 years	**9 months to 9 years**	**18 months to 15 years**
RMS of residuals	.031	.020	.018

from the simplex pattern. When these two ages were reinserted in the least-squares analysis, the RMS residual became more than 50% larger than the same statistic in the second and third samples. Measured individual differences at 3 and 6 months are not congruent with the simplex model of continuity from one occasion to another.

It is not unreasonable to assume that a continuous process might not start at very early ages, but it is unreasonable for it to stop at 7 years when continuity has been demonstrated between 11 and 17 years by Humphreys and Parsons (1979) with closer fits of the model. Therefore, we deleted 3 and 6 months and added 8 and 9 years in the second analysis and found an acceptable fit. Residuals are relatively small and do not form the consistent pattern that one would expect if there were a qualitative break in the continuity of development. Both the pattern of residuals and the size of chi squares reveal the qualitative break in the growth of height (Humphreys et al. 1985) during the adolescence growth spurt. Although clustering of large residuals of the same sign was not observed, the 9-month occasion produced more than its share of large residuals (RMS = .034).

For the oldest sample, the several chi squares reported are larger, but the residuals are, on average, smaller than in the second analysis. It is probably not coincidental that the largest residual occurred between 18 months and 15 years where sample size was only 54. The 15-year occasion also produced a RMS of .032. The inclusion of 15 years provides the only overlap with the data of Humphreys and Parsons in which the fit of the model was excellent. There is more overlap with the intelligence test data of Humphreys et al. (1985) but the fit of the model was inadequate in the latter case. There may be valid reasons for the inadequate fit in the multiplicity of intelligence tests represented, the absence of

a constant sample size, and the inappropriateness of group tests in years 7 to 9, but the simplex model could not be fit satisfactorily to the observed data.

Parameter Estimates

Table 3 presents reliabilities and adjacent-occasion stabilities for each of the three analyses in which the least-squares criterion of goodness of fit was used. Values of the reliabilities and stabilities of the first and last occasions are deleted from this table because they could not be freely determined.

For occasions where there are two or more reliability estimates, the agreement found is generally satisfying. The largest discrepancy, .71 to .75 at 12 months, includes the occasion in the first analysis most heavily influenced by correlations with 3 and 6 months. The progression in size of reliabilities from early to late occasions is not monotonic, but the trend is in the expected direction.

In interpreting the adjacent-occasion true-score stabilities, one must keep in mind the length of the interval between occasions. The following rule of thumb is helpful: To compare shorter intervals with a full year, take the fourth power of the 3-month and the second power of the 6-month stability. Starting at 12 months, the stabilities are almost monotonically increasing, but those at 6 and 9 months do not fit the pattern. Also, starting at 12 months, there are two or more stabilities available until ages 8 and 9. The agreement is encouraging.

In the research by Humphreys and Parsons (1979) previously mentioned for the one occasion, grades 7 to 9 not influenced by the arbitrary fixing of terminal reliabilities, the stability of canonical composites not corrected for shrinkage was .968. We have since obtained the stability of unit-weighted composites. This is .956. To compare this value with those in Table 3, it is necessary to obtain the square root, namely, .978. This stability can be applied between 13 and 14 years and between 14 and 15 years. It appears to be a reasonable extrapolation of the estimates in Table 3.

If the defects in the data of Humphreys et al. (1985) merely added random noise so that a good fit could not be obtained, it would still be possible for the year-to-year stability estimates to be reasonably accurate. With the exception of the stability of intelligence from age 8 to 9, the remaining stabilities tend to be only trivially smaller than the ones in Humphreys and Parsons (1979) and are also reasonable extrapolations from the present data.

Several considerations place limits on the size of the subjective confidence intervals around these stability estimates. In the first place, the stabilities cannot approach unity in the later ages much more closely and still be congruent with the correlations observed for more remote occasions. Secondly, true-score stabilities cannot be lower than observed correlations. For the later ages, again, this places a lower bound on the stability that is not far from our estimates.

This reasoning is especially cogent in the data of Humphreys and Parsons (1979). The reliabilities of their unit-weighted composite were .96 in grade 7 and .94 in grade 9. It takes a very modest decrease in these reliabilities to push

TABLE 3
Estimates of Reliabilities and Adjacent-occasion True-score Stabilities in Three Age Ranges[a]

	Reliabilities			Stabilities[b]		
	3 months to 9 years	9 months to 9 years	18 months to 15 years	3 months to 9 years	9 months to 9 years	18 months to 15 years
6 months	67			89		
9 months	63			87		
12 months	71	75		74	72	
18 months	83	83		82	82	
24 months	79	81	81	94	92	91
30 months	78	78	77	94	95	96
36 months	85	87	87	90	90	91
4 years	81	83	84	95	93	92
5 years	89	90	89	96	96	96
6 years	93	93	92	94	94	95
7 years	88	88	87	98	98	96
8 years	90	91	92			97
9 years			95			

[a]Values are not entered for the initial and final occasions in each analysis because they could not freely be determined.

[b]Each entry represents the estimated correlation with the occasion immediately following.

the estimate of the true-score stability over 1.00. If there were no measurement error in the composites, on the other hand, the estimate of the true-score stability over 2 years would be .91. Even though the simplex model may not be precisely true, it cannot be far wrong.

DISCUSSION

An important conclusion from these data does not depend on model fitting. Within the range of ages studied there is no indication from the observed correlations that relative intelligence (IQ) has become stable within individuals. Correlations do become larger over the same amount of elapsed time during development, but there is no break in the continuity of change. Only the rate changes. A conclusion that tests become measures of intelligence at some point in time, whereas they did not measure intelligence earlier, is not possible.

Infant tests do differ in content from later tests, but the correlations indicate that change continues when test content becomes stable. During the school years, intelligence shows a degree of stability that, on a short-term basis, can be confused with the stability expected of a fixed trait; but, on a long-term basis, an annual stability in the mid-'90s allows for a great deal of change in the position of individuals in their relative intelligence (IQs).

As a matter of fact, the genetic and environmental substrates for intelligence

provide for change, not for stability. It is widely assumed that the genetic substrate is polygenic. The large number of genes involved do not fire simultaneously at the moment of conception to produce a unitary entity. Instead, it is probable that the multiple genes fire at different times during development and do so, independently. The environmental substrate is also a complex of many determinants that impinge on the organism at various times during development. To some extent, the multiple environmental determinants are correlated through the mechanism of parental social status, but there is also a good deal of independence. Parental status has only a little to do with the ability of a given teacher and even less to do with that teacher's personality. There is little basis in either substrate for the conception of intelligence as a stable entity.

The reasonably good fit of the model between 12 months and 9 years allows a second conclusion. The hypothesis that tests designed to measure individual differences in cognitive development are measuring the general factor of intelligence at all ages within this range is a viable alternative to the conventional wisdom. Change does take place in relative standing on the general factor, but that change is continuous. Change is more rapid in the preschool years, but intelligence over the age span designated can be considered growth along the same dimension.

There is a simple explanation available for early rapid change and later slower change. Humphreys (1971) defined intelligence as the repertoire of knowledge and skills falling within the cognitive domain available to the person at a particular point in time. (The domain is defined by consensus among specialists in the area.) An infant's repertoire is relatively small. Increments to the repertoire that have near zero correlations with the initial base will change the ordering of individuals on measures of the repertoire quite rapidly. Increments are to larger and larger bases during the preschool period, thus producing the increasing stability from one occasion to another. The almost flat level of stabilities during the school years seemingly requires increasing variance of the increments or a small change in the correlation between increments and bases.

There is also continuity in the methodology of test construction from infancy to adulthood. One starts with a wide variety of problem-solving items appropriate to the level of development of the population to be tested. Items are selected that show a steep increase in proportion passing with increases in chronological age and substantial positive correlations with other similar items. Tests produced using these criteria will define a general factor. If the items in infant tests of development are good measures of a general factor among problem-solving items at a given stage of development, they also meet the objective criteria for intelligence test items. The test those items define does not have to be highly correlated with measures of a general factor several years down the road.

The general factor early in development may be more general than it becomes later. A personal communication from Lipsitt (October, 1985) contained unpublished data of Lipsitt and Buka for the Bayley mental and motor scores on

almost 3000 cases. The correlation of the two scores with each other is .611 at 8 months and the correlation of each with the Stanford–Binet at 4 years is .214 and .224, respectively. Correlations drop to .162 and .169 with Wechsler Verbal at 7 years and are .192 and .226 with Wechsler Total at the same age.

Threats to the Model

The observed correlations of 3 and 6 months with subsequent occasions constitute a threat to the validity of the model if those correlations are replicable. The model cannot tolerate correlations with those ages that are lower at 4 years than at 6, let alone 9. Because the correlations are based on maximum *n*s that typically involve both members of the twin pair, a defensible test of statistical significance is not possible. Our professional judgment, based on an analysis of several limiting cases, is that the differences are not large enough to allow rejection of the null hypothesis.

The current interest in developmental psychology in measures of recognition memory of the sort pioneered by Fagan (1974) constitutes a second threat. Relatively large correlations have been reported on these measures obtained within the first few months and measures of intelligence several years later. These correlations are based on the small samples endemic in a great deal of developmental psychology. Small correlations tend not to be reported when a new experimental paradigm is being rapidly exploited and enthusiastically received.

The requirements for refutations of the model in any age range are simply stated. One finds a well-defined general factor among problem-solving type items at a particular age that research workers agree are appropriate to that age. A new measure administered at that same age has a relatively small correlation with estimates of scores on the general factor at that age but relatively high correlation with the general-factor score estimates obtained at a later age. Because so much intellectual change is gradual, large samples will be required to establish statistical significance. Approximately equal correlations with the general-factor estimates at two different ages that are based on small samples are not adequate.

Recommendations for Research

The research just described that is required to reject the hypothesis of continuous change in intellectual development is the sort of research that should be widely undertaken even if the simplex model is rejected on the first try. The best theory, in our opinion, is derived from dependable data. There must be an adequate empirical base of what goes with what. Suppose that individual differences in attentive behavior during the first 2 months are more highly correlated with the Stanford–Binet at 4 years than with the Bayley at 12 months. How large will the correlation be with an equivalent intelligence test at 6 or 16 years?

If one accepts the existence of a "real" intelligence that is measured somewhat fallibly at all ages, but especially so during infancy, it is necessary to

specify an age in which the measurement is least fallible. It is clear that there is no age at which a standard test becomes a stable measure of this construct. An infant test at 12 months is a very imperfect predictor of scores on a standard test of intelligence at 6 years, but the test at 6 years is very imperfect as a predictor of intelligence at age 18. Is an intelligence test in early adulthood the least fallible measure of the hypothetical real intelligence with earlier tests showing a gradual progression in the validity with which this construct is measured?

These contrasting conceptualizations can be tested against each other in a simple design. Surely, the earliest evidence of what is measured by later standard tests of intelligence is aurly comprehension of language. Let this variable, perhaps measured by accuracy in following oral directions at 18 months, be the criterion. Will an infant test of development at 12 months be more highly correlated with the criterion than a standard test of intelligence administered at 17 years? The answer seems so obvious that we may be accused of setting up a straw man, but this seems to be a necessary consequence of the construct of a real intelligence.

It is more realistic to design studies in which the correlates of measures of the general factor at various points during development are determined. One can predict confidently that comprehension at 18 months will be more highly correlated with a measure of the general factor at 24 than at 12 months, simply because individual differences are becoming more stable. At what point in later development will the correlation between aural comprehension at 18 months and a measure of intelligence drop below the correlation obtained predictively at 12 months?

It would also be informative to use the predictive–postdictive design in the early grades. At the college level, Humphreys (1960) showed that the intercorrelations of independently computed semester grades, controlling for enrollment within colleges of the university, formed quasi-simplex matrices. Later (1968) he replicated the quasi-simplex form and showed that college entrance tests and high school rank in class showed a sharp decline in predictive accuracy from the first to the eighth semester, controlling for the reduction in range of talent caused by selective dropping out. Still later, Humphreys and Taber (1973) showed that the Graduate Record Examination Verbal and Quantitative scores obtained in the senior year in college reproduced postidictively the predictive validities of the American College Test administered at the end of high school. Senior tests were most highly correlated with freshman grades, not with senior grades.

At the college level, the changes producing the quasi simplex for grades are not those that are presumably producing relative change, perhaps at a reduced rate as compared to earlier years, in college "aptitude" test scores. The GRE advanced tests, however, show the largest correlations with sophomore and junior grades. Apparently, opportunities for specialization in college make senior grades less dependent on general intelligence. At the grade school level, however, we would predict that third-grade intelligence test scores would be more

highly correlated with third- than with sixth-grade achievement and sixth-grade test scores would be more highly correlated with sixth- than with third-grade achievement. Educational experiences are broader, with more even access for the students, in primary education than in college. The expectation, therefore, is that changes in intelligence test scores would go hand in hand with changes in academic achievement.

When viewing the development of abilities from this point of view, one looks for the correlates and for the changes in correlates of intelligence during development. Each of the 16 component tests of the intellectual composite used by Humphreys and Parsons (1979) forms a quasi-simplex matrix between grades 5 and 11 (Humphreys, Parsons, & Park, 1979). When one of these, a measure of aural comprehension, is correlated with a composite of the remaining 15, the cross-correlations suggest that individual differences in aural comprehension anticipate changes in individual differences on the composite (Humphreys & Parsons, 1979). Similarly, individual differences in height and 8 and 9 years anticipate changes in individual differences in intelligence for girls at 11 and 12 years (Humphreys et al., 1985).

Some of Cattell's hypotheses (1971) concerning differences between fluid and crystallized intelligence can be tested by the same design. Broad measures of both will show the quasi-simplex form. Will fluid intelligence be more stable than crystallized? Will individual differences in fluid intelligence anticipate later individual differences in crystallized?

A good deal of evidence suggests that changes in relative intelligence are not completely Markovian although there is a good deal of randomness involved in intellectual growth. For one thing, Wilson (1983) has shown that the expected relationships between the two types of twins are gradually established. Anticipation of individual differences in intelligence by measures of aural comprehension and height also suggest nonrandomness. Beliefs in the stability of intelligence beyond the preschool period, however, have deterred research on the prediction of change.

CONCLUSIONS

We conclude from these analyses that continuity in intellectual development from 12 months to 9 years along a dimension appropriately called general intelligence cannot presently be rejected. The noticeable differences in content between infant tests of development at 12 months and later tests of intelligence are not in themselves sufficient to reject this conception. Rapidity of change of relative intelligence in infancy is explained by the addition of sizable increments of skill and knowledge to initially small cognitive repertoires over short periods of time.

The simplex model assumes that true-score gains are independent of the true-score base at the beginning of a given period of time. This assumption is cer-

tainly not precisely true, but the model still provides a reasonably accurate approximation to the intercorrelations of measures of the general factor over numerous occasions during development. The obtained correlation in the Louisville data between 12 months and 8 years is .32, and the expected value is .313. This is derived from an expected correlation between true scores over that interval of time of .379 multiplied by the square root of the product of the two estimate reliabilities.

If the model is approximately accurate, one should be able to extrapolate to longer time intervals. McCall (1979) reports a median correlation between infant tests administered from 7 to 12 months and intelligence tests between 8 and 18 years of .26. If we assume that true-score stabilities level off at about .97 beyond 8 years, the correlations between 8 and subsequent years will drop slowly to a true-score correlation of .28 and 18. The expected amount of measurement error reduces this value to .23. The model nicely brackets McCall's obtained correlation.

The relatively small amount of change in the prediction of later IQs from 12 months over the 10-year period from 8 to 18 cannot be interpreted, however, as relative intelligence becoming almost stable. Using the same values of adjacent occasion stabilities, the expected true-score correlation between IQs at 8 and 18 years becomes .74. If reliability at 18 years is equal to the .91 at 8, the expected observed correlation is .67. There is approximately 50% of common variance between 8 and 18 years predicted by the model.

REFERENCES

Anderson, J.E. (1939). The limitations of infant and preschool tests in the measurement of intelligence. *Journal of Psychology, 3,* 351–379.

Anderson, J.E. (1940). The prediction of terminal intelligence from infant and preschool tests. In G.M. Whipple (Ed.), *Intelligence: Its nature and nurture, Part 1. The 39th Yearbook of the National Society for the Study of Education,* Bloomington, IL.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *Journal of Genetic Psychology, 25,* 165–196.

Cattell, R.B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Fagan, J.F. (1974). Infant recognition memory: The effects of length of familiarization and type of discrimination task. *Child Development, 45,* 351–356.

Fleishman, E.A. (1960). On the relation between abilities, learning and human performance. *Journal of Experimental Psychology, 60,* 162–172.

Fleishman, E.A., & Hempel, W.E., Jr. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika, 19,* 239–252.

Fleishman, E.A., & Hempel, W.E., Jr. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology, 49,* 301–316.

Fleishman, E.A., & Rich, S. (1963). Role of kinesthetic and spatial–visual abilities in perceptual-motor learning. *Journal of Experimental Psychology, 66,* 6–11.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, IL: Free Press.

Humphreys, L.G. (1960). Investigations of the simplex. *Psychometrika, 25,* 475–483.

Humphreys, L.G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology, 59,* 375–380.

Humphreys, L.G. (1971). Theory of intelligence. In R. Cancro (Ed.), *Intelligence: Genetic and environmental influences* (pp. 31–42). New York: Grune & Stratton.

Humphreys, L.G., Davey, T.C., & Park, R.K. (1985). Longitudinal correlation analysis of standing height and intelligence. *Child Development, 56,* 1465–1478.

Humphreys, L.G., & Parsons, C.K. (1979). A simplex process model for describing differences between cross-lagged correlations. *Psychological Bulletin, 86,* 325–334.

Humphreys, L.G., Parsons, C.K., & Park, R.K. (1979). Application of a simplex process model to six years of cognitive development in four demographic groups. *Applied Psychological Measurement, 3,* 51–64.

Humphreys, L.G., & Taber, T. (1973). Postdiction study of the Graduate Record Examination and eight semesters of college grades. *Journal of Educational Measurement, 10,* 179–229.

Joreskog, K.G., & Sorbom, D. (1978). *LISREL IV, a general computer program for the estimation of linear structural equation systems by maximum likelihood methods.* University of Uppsala, Department of Statistics, Uppsala, Sweden.

McCall, R.B. (1979). Stability–instability of individual differences in mental performance. In J.D. Osofsky (Ed.), *Handbook of infant development* (pp. 707–741). New York: Wiley.

McCall, R.B., Hogarty, P.S., & Hurlburt, N. (1972). Transitions in infant sensori-motor development and the prediction of childhood IQ. *American Psychologist, 27,* 728–746.

Roff, M. (1941). A statistical study of the development of intelligence test performance. *Journal of Psychology, 11,* 371–386.

Wilson, R.S. (1983). The Louisville twin study: Development synchronies in behavior. *Child Development, 54,* 198–216.