Mr. Binet’s Test 70 Years Later

ROBERT L. THORNDIKE
Teachers College, Columbia University

Unlike a human infant, a psychometric device has no clearly defined and annually celebrated birthday. The moment of conception is unspecified, and attended by no particular ecstasy; the period of gestation is variable and ill-defined; and the date of parturition is usually unknown. So in looking for a birthdate for the Binet Intelligence Scales, we have some room for maneuver. However, a reasonable case can be made for 1905, the year in which L’Annee Psychologique published the account of a workable version of an individual intelligence test. And so, we may define 1975 as the 70th anniversary of that event—an event that, as much as any in the early years of this century, determined the shape and course of subsequent psychological and educational research, and of psychological and possibly educational practice. For my remarks today, I have chosen to look at a few aspects of that 70-year career.

That first set of tasks proposed by Binet stemmed from no elegant theory, nor from any precise definition of intelligence. Binet made no assumptions about the causes or origins of the intellectual deficits that he proposed to study (for his initial concern was primarily with intellectual deficiency), nor about their prognostic significance for future development. He undertook to describe objectively, in the here and now, levels of development of what he described in a very general way as “judgment.” The basis of his selection of tasks was very largely empirical—try-out with groups of children of various ages in regular school, and comparison of their performance with that of those who were in institutions for the mentally retarded.

In his own country of France, Binet’s efforts appear to have been received with what our late colleague Alexander Wesman used to describe as “modified rapture,” but in the United States the rapture was more immediate and complete. By 1916 there were at least three translations and adaptations of Binet’s scales—one by Goddard at the Vineland Training School, one by Kuhlmann at Minnesota, and one by Terman at Stanford—but of these only Terman’s Stanford-Binet has survived to the present day. It has not only survived but flourished. According to figures provided by the publisher on sales of the test record form, Terman’s version appears to have been administered to about 150,000 persons a year during its first incarnation from 1916 to 1937, to about 500,000 a year during the reincarnation from 1937 to 1960, and to about 800,000 a year during the third coming from 1960 to 1972.

It is interesting to speculate why this version won out in the competition. Was it because it was a better test, better standardized, with a more discriminating, more administrable set of tasks? Was it that it provided adequate “top” so that it was able to assess the ability not only of the deficient but also the intellectually gifted? Or was it due to Terman’s professional reputation and personal enthusiasm, together with his massive and widely publicized “Genetic Studies of Genius” in which he followed over 1,000 very bright California youngsters through 35 years of their lives and all that remained of his?

Be that as it may, it was the Stanford-Binet that survived, that was revised in 1937 and again modestly in 1960, that was re-normed in 1972, and that has been for most of the past 60 years the workhorse of psychometric appraisal of cognitive development,
the standard against which other tests of cognitive abilities have been evaluated, and more recently a prime target for the social critics of ability testing.

Binet was concerned with an over-all appraisal of mental functioning. It was not that he failed to recognize that functions were involved other than the “judgment” to which he accorded a central role—functions of perception, discrimination and memory—and that each might be separately evaluated. It was rather that he felt it most important to provide a unitary over-all appraisal of level of mental functioning, and believed that a judicious pooling of a variety of different tasks provided the best basis for such an appraisal. This unitary over-all appraisal has been carried forward to the present as the central feature of the successive versions of the Stanford-Binet—the lineal descendant of Binet’s original efforts.

A good deal of test development in the past 40 years has tended to move away from single-score composite measures, just as factor analyses of abilities have moved their focus away from Spearman’s original conception of a general intellectual g factor of perceiving relationships. But it is worth standing back for a moment and asking whether the latter-day urge to fractionate abilities into smaller and smaller splinters always represents either sound theory or effective practice.

Since the basic pool of exercises used in the Binet includes tasks of a spatial nature as well as verbal tasks, tasks calling for short term retention as well as tasks demanding the application of more remote past experiences, it is possible both rationally and statistically to identify several components in the Binet score. Factor analyses of the exercises of the 1937 and 1960 versions of the Binet, starting with McNemar’s original 1942 analyses of the revised test, have generated multiple factors that have been more or less interpretable. But it is also true that in these analyses a very large proportion of the common-factor variance (approximately 80% in some analyses that we have recently carried out) is extracted by the first factor. This large common core is what gives meaning to a single score and support for Binet’s original notion of pooling performance on a diverse array of tasks into a single assessment of cognitive ability.

One feature of the over-all assessment by the Binet that has given it practical significance to users has been its stability in an individual over a span of years. My very first article published as a graduate student 42 years ago related that stability to the interval over which the forecast was made, and a host of subsequent analyses affirm the dependence of the stability on not only time interval but also the age level at initial testing. As we consider fractionating cognitive abilities into more narrowly defined bits and pieces, it is appropriate to ask whether those narrower abilities show comparable stability, and especially whether the individual patterning of strengths and weaknesses is maintained from one time to another, especially in young children.

This question is better attacked using those tests that are specifically designed to yield more than a single score, such as, for example, the Wechsler Intelligence Scale for Children.

Thanks to the kindness of Dr. Robert Osborne of the University of Georgia, who made available his data on repeated WISC testing of children in grades 1, 2, 4, and 6, we have been able to calculate the stability of the differences between various part scores on the WISC. Thus, one might look at the difference between Verbal and Performance IQ and undertake to interpret it in some way in relation to the potential or personality of a child. How stable is the difference score on whose basis such an interpretation is made? Comparing grade 1 with grade 6, Osborne’s data yield a correlation of 0.38.

Clinicians have from time to time been tempted to interpret even more specific aspects of patterning in Wechsler profiles. How stable are these in young children? We calculated the stability coefficients for all possible pairs of sub-test differences over this grade 1 to grade 6 time span. The correlations range from -0.08 to 0.39, with a median value of 0.15. These very modest relationships may be compared with the stability of WISC total IQ for these same children over the same time span of 0.79.

Of course, over shorter time spans and at later ages the stability of the differential measures is a bit better, as is that of total IQ, and the distinctions between some pairs of sub-scores may have value as a source of tentative hypotheses for the clinician studying a specific child. But the real stability of cognitive measures lies in the common core of general cognitive functioning. Especially with young children, Binet may have shown good judgment in focussing upon “judgment,” this common core.

In his early test series, Binet was content to express individual cognitive performance in rather crude age equivalents, reporting to the nearest year the developmental level of the children whom he tested. However, as the tests were taken over in the United States and developed by Terman and others the attempt was made to achieve greater precision, reporting mental ages in years and months rather than in simple year units, and the ratio concept proposed by Stern in Germany was adopted as a form of score that displayed the individual’s progress relative to his own age group. Thus, the IQ came to occupy a central role in our conceptions of testing of cognitive performance, and is perpetuated in our language to this day in that somewhat unhappy phrase “IQ tests.” Next I will consider the wandering IQ in the special context of Terman’s development of the Binet Scales.

In his 1916 test, Terman, by a judicious assignment of tasks to age levels and specification of scoring standards for evaluating performance, arranged things so that the average mental age on his test series would correspond to chronological age at each age level. Thus, the average IQ was made to remain firmly at 100, and appeared to do so rather consistently over the whole...
range of ages at least from 6 to 14. Beyond that age range, the 1916 test was not really adequately normed and it had inadequate top to tap the abilities of older children of high ability.

It so happened, not by any special plan or design, that the standard deviation of these 1916 IQ's was approximately twelve points, so that the middle fifty percent of youngsters fell between roughly 92 and 108 on this ratio index. Fortunately, this spread of scores turned out to be quite uniform over the age range, so that a given IQ value could be interpreted in much the same terms whether the individual was six years old or ten years old or twelve years old. Thus, the IQ became in effect a type of standard score with a mean of 100 and a standard deviation of twelve.

When the 1937 revision was produced, once again the tasks were chosen and scoring adjusted so as to yield an average IQ of approximately 100 at each age level in a representative population sample. The equivalence of the mean IQ's on the 1916 and 1937 versions of the test was fairly well documented by a study by Merrill in which some 1500 youngsters who had previously taken the 1916 test were tested with the 1937 version, and this is supported by the careful national surveys carried out in Scotland. To a very close approximation, on average the individual who earned an IQ of 100 on the earlier test also earned an IQ of 100 on the later test. However, for reasons which have never been quite clear to me, the standard deviation of the new test in the US was no longer twelve but somewhere in the range of 16 to 17. The spread of scores was increased by about a third. In the Scottish surveys the comparable figures were 15 and 20.

I've brooded a good deal about how this might have come to be. In both the early and late tests, the authors required that the items they retained in the test show an increase in percent passing from one age to the next. However, in both forms, attention was also paid to having test tasks that might reasonably be considered to have an intellectual component, and to including tasks that showed a substantial relationship to the total score for individuals of a given age. The last criterion corresponds to the conventional type of item analysis that has become completely familiar in test development over the past fifty years. It seems to me that there must have been a good deal more emphasis upon the item's relationship to total score within an age group in the development of the 1937 test than there had been in 1916. This, combined with a need to generate more than twice as many tasks in order to have two forms of the test, L and M, probably resulted in items that were less related to chronological age and that showed flatter profiles of improvement as one went from younger to older children. Items of this sort would tend to show a greater spread of score values for children of a single age and to yield the larger standard deviation of scores.

The shift in variability of scores would have only a modest impact on the typical youngster who fell fairly close to the national average. However, the impact upon individuals at the extremes would be quite marked. This can be illustrated by instances at both extremes. For example, in Terman's longitudinal study of very bright children, to which I made earlier reference, the average IQ of the original gifted group was approximately 150 based upon the 1916 form of the Binet. The average IQ for their children based upon the 1937 form of the Binet was approximately 132. The difference between 150 and 132 suggests a relatively modest regression towards average in the offspring generation. On the other hand, if one realizes that the 150 on the 1916 form was fully four standard deviations above the general population mean, and represents approximately the equivalent of 165 on the 1937 form, the regression from the highly selected parent generation to their offspring becomes considerably more marked, the children being only about one half as extreme as their gifted parents.

At the other end of the scale the Binet critical score for defining mental deficiency had been set on the 1916 test as approximately an IQ of 70, or minus two and a half standard deviations. If one translates that into the scale of the 1937 Binet, it corresponds approximately to an IQ of 60, and educational and legal interpretations would need to be adjusted to correspond. I suspect that they were in part, but I am not sure that they ever were completely.

The 1960 edition was a consolidation of the two forms developed in 1937, where the consolidation took account of data accumulated during the 1950's but the basic standardization population was the same one that had been used in 1937. In 1960, frank recognition was made of the fact that an IQ functions as a standard score, and the IQ equivalents of mental ages were adjusted so that at every chronological age level the mean would be 100 and the standard deviation would be 16. This adjustment also took account of the fact that the original 1937 standardization population had been somewhat disproportionately drawn from middle and upper socioeconomic levels and that the obtained mean IQ's in the 1937 standardization sample were running two or three points above the 100 level.

The final journey of this wandering IQ occurred in 1972 when the form of the test assembled in 1960 was re-standardized on a new and current sample of cases. The standardization procedure was somewhat different from the one that had been used previously. The 1972 sample was based on the year before standardization of the Cognitive Abilities Test on a large national sample, rather than being chosen to proportionally represent socioeconomic categories. A stratified sample of children was selected to be representative of those who had been included in the 20,000 per-grade group-test standardization. That child and/or one or more siblings were tested on the Binet. The sample was so chosen that at each age level the deviation IQ's of the group-tested siblings had a mean of 100 and a standard deviation of 16. If we can assume that the large-scale group test standardization was appropriately carried out, then it would seem that by this procedure the individual tests should
have been rather effectively anchored to a much larger base population than could possibly be tested directly with an individual test. Since the correlation between the group test and the Binet was of the order of .70, a good deal of control over the characteristics of the Binet sample was achieved.

The first step was to look at the distribution of IQ's based upon the 1960 standard scores. These, it should be remembered, had been based upon the standardization testing for Form L and Form M in the early 1930's.

The shifts were fairly dramatic. At the preschool ages, the average deviation IQ of the 1972 sample was running close to 110. This then, gradually dropped off to an IQ of about 102 at age 10 and gradually crept up again to 105 or 106 by the adolescent years. What makes these higher average values even more impressive is that the sample in 1972 included ethnic minorities as well as the white group that was the only group tested in the 1930's. If one allows for this fact, the actual shift is probably two or three points greater than it appears on the surface.

The general rise in IQ level is not surprising. Such a rise has been reported by test-makers on all sorts of ability tests both in this country and abroad. The thing that is perhaps a little surprising is the size of the shift in the preschool group. On the other hand, perhaps this isn't surprising when one considers how life has changed for a preschooler between 1932 and 1972. Certainly, the amount of verbal and visual stimulation that the preschooler of 1970 was getting was enormously more than that available to the typical preschool-age child in the 1930's. The child of 1970 was probably watching television three or four hours a day. Furthermore, he had parents with two or three years more education, on the average, and had a much wider and more varied stock of books, toys, and other materials available to him.

The question we immediately faced with respect to these preschoolers was whether they represented the "Sesame Street" generation who were brighter than their older siblings, and would maintain their higher IQ's as they got older, or whether longitudinal data would reproduce the cross-sectional results with the drop that we observed in 1972. Thanks to the Spencer Foundation, we have now finished re-testing, after an interval of roughly three years, about 80% of the 750 children in the preschool norming group who were between three-and-one-half and six years old at the time of the original testing, and I can report in a general way what has happened to these youngsters over a three-year period. The best quick summary is that they have lost about one IQ point per year, so that in this group the drop from age 4 to age 6 would be about five points; and this would come close to the difference that was observed in the cross-sectional data. The greater change from the 1930's to the 1970's at the pre-school level appears to be a genuine phenomenon and not a deficiency in our sampling procedures, or a very recent bulge in measured ability.

Some have been inclined to point a finger of reproach at our schools and ask: Why isn't the preschool acceleration maintained through the school years? Perhaps, instead, the finger should be pointed at our forebears with the question: Why didn't you provide a stimulating world for your toddlers? But more realistically, we should recognize that with TV, the world of the child has changed, and that the beneficial aspects of that change are experienced primarily, perhaps even exclusively, in the preschool years.

These re-norming results are both intellectually interesting and practically significant. They indicate the extent to which the total impact of a changing culture can have an influence at least on the test performance that youngsters display, and probably on their basic cognitive functioning, and they reinforce our recognition that the tests that we have produced measure developed abilities, developed in a particular cultural setting.

Of course, individual differences persist in the midst of this general shift, and the variability of IQ's in 1972 has remained at least as great as that established in 1960, based upon testing carried out in the period from 1932 to 1937. That is, the changes that have produced a general elevation of abilities have not reduced individual differences, but the current culture has maintained them at least at the level that they displayed in the earlier period.

We were interested to inquire whether the shifts that have taken place over the forty-year period have affected the test items uniformly, or whether some systematic and meaningful description could be made of the items that show large changes and the items that show small ones. Dr. Robin Garfinkle has just finished a dissertation on this topic based upon the preschool results, and rather to our surprise the largest shifts in the direction of items becoming easier have taken place in the non-verbal, pictorial, perceptual, and memory items, rather than the ones that involve a good deal of semantic content. This phenomenon appears clearly only at the earlier age levels, that is, from about age three to about age four-and-one-half, but it suggests that the impact of cultural change on the preschooler has not been primarily and fundamentally a verbal impact, but has been at least as much a visual and perceptual enrichment.

Binet developed his scales 70 years ago in a France that was ethnically and culturally relatively homogeneous for a school system generally elitist in its orientation and unsympathetic and unresponsive to individual variability. In that setting he sought a procedure that could add objectivity and impartiality to the identification of children to be screened out of the regular school system for more appropriate placement in special classes. When Terman adapted the scales for use in this country, completion of high school was still the exception rather than the rule here, and schooling through most of the country was oriented toward the white English-speaking majority. Even in the 1930's it seemed to Terman more important to have a clearly defined all-white norming group than to have representation of the totality.
of the school-age population.

The United States of 1975 is different from France in 1905—or even the United States of 1930. We recognize and try to adapt to our ethnic and cultural pluralism, though not in full accord as to how best to do it. We undertake to provide education for all citizens at least through the secondary school years, and in many locations through some type of post-secondary program. We do this in a world in which the only thing that is constant is change at an ever-increasing rate. How must our conception and use of Mr. Binet’s test be modified for the educational scene of 1975?

Clearly, the normative interpretations must be adjusted to the changing times. But it seems equally clear that the prognostic interpretations must be adjusted to the particular place—that is, the setting and sub-culture in which a child has been reared and educated. Though the temporal change in a culture may facilitate test performance, sub-cultural differences may inhibit that performance. As a consequence, any rigid specification of level of test performance as the basis for decision or action—whether specified by law or by administrative ruling—seems unwise and perhaps pernicious. Binet’s test—or any other—must guide, and not replace informed judgment.

Accepting the goal of maximizing the effectiveness of education for all children and youth, we must face up to the problem that we have long acknowledged but seldom deal with effectively—the problem of providing for each individual the educational treatment that will be most effective in developing that person’s potential. A good measure of scholastic aptitude is not automatically a good guide to the optimal educational treatment. Binet’s test, like others used in education, must be judged in terms of its ability to facilitate constructive adaptations of educational programs for individuals. This is the challenge for the next 70 years.

Notes


Federal Perspectives on Education

GILBERT R. AUSTIN
University of Maryland

The economy is the policy issue on which the federal government is seeking information today, according to a panel of congressional aides and HEW officials speaking during the recent AERA Annual Meeting in Washington. Whether educational researchers can make a contribution to economic policy formulation was not clear, however.

The symposium from which this consensus emerged debated “Federal Perspectives on Educational Research.” Members of the panel were Robert Andringa, House Committee on Education and Labor; Lila Carroll, NIE; Harley Dirks, Senate Subcommittee on Appropriations for Labor and HEW; John W. Evans, USOE, Office of Program Planning, Budgeting, and Evaluation; Allen Ginsberg, Office of the Secretary of HEW; and Roy Millenson, Senate Committee on Labor and Public Welfare. Reacting to the panel were Robert Emans, University of Maryland, and Michael Timpane, Rand Corporation.

Other pressing policy issues on which the federal government is seeking information are: the study of income maintenance and how to get cash benefits to low-income people; long-range school finance and equalization of expenditures; the question of desegregation and a need to synthesize what we have learned as a result of ten years of efforts; and the problem of helping the disadvantaged which now is considered to include physically handicapped children.

In a number of instances, the panelists recommended that the educational research community give more attention to the re-analysis of completed research. They also warned that both educational and biomedical research would probably have to take a back seat to the more pressing problem of how to get the economy back on its feet.

One question asked of the panel concerned how priorities are set for educational R&D, especially with respect to applied vs basic research, the funding of individuals vs institutions, and a balance between fed-