A Century of Ability Testing

Robert M. Thorndike
Western Washington University

with **David F. Lohman**
The University of Iowa

The Riverside Publishing Company
Chicago

Chapter–opener photo credits: Pages xii, 40, 62, Historical Pictures Service, Chicago

ABCDEFGHIJ-FFG-93210/89

ISBN: 0–8292–5156–1

Printed in the United States of America

To ELT, RLT, and EST,
each of whom has served as
an inspiration for me in many ways

Preface

There are many reviews of the several revisions of the Binet–type scales. Reviews of the earlier revisions are available in a number of other places, including Freeman (1962) and various editions of the measurement texts by Anastasi and Cronbach. Although it is too soon to render a verdict on the most recent revision of the Stanford–Binet, a comprehensive comparison of the Fourth Edition with its predecessors is available in Sattler (1988). The purpose of this book is not to present a detailed account of the Stanford–Binet, or any other intelligence test. Instead, at the request of The Riverside Publishing Company, I set out to trace some of the historical forces that shaped the development of the measurement of intelligence, in particular as it has been defined by the Stanford–Binet.

For the last 25 years, which roughly spans my career in psychology, psychological testing has been under more or less continual attack, often for events that took place in quite a different time and climate. Because members of my family have been involved in the measurement of human abilities and other characteristics almost from the beginning of the enterprise, they have been part of the conflict. The research that went into this little book has at times taken the form of a personal voyage of discovery into my past. I found out many things about E. L. Thorndike, my grandfather, that I had not known, and on occasion I tried to bring an element of familial perspective into the narrative. But more than anything, this book tries to understand and explain the course of development of the measurement of intelligence within the changing context of the times.

Many of the prominent figures of the early phases of the testing movement in the United States have been portrayed by some recent writers as heartless fiends whose only goal in life was to persecute the less fortunate. It would be hard to distort their motives more. These were men and occasionally women who went about their business with the best of scientific intentions in an era when many people expected science to solve all the world's problems within the next few years. They were not completely successful, but that was not for lack of effort. And, as the record of psychometrics in educational, industrial, and military affairs will attest, tests have been successfully used to reduce bias and improve efficiency in meaningful ways. That bias and prejudice may still exist and may enter into the use of tests is more a function of the society as a whole than it is of the testing movement.

There are several areas of controversy that I have tried to avoid or on which I have withheld judgment. In this book I do not take any position on the fundamental nature of intelligence. At some points I note that the evidence or a

popular theory seems to be heading in one direction or another, but it is not my intent to define intelligence or to take sides on such issues as the heritability of intelligence.

Tests merely provide operational definitions of intelligence; they do not reveal anything more basic about it. They are systematic ways of collecting samples of behavior which may be compared to other samples of behavior. When proper precautions have been taken, certain kinds of conclusions or predictions may be justified based on observed regularities. As is always the case when dealing with individual differences among human beings, those predictions or conclusions may be in error to a greater or lesser degree. With the exception of some of the new research on cognitive science and intelligence, which David Lohman has summarized well in chapter 6, testing is a correlational enterprise, and causal inferences are hazardous at best. Of course, some explanations account for observed relationships better than others do, but theories of intelligence are generally descriptive rather than deductive.

The book is divided into six chapters that cover periods of varying duration and themes of varying scope. The first chapter sets the stage and attempts to show how Binet's discovery was a logical outgrowth of his own inventiveness and persistence and the intellectual direction of the times. Methods to measure intelligence were about to be discovered, and Binet was the right man in the right place at the right time. Chapter 2 describes the work going on in the United States and England between 1900 and the start of World War I, including the development of the first Stanford revision of Binet's scale. Chapter 3 covers events surrounding the war and some of the controversies that publications from the army testing program created. Chapter 4 gives a brief review of some of the arguments about the nature of intelligence that made the journals of the period so interesting. The debates that centered on Spearman's two–factor theory of intelligence take center stage, along with the development of factor analysis. This chapter ends with the 1937 revision of the Stanford–Binet and the introduction of the Wechsler scales.

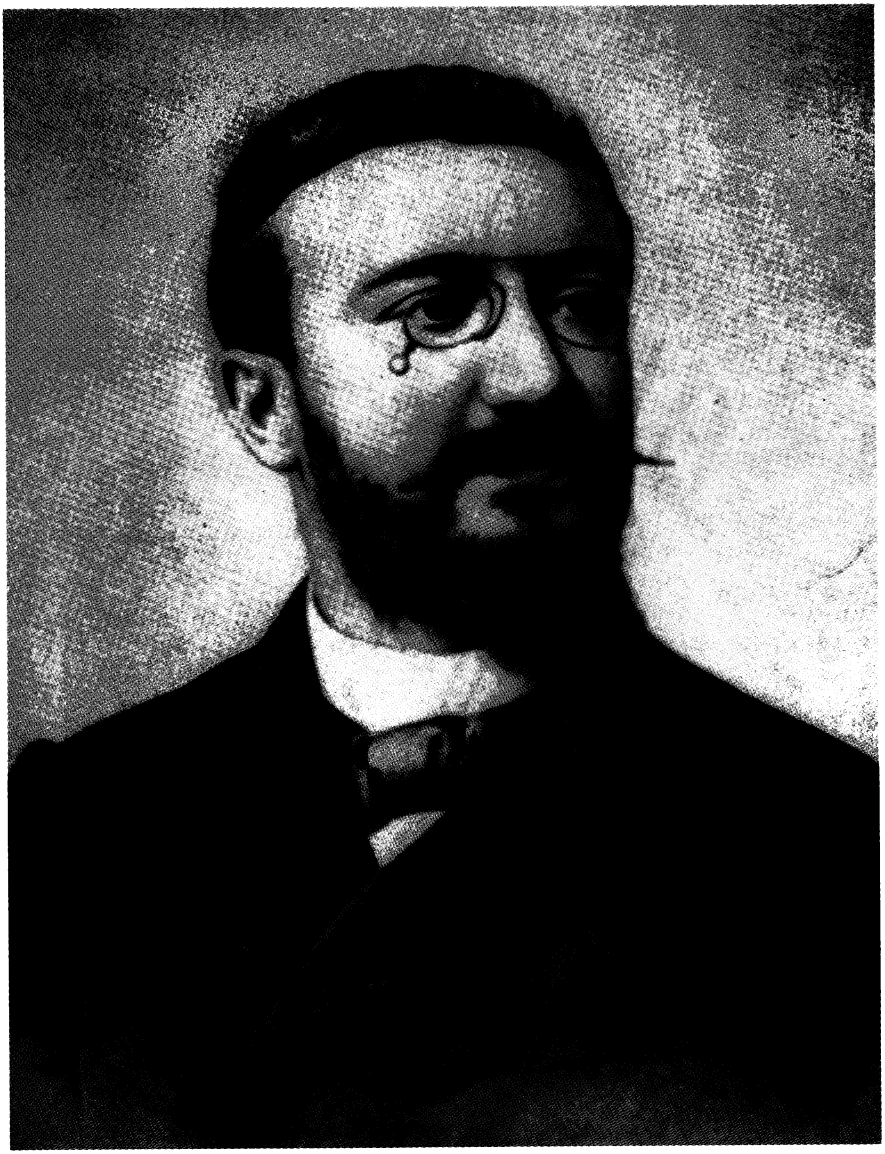
The first four chapters cover the first 50 years of ability testing, while chapter 5 covers the next 50. This rather uneven distribution of coverage is due to an equally uneven rate of development. As Oscar Buros observed in reviewing his 50 years of activity in the field, not much of note has happened in the development of psychological measurement in this later period. World War II, the development of test batteries, introduction of the third and fourth revisions of the Stanford–Binet, and the developing concern about bias in tests and testing practices form this chapter. Finally, David Lohman presents in chapter 6 a brief overview of the most recent developments in the search to understand what intelligent behavior involves and how it can be understood. This chapter reviews

the links between tests and the rapidly developing field of cognitive psychology or cognitive science. As Dr. Lohman points out, the cognitive revolution in psychology has returned intelligence to a central role in the study of human behavior, and this renewed interest and fresh perspective may well result in significant advances in both the theory and measurement of intelligence in the future.

I would like to express my thanks to The Riverside Publishing Company for suggesting this project to me. It is something I had in the back of my mind to undertake sometime, but their encouragement made sometime now and they gave me free rein to develop the topic as I saw fit. They also introduced me to Dave Lohman, who brought his very considerable knowledge of the fields of cognitive psychology and testing to the book in the form of chapter 6. I would also like to thank Western Washington University, which provided professional leave for me to work on this book and the facilities for me to finish it. A special thank you also goes out to Dr. Lloyd M. Dunn who donated a copy of Terman's personal copy of the Kite translation of Binet's major papers. His gift to Western's library arrived while I was in the middle of the research, and I fell on the book immediately. Dr. John Richardson of Western's Sociology Department and my father, Robert L. Thorndike, both read late drafts of the manuscript and offered valuable corrections and suggestions. My colleagues in the Psychology Department at Western have been willing to listen to my endless "discoveries" and to give me moral support; and, my wife Elva has once again born the inevitable and unenviable burden of reading (and correcting) my efforts with patience, skill, and grace, all the while tolerating the artist's temperament. This book has been fun for me to write, and I hope it will be enjoyable for others to read.

Robert M. Thorndike
Bellingham, Washington
June 1989

Contents



Alfred Binet (1857–1911)

1

Beginnings of Mental Testing

A UNIVERSAL ENDEAVOR

As is the case with most other areas of intellectual endeavor, it is not really possible to identify a particular date or event that marks the beginning of intelligence testing. DuBois (1970) described a "test–dominated society" in China that had its origins around 2200 B.C. Tests of proficiency in various subjects, such as archery and poetry, were used as measures of one's education and general ability. Development of such tests continued over the centuries as part of the Chinese civil service system. Cheng (1922) noted that "the true–false test, the ingenuity test, the picture–completion test, the verbal question test, all have fore–runners" in the period of the Han and Wei dynasties around the dawn of the Christian era (p. 7). Other authors such as Peterson (1925) and Pintner (1931) have described the periodic rise and fall of concern with individual differences in ability from the time of the Greeks to that of their intellectual descendants in Europe and North America. Virtually every society has been aware of individual differences in talent of some sort among its members and has developed ways of identifying those differences through the assignment of rank or role in the society.

"Mental Test"

Cattell's paper. To begin our study of the history of intelligence measurement, I have (somewhat arbitrarily) chosen a paper by the pioneering American

psychologist James McKeen Cattell (1890) in which the term *mental test* appeared in print for the first time. Cattell's paper, in a sense, marks the beginning of the modern era of human abilities and personality measurement. In his paper, Cattell issued a call for the investigation of mental phenomena through the use of mental tests. He argued that the greatest progress would be realized if each of the psychological laboratories then being founded in the United States and elsewhere would apply a uniform set of measures to all subjects, using standard procedures, until a base–line pool of data was developed. He suggested ten tests—mostly measures of sensory sensitivity:

1. dynamometer pressure
2. rate of movement
3. sensation areas
4. pressure causing pain
5. least noticeable difference in weight
6. reaction time for sound
7. time for naming colors
8. bisection of a 50 cm line
9. judgment of 10 seconds time
10. number of letters remembered on one hearing

The selection of measures was consistent with the scientific thinking of the time, which held the notion (derived from the writings of John Locke) that the mind is a blank slate at birth, that it is written on by experience, and that the amount and quality of what is written depends on the quality of sensory input. Sir Francis Galton, with whom Cattell had worked following the completion of his doctorate in Wilhelm Wundt's Leipzig laboratory, had used similar measures in his South Kensington laboratory and had found a wide range of individual variation in these as well as in anthropometric variables. DuBois (1970) gave detailed descriptions of the tests suggested by Cattell.

Galton's analogy. Cattell's 1890 paper was published in the prestigious British journal *Mind* and was followed by a lengthy postscript written by Galton. In his comments Galton made two points that are almost as important as those made by Cattell in the main paper. He likened the mental tests to shafts being sunk at a few critical points; the results of these explorations would help determine the quality of the mind. His analogy is not dissimilar to the principles

of present testing practice. In addition, Galton noted the need for test validation, although he did not use the term. He pointed out that "the sets of measures should be compared with an independent estimate of the man's powers. We thus may learn which of the measures are most instructive" (Galton in Cattell, 1890, p. 380).

Of great importance in understanding any set of historical events is an appreciation of the social context of the times in which those events occurred. It is easy, with 20/20 hindsight, to see that Cattell's proposed tests would not measure what we now think of as intelligence. Likewise, we encounter people at later periods whose actions and attitudes may seem almost foolish or malicious by current standards. However, these people were not working in today's social and intellectual environment, and what they did and thought was generally appropriate to the climate of their times. It has been common sport in recent years for writers on the history of intelligence measurement to malign the attitudes and practices of the founders of mental testing because they do not conform to currently accepted norms. Unless fashions of thought stop evolving, the attitudes and beliefs of these contemporary writers of history will look as foolish and misguided to future generations as those of the past do to them today. Ethnocentrism has a temporal as well as a sociocultural dimension.

Converging Trends of Thought

Individual differences. In America and Europe in the late 1800s, three general lines of development converged into what became known as the mental testing movement. One of these lines, epitomized by the work of Galton and Cattell and carried on largely in England, concerned an interest in individual differences and their distribution in the human population. This interest led, through the work of Karl Pearson and Charles Spearman, to the discovery of statistical techniques for describing the extent of variation among individuals and covariation among variables. These statistical methods, while they did not have an impact on the very first modern measures of intelligence, were quickly brought into play in the analysis and comparison of tests. Later, complex statistical procedures assumed a major role in the design and construction of tests.

Educational reforms. The second line of development leading to the testing movement may be found in education. During the second half of the nineteenth century formal education became much more accessible than it had been. Universal compulsory education laws were passed in many countries, where previously only the children of the upper class (and to an increasing extent, the growing middle class) had been able to afford an education. This

brought individual differences in educability to the forefront of popular and academic consciousness. (When education was voluntary, differences in means, ability, and interest tended to select those who would continue their education.) Universal education also made it clear that formerly acceptable methods of academic selection would not work with such heterogeneous groups. At the same time, educational reformers were beginning to question the prevalent practices in education, and interest was being shown in educational research. Measures of educability were clearly needed.

Warner (1890) provides an interesting perspective on some of the then–current thinking about education. As a medical doctor lecturing to prospective teachers at Cambridge, the good Doctor Warner made such comments as "Good mental training diminishes the amount of subsequent brain wear" (p. 24) and "A well–made brain in a well–made body is likely to give the best results under good and wise training. The less good the physiognomy may be, the more the need for good education" (p. 33). Galton, in his 1890 note on Cattell's paper, referred to a "medical man" who was astute at judging "a man's powers" and Warner may have been that doctor.

Mental illness. The third converging trend that directly affected the development of the first tests of intelligence was the concern, most widely felt in France, for the plight of the mentally ill. In the nineteenth century the French led the way in the humane treatment and care of these people. Schemes for classification and methods of treatment were proposed. One major concern was the differentiation between those who had once possessed the power of rational thought and lost it (*dements*) and those who had never developed such capacities (*aments*). At first, tests such as those advocated by Cattell and Galton were used to differentiate these groups (and subgroups within the aments), but without much success. Later, developments in testing and improved descriptions of symptoms of mental illness reached the point where clearer diagnoses became possible in most cases.

An additional force that affected, but did not lead directly to, the testing movement was the development of experimental psychology, particularly as practiced by Wilhelm Wundt and his students. This branch of psychology was concerned with finding general laws of the mind and mental experience. Early testing procedures *were* the procedures of the psychological laboratory. They involved elaborate apparatus and large amounts of testing time on just a few subjects. There were a number of attempts to apply the methods of this psychology (sometimes called "brass instrument psychology" by clinicians) to the study of individual differences in intelligence, but the efforts generally

met with little success because they did not relate in expected ways to commonly held beliefs about intelligence.

There were other lines of inquiry or interest that also affected the development of the measurement of intelligence. One of these, which was related to Galton's concern with distributions of characteristics, was the eugenics movement. Galton was a cousin of Charles Darwin and was very interested in the mechanisms of evolution. According to Darwin's theory, variation among individuals is a prerequisite for evolution. Galton reasoned that if a person's (or any organism's) condition with respect to an inherited characteristic could be measured, it would be possible to direct the course of evolution. Improvement of the human race by controlling evolution came to be called eugenics, and the development of measures that would accurately identify those individuals who would benefit the human race, either intellectually or otherwise, was considered necessary if eugenics was to succeed.

THE WORK OF ALFRED BINET

From the above discussion it is clear that there was active and widespread interest in developing tests to measure human capacities during the latter part of the nineteenth century. One of the most active—and certainly the most productive—workers in the field was the French psychologist, Alfred Binet. His impact, direct and otherwise, has been so great that his name was for many years almost synonymous with tests of intelligence.

Early Work of Alfred Binet

The importance of objectivity. Binet began his career in psychology working with Jean Martin Charcot on problems of suggestibility under hypnosis. Binet's biographer, Theta Wolf (1973), gave a detailed account of this research and of how Binet's devotion to the ideas of Charcot, the laboratory's leader, led Binet, in collaboration with Charles Féré, to take an extreme position based on very weak evidence. In 1885 they published a series of papers on animal magnetism, which focused on the effect that magnetic fields might have on hypnotic phenomena. The papers drew sharp criticism from a Belgian psychophysicist named J. L. R. Delboeuf. After a series of caustic exchanges in which Delboeuf pointed out the lack of even the most elementary controls in Binet's experiments, Binet was totally vanquished. He had, as Delboeuf mercilessly pointed out, been guilty of interpreting his results in light of what he expected to find rather than what the data actually showed. (Gould, in 1981, gave other examples of the tendency to let preconceptions dictate the interpretation of data during the search

for measures of intelligence.) For the rest of his career Binet paid particularly close attention to the need for objectivity on the part of the experimenter and for control of extraneous variables in the testing situation. Never again did he wander away from where the data led, and the science of psychology is almost certainly much richer for his unfortunate early experience.

Individual differences. Shortly after his humiliation, Binet turned his attention to the study of individual differences, an interest that was to absorb him for the rest of his life. Binet was a complete empiricist in his search for ways to assess the character and intelligence of people. He was willing to try anything, and if something worked he used it. He actively pursued studies of graphology, cephalometry, and anything else he could think of in his search for appropriate measures. It was probably his open–mindedness and resistance to direction by theory, learned the hard way from his encounter with Delboeuf, that enabled him to achieve the insight that came in 1905.

Binet's interest in individual differences became manifest in his experimental studies of child psychology, a field of which he may be considered a founder. Beginning with a series of papers in the *Philosophical Review* in 1890, Binet reported studies of his two daughters over a period of years, which anticipated the work and methods of Piaget. However, for the future of mental measurement, the crucial feature of these studies is that Binet developed a great variety of simple experiments that could be done without elaborate apparatus and that tested complex mental functions. These studies probably led Binet to the conviction, expressed in his influential 1895 paper in collaboration with Victor Henri on individual psychology, that the only appropriate way to study the nature of intelligence was to use complex tasks that manifestly required the application of intelligence for their completion. (This paper was, in part, a reaction to some of the work that had been going on at Columbia University under Cattell's direction, in which measures of sensory sensitivity and reaction time were used as tests of mental ability. See Cattell & Farrand, 1896.) Binet and Henri's work set the agenda for much of the psychology of individual differences and for the development of tests of intelligence, although its impact on intelligence testing would not be seen directly for another 10 years. (Recent research on the relationship between intelligence and reaction time, which is summarized in chapter 6, suggests that Cattell may not have been as far off the mark as Binet believed him to be.)

Examining the extremes. In the closing years of the nineteenth century, Binet also turned his attention to the problems of the schools. Although he educated his own children at home, he was concerned about the problems of the mentally retarded in public schools and in asylums. He viewed these individuals as being at the bottom end of a complex continuum of mental

characteristics, and felt that the best way to understand human mental functioning was to study individuals at the extremes. (In later years there would be attempts to decompose Binet's concept of intelligence into a number of unidimensional constructs, but these tended to lose much of what Binet meant by the term *intelligence*.)

The problems that particularly attracted Binet's attention involved differentiating the classes of mental retardation. At the time, three general classes were recognized: *Idiots* were at the bottom of the scale, *imbeciles* came next, and those at the highest level were called *debiles* in French (literally "weak," or those of weak mind). (This last category was given the name *moron* by Henry Goddard, and this became the standard translation for the French term.) There were also several subcategories within each class.

Simon and La Société. Two events occurred just before the turn of the century which had a significant impact on Binet's work. One was the arrival of a young psychiatrist named Theodore Simon at the colony of Perray–Vaucluse, a residential facility for the mentally retarded. Simon presented himself to Binet and asked to work with him. This contact gave Binet access to one source of the subjects he needed for his research, and for the remainder of his life Binet collaborated with Simon on many studies.

The founding of the *Société Libre pour l'Étude Psychologique de l'Enfant* (often referred to as La Société) in 1899 was the second event. La Société was composed of school people—predominantly principals, teachers, and parents, who were interested in the scientific study of education. Shortly after the founding of La Société, some of the members asked Binet, already well known from popular books he had written on child psychology and intelligence, to become a member. In accepting membership, Binet, as he so often did, attacked the issues of the organization with an almost boundless energy and enthusiasm. He was soon elected president and continued to be the guiding force behind La Société until his death. As was the case with his relationship with Simon, membership in La Société provided him with access to subjects and to willing workers and collaborators.

Resolution on mental testing. Binet was a prodigious worker. In addition to his own research, he founded and edited a journal, *L'Année Psychologique*, which provided both a ready medium for his own substantial output (some early issues of the journal are composed almost entirely of Binet's work) and reviews in French of many of the books and articles that appeared in other languages. He also served for several years as the editor of the *Bulletin* for La Société and was very active in the organization's affairs, including the society's Commission for the Retarded. It was this latter group, founded in December 1903 as an expansion of the Commission on Graphology, that took

the initiative in proposing a resolution to the French government that a method be developed to differentiate those who *could* not benefit from normal instruction from those who *would* not. The substance of the resolution read:

> That in the primary schools, the children judged refractory to education, to teaching, or to the discipline of the school should not be sent away without being submitted to a medico–pedagogical examination, and
>
> That these children, if considered educably retarded, should be grouped in special classes annexed to the regular school, or in a special establishment, and
>
> That a special class for the educable be opened for the present in one of the Paris schools, as a demonstration. (Wolf, 1973, p. 165)

This resolution, developed with Binet's strong support at the February 1904 meeting of La Société and delivered by three La Société members to the Ministry for Public Instruction shortly thereafter, was the direct cause for the appointment of the historic Ministerial Commission for the Abnormal by Minister Chaumie, in October 1904. Binet and three other members of La Société served on the commission.

The major events of the next six years are described in detail in most books on testing and many introductory psychology texts. An initial scale of graded intellectual tasks was published in 1905 by Binet and Simon and received a modest reception. It was superseded by the 1908 Binet–Simon scale, in which the concept of mental level was introduced. (Recent critics of mental testing have suggested that a mistranslation of the term *level* as *age* has caused many of the problems with intelligence tests and therefore misinterpretations of the scores. Wolf [1973, p. 202] credits Rene Zazzo and Guy Avanzini with pointing out in the 1960s that the correct translation of Binet's original term is *level*, not *age*. However, since Goddard speaks of mental levels in his 1920 book on the relationship of intelligence to human efficiency, it seems likely that the distinction had been made before and ignored as irrelevant.) The 1908 revision had a major impact, particularly in the United States, and the minor changes that were made by Binet and Simon in 1911 did little to modify the influence of the 1908 scale. To put these three tests in perspective, it is helpful to first examine the environment and influences surrounding their development.

Binet's Work in Context

Influences on Binet. When one reads some of the modern accounts of the beginnings of intelligence testing, one comes away with the impression that

Binet was working alone on the problem and that his scales were the result of consummate genius and sudden insight. While there is no denying the importance of Binet's personal contribution, an adequate appreciation of the history of intelligence measurement must put his work into a context of active and widespread research by many investigators in North America and Europe. Binet, who read widely in the international literature, was undoubtedly influenced by many of these studies. In fact, he commented on many of them in his reviews for *L'Année*.

Rudolph Pintner (1931), who had long been active in the testing movement in the United States, gave the following assessment of Binet's role:

> The work of Binet is important and merits special consideration because of the great stimulus he gave intelligence testing. . . . We are still elaborating upon the ideas that he set forth, and his concept of intelligence is essentially the one that is held at the present time by psychologists. This does not mean that the measurement of intelligence would not have been attempted without the work of Binet. . . . The work in mental tests started by Cattell, and particularly the work of [E. L.] Thorndike in educational measurement would undoubtedly have culminated in the testing of intelligence as we know it today. . . . But . . . without Binet this development would have been much slower and would probably not have taken the decidedly practical turn at the outset that the work of Binet gave it. (p. 21)

Wolf (1973) also noted that "for over two decades some such instrument to differentiate children and adolescents on the basis of their ability to learn had been the objective of researchers in many countries, but everywhere this work seemed to lead to no useful results" (p. 139).

Peterson (1925) provided a good review of the various lines of development that were taking place when Binet was finalizing his ideas. Cattell (1890) had pointed out the need for a normative data base, and an example of such a data base was provided by the anthropologist Franz Boas, who, in 1891, "obtained anthropological measurements of about 1500 school children, and tested them as to vision, hearing and 'memory.' He also secured from their teachers estimates of their 'intellectual acuteness.' . . . This is probably the first attempt to make a comparison of test scores with independent estimates of the subjects by other persons" (Peterson, 1925, pp. 83–84). Binet also compared students' test performances with teachers' judgments of their intelligence as a check on his scales.

Two studies by Gilbert, one published in 1894 and the other in 1897, are also cited by Peterson as having an impact on Binet. The first study used simple sensory, reaction time, and memory tasks and found that as children got older

their performance improved. The second study showed similar developmental results. Both studies used large samples and related test performance with teachers' estimates of intellectual ability. In his review of these studies Binet objected to the types of tests used but expressed interest in the method of age–grading (Peterson, 1925).

Adaptations of Binet's proposals. Binet and Henri's (1895) proposals for the development of a psychology of individual differences stimulated a number of other investigations. One such study was conducted by Stella Sharp (1898–1899). She reviewed the various types of test tasks being used in France by Binet, in Germany by Emil Kraepelin and Axel Oehrn, and in the United States by Cattell and others. Working in E. B. Titchner's laboratory at Cornell, she then administered a large number of tests that were similar in nature to Binet's to a group of college students. The tests included evaluations of memory, mental images, imagination, attention, observation and description, and taste preference tendencies. Since the subjects on whom data were reported were seven advanced psychology students, the conclusions were far from definitive and did not reveal the kinds of individual differences that Binet sought. This is not surprising, considering the nature of the subject pool. (Titchner, as an advocate of introspection, used only highly trained subjects, as was the practice also in Wundt's laboratory.) However, Sharp (1898–1899) did conclude "that individual psychical differences should be sought for in the complex rather than in the elementary processes of mind and that the test method is the most workable one that has yet been proposed for investigating these processes" (p. 390). In addition, Sharp noted the need for careful studies of the reliabilities of mental tests.

Alternative approaches. The approach advocated by Galton and Cattell of using sensory sensitivity and reaction–time tests as measures of intelligence proved to be unsuccessful. This was pointed out clearly by Clark Wissler's (1901) study in which he applied Karl Pearson's new method of product moment correlation to mental test data for the first time. Wissler found that Cattell's tests did not correlate substantially with grades subjects had received in their courses at Columbia. (However, a review by Eysenck in 1986 pointed out that Cattell's and other reaction–time studies conducted during this period used short and unreliable tests and had other shortcomings. Eysenck also describes more recent work that has yielded results more consistent with the expectations of Galton and Cattell.)

Peterson (1925) suggested that the results obtained by Sharp and by Wissler cooled American ardor for psychological testing that used Binet's methods. At Cattell's urging, the American Psychological Association had appointed in 1895 a committee on mental tests to coordinate and foster research on tests. Clearly, there was interest. However, Sharp's and Wissler's studies

yielded what were viewed as largely negative results. This, in the face of the influence of William James's support of introspective methods and his opposition to mental testing, refocused much American attention at the dawn of the new century on achievement testing.

In other countries, there were personalities and events influencing Binet's ideas that resulted in the 1905 scale. Binet was very aware of the work going on in Germany, particularly that by Kraepelin and his students on the description and diagnosis of mental illness, but Binet found Kraepelin's tests to be as inadequate as Cattell's. A second German influence on Binet was Hermann Ebbinghaus's work on problems of mental fatigue in schoolchildren. In the late 1890s, Ebbinghaus was using mutilated sentences, from which one or more words had been deleted, to study learning and memory. (Both Ebbinghaus and Binet considered memory to be a component of intelligence.) The subject's task was to fill in the missing words to complete the sentences. From this work Ebbinghaus concluded that the ability to combine information, which Spearman (1923) would later call the eduction of relationships, was a major component of intelligence. In fact, Ebbinghaus's results conformed closely to Binet's own criteria for a good test of intelligence—that the scores show an increase with age and that students judged brighter by their teachers should earn higher scores.

Publication of the Binet–Simon Scales

The Blin–Damaye precedent. Clearly, the climate was right and there was an evident need for an intelligence test. Binet and his associates in La Société had created the necessary political atmosphere in France with their resolution to Minister Chaumie, and his subsequent appointment of the Ministerial Commission for the Abnormal furthered the trend. Nevertheless, Binet and Simon (1905a) credited another pair of Frenchmen with "a first attempt to apply a scientific method to the diagnosis of mental ability" (p. 28).[1] Dr. Blin of the staff at Perray–Vaucluse and his student Dr. Henri Damaye reported their work in Damaye's doctoral thesis in 1903. According to Wolf (1973), "Binet's review of this monograph [in 1904], while critical, provides enthusiasm for the method" (p. 173). The Blin–Damaye study was first publicly presented at the same meeting at which Henri Beaunis, serving as Binet's representative, read the first preliminary report on the Binet–Simon scale. The

[1] The 1916 translation by Kite of Binet and Simon's papers on the intelligence scale was used as the primary source. To keep the chronology clear, references to Binet and Simon's work give the date of original publication. However, page references for quotations are to the pages of the 1916 translation.

presentations took place at the Fifth International Congress of Psychology in Rome in April 1905.

The method proposed by Blin and Damaye used a set of 20 standard questions presented in an established order. Its objective was to identify those individuals who should be institutionalized as mentally deficient. The answer to each question was scored on a scale of 1 to 5 points. Binet and Simon (1905a) criticized the subjectivity of the scoring and noted the need for inter–rater reliability studies. In addition, they objected that "the whole system constitutes a scale established *a priori*" rather than one based on empirical evidence of item difficulties (p. 35).

Having criticized the Blin–Damaye effort, Binet and Simon (1905a) went on to present the features that they believed an intelligence scale should have. After noting that it was important to protect the child from the stigma of being classified as mentally deficient unless truly justified, the authors argued that the problem of misclassification was due to lack of "a precise basis for differential diagnosis" (p. 14). They argued that the diagnosis of degree of deficiency must be based on a finely and objectively graded series of competencies. "*Quantitative differences . . . are of no value unless they are measured, even if measured but crudely*" (p. 24). Binet and Simon's commitment to an empirically derived ordering of the tasks was plainly stated:

> A distinction of this nature ought to be made only from observations taken from life. The intellectual functions which are the first to develop should be sought out, how they arrange themselves, in what order they appear, how they coordinate. This is the true, the only method. (p. 25)

And this was the method Binet and Simon used.

The 1905 scale. The original scale was formally published in three consecutive papers that together filled 173 pages of *L'Année*. The first paper (Binet & Simon, 1905a) set the background by reviewing previous work, including a detailed description of the 1903 Blin–Damaye scale, and by stating Binet and Simon's philosophy. The second paper (Binet & Simon, 1905b) presented the authors' beliefs about proper diagnostic procedures and gave detailed descriptions of the 30 tasks (or "experiments," as the authors called them) that composed the scale. It was in this paper that Binet and Simon stated that identification of mentally retarded persons should be based on diagnosis by medical, pedagogical, and psychological methods. Their scale was an example of the psychological method, and of it they said, "We believe that we have succeeded in completely disregarding the acquired information of the subject. . . . It is simply his natural level of intelligence that is taken into account" (p. 42). The pedagogical method, which was not seen as a substitute for the psychological

method, involved "an inventory of the total knowledge of the subject" (p. 70) compared with the total knowledge of the normal subject. The proposed medical method used objective screening variables such as height, body temperature (morons were believed to have a lower temperature), and other "stigmata" that were found after blind diagnosis to differentiate the mentally retarded from the normal population.

The third paper (Binet & Simon, 1905c) detailed the methods used to select and norm the tests in the scale, including what types of responses could be expected from children at each level of deficiency. The authors anticipated their later development of the concept of mental level when they noted: "Since we possess a nearly complete series of the results of the tests for each age of normal children, it is easy to find the place of the candidate in such a series. The subsequent consideration of his age permits us then to know if he is backward, and how much above or below average" (p. 170).

The 1905 scale did not create a landslide of attention for its authors. In Belgium, Decroly and Degand (cited in Binet & Simon, 1908) applied the tests to a group of students and found them to be too easy for their subjects. They did recommend the test for general use, noting that it was the best instrument available, but they found several flaws in it. They communicated their objections to Binet and Simon, who used the criticisms in their revision of the scale.

The 1908 scale. The 1908 Binet–Simon scale was presented in a single 90–page paper (Binet & Simon, 1908). While the tests of the 1905 scale had been arranged in approximate order of difficulty, those of the 1908 scale were grouped by the ages at which normal children passed them. The scale was expanded to include 54 tests; only 14 of the original 30 remained unmodified. Four or more tests were provided for each age from 3 to 12 years, with three tests for 13–year–olds.

The 1908 scale was the first to yield a score for mental level. A child's mental level was determined by the highest age at which he or she passed all (or all but one) of the tests (this was the basal level), with an additional year credited for each five tests passed beyond the basal year. The 1905 scale had been more of a clinical interview than a standardized test. Binet and Simon had given cutoff tests for various levels of mental deficiency (for example, Test 6 was the upper limit for adult idiots), but they had emphasized the method by which the tasks were solved as much as the correct solution itself.

Interest in the 1908 scale. The change to include mental level expressed in terms of years, coupled with the greater range and organization of the tests, dramatically increased the attention paid to the scale outside France. In Germany, Meumann and Bobertag both prepared and applied translations. Decroly and Degand repeated their study in Belgium, and in England both Johnston and

Burt tried out versions of the scale. But the most important event that led to Binet's enduring fame was the notice his scale finally attracted in America.

Henry Goddard, who had been appointed research director at the Vineland Training School in New Jersey in 1906, was on a tour of facilities for the retarded in Europe when Decroly brought Binet's 1905 scale to his attention. Goddard translated and published the scale in English as a short pamphlet in December 1908. He later noted (Goddard, 1916) that the 1905 scale had met with so little enthusiasm in America that his search of the literature in 1906 had turned up no mention of it. So when Goddard went to Europe, he was unaware of the scale until he happened to meet Decroly. Goddard's translation of the scale did not result in any immediate fame for Binet either.

When the 1908 issue of *L'Année* arrived in America in 1909, Goddard learned of the revised scale, which he did not believe could work as described. Although originally skeptical, he eventually tried the scale and was so pleased with the results that in 1911 he published a 16–page outline of the tests and procedures, which, by 1916, had seen 22,000 copies distributed (Goddard, 1916). Goddard was the first champion of the Binet scales in the United States, but others were soon to follow.

The 1911 scale. The final revision of the scale was published in the 1911 issue of *L'Année* under Binet's name alone (Binet, 1911) and separately as a joint work with Simon in the *Bulletin* of La Société (Binet & Simon, 1911/1915). The latter work is a much more complete statement of the scale because it does not rely on the 1908 paper. The 1911 revision contained only minor changes. Some tests were relocated, and the number of tests at each age was set at five. Both Peterson (1925) and Wolf (1973) have suggested that by this time Binet's attention had shifted to other topics. Binet was also very ill. He died on October 18, 1911, just six months after the last version of the scale was published.

It has been suggested by some critics of subsequent developments in mental testing that, had Binet survived, the course of mental measurement would have been profoundly different. They imply that his concern for the individual and his clinical orientation would have prevented the emphases on group testing and the intelligence quotient that swept the United States after World War I. Although we will never know for certain, there are reasons to see this as highly unlikely. Binet's method had already been adopted by those who influenced the future of testing in America, and they already had their own agendas. While he might have protested vigorously, it is by no means certain that anyone on this side of the Atlantic would have listened, and Binet

lacked any real base of power to affect the course of events in Europe. His refusal under any circumstance to leave France (he seems seldom to have ventured far from Paris) prevented him from forming close contacts with other major figures in psychology, and without these personal contacts his influence was primarily secondhand. The fact that he did not have a university teaching position left him without a large cadre of students to carry on his work, such as those who augmented the impact of Wundt and others who held influential professorships.

Binet's Concept of Intelligence

The nature of intelligence has been one of the most hotly debated topics in the history of psychology. It is not my purpose in this book to take sides on any of the questions that have been raised regarding this issue. However, an adequate appreciation of the events surrounding the development of ability tests in the United States and elsewhere requires some understanding of what the people involved believed intelligence to be. Binet's ideas on the subject certainly conditioned his choice of tests, and his statements have been used as ammunition in subsequent debates.

Inconsistencies. Binet was far from crystal–clear in his writings about the nature of intelligence. At times he wrote like a faculty psychologist, discussing specific types of abilities: faculties of memory, attention, adaptation, and other dimensions. But when forced to make a definitive statement of his beliefs in the presentation of his scales, he generally came down on the side of a single entity which he referred to as *the* intelligence. In his presentation of the 1905 scale he gave us the oft–quoted passage:

> It seems to us that in intelligence there is a fundamental faculty, the alteration or lack of which, is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances. To judge well, to comprehend well, to reason well, these are the essential activities of intelligence. A person may be a moron or an imbecile if he is lacking in judgment; but with good judgment he can never be either. Indeed the rest of the intellectual faculties seem of little importance in comparison with judgment. (Binet & Simon, 1905b, pp. 42–43)

All or parts of this passage have been quoted by many authors, including Lewis Terman and David Wechsler. Terman's personal copy of the Kite translation of Binet's papers (Binet & Simon, 1916) on the intelligence scale (see p. 17) contains marginal notes indicating interest in this passage.

Binet's conception of *the* intelligence was sometimes so global that he considered all problems of differential psychology to be problems related to intelligence. In some places he used the term *intelligence* almost synonymously with character or personality; in others he focused on relatively narrow definitions such as attention; however, the term was always used to indicate a relatively undifferentiable quality of the person. It was probably his belief in the unified and global character of intelligence that explains why he used the simple sum of performances on a wide variety of tests, coupled with clinical observations from the testing session, to reach a rough quantitative assessment of a subject's intelligence. As Pichot (1968) has observed, "Binet implicitly assumed the existence of a 'general intelligence,' a hypothesis which ran counter to the psychology of 'mental functions' that he had himself used in his other works" (p. 76) and that sometimes seemed to have infiltrated his writings on his scales as well.

Nature or nurture. The scientific atmosphere at the turn of the century was dominated, particularly in England but also in other countries, by the evolutionist ideas of Darwin and the hypothesis of genetic transmission of traits that his theory implied. Binet spelled out his position on this aspect of intelligence in a book he wrote for popular consumption in 1909, *Les idées modernes sur les enfants*. He made it clear when he proposed a set of "mental orthopedics" to help strengthen weak minds that he viewed observed intelligence as modifiable. This program included a series of exercises designed primarily to help children diagnosed as "defective" to strengthen their powers of attention to detail and to the world around them. For example, one exercise required the child to carry a full bowl of water without spilling any (see Wolf, 1973, p. 207). In commenting in *Les idées modernes* on the claims of others that intelligence was immutable, Binet wrote: "We must protest and react against this brutal pessimism" (cited in Kamin, 1974, p. 5).

Statements such as the one above have been offered as proof that Binet believed intelligence to be environmentally determined. While it is clear from Binet's writings that he believed the environment had an impact on one's performance on his scale (he used social class as an explanation for the difference in results obtained by Decroly and Degand from those he found), it is also apparent that he saw inheritance as a factor. The following quote from *Les idées modernes*, which appears on the page following the passage quoted by Kamin, makes his position quite clear:

> Anyone's intelligence is susceptible to development; with practice and training, and especially with appropriate methods [of teaching] we can augment a child's attention, his memory, his judgment—helping him literally to become more intelligent than he was

> before . . . right up to the moment when he arrives at his limit. Thereafter progress is ruled by a remarkable law of fixity; the ordinarily great progress at the beginning diminishes little by little . . . and despite great efforts, the moment arrives when it becomes practically equal to zero. At this point the person has attained his limit, for incontestably there is a limit. It varies according to the persons and the functions under consideration. (cited in Wolf, 1973, p. 207)

Binet had offered a similar view of this position in his 1908 paper presenting the revised scale. Although translators disagree on the precise wording, Peterson (1925) gave a reasonably accurate rendering of the passage:

> The person of great innate qualities shows his superiority in the repetition of numbers, the repetition of sentences, the drawing of a design cut in quarto–folded paper, the arrangement of weights, the interpretation of pictures, etc.; and it is especially the province of these tests, when this need is evident, to isolate from the scholastic effects the real native intelligence. (p. 205)

Kite translated the last phrase as "to free a beautiful native intelligence from the trammels of the school" (Binet & Simon, 1916, p. 259), but Terman (who was fluent in both French and German) revealed by his comments in the margin of Kite's text that he considered Kite's rendering inaccurate. At any rate, it is clear that Binet's contemporaries represented him as believing that intelligence was not completely malleable and dependent on experience, and recent translations agree with this depiction. (Binet's proposal [Binet & Simon, 1905a] to separate the measurement of psychological [native] from pedagogical [learned] aspects of intelligence [see p. 12] offers further evidence that he believed some aspects of intelligence were inherited.)

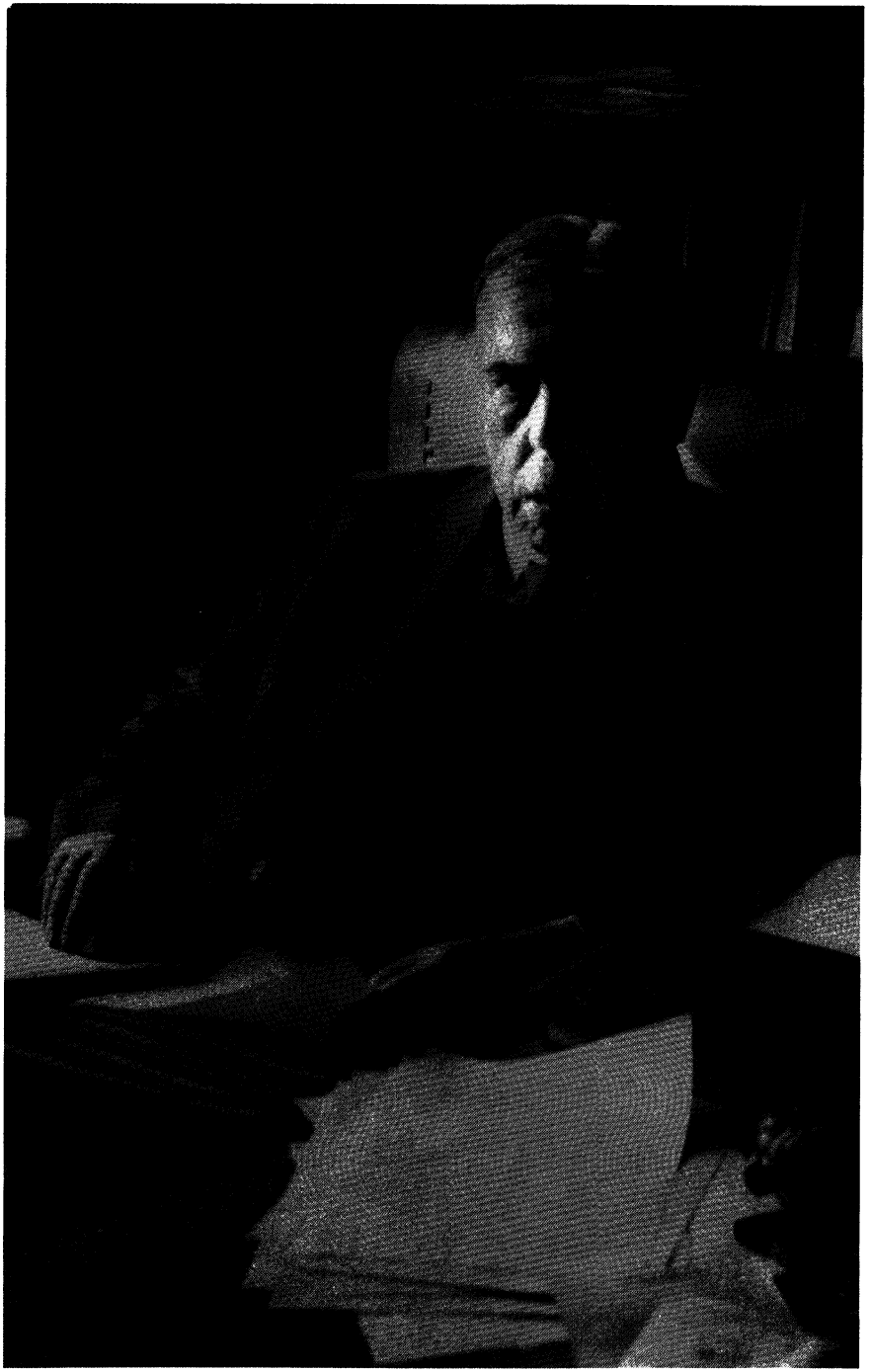
Binet's description of intelligence seems quite consistent with the concept of reaction range proposed by Dobzhansky (1962). This position holds that an organism's observed level on any trait will fall in a range limited by genetics, but that the precise manifestation of the trait within that range is a function of environment and experience. Binet's position appears to be well suited to a scientist totally committed to following the data where they lead and to entering each investigation determined not to let preconceptions color perceptions.

Assumptions Behind Intelligence Measurement

Binet was able to create his scale because he made certain assumptions about the nature of intelligence, assumptions that we have already encoun-

tered in the research that he and others had completed prior to 1905. The first assumption was that the functions involved in intelligence showed an increase with age. That is, whatever intelligence is, it is something that shows a normal and fairly consistent course of average development: Older children have more of it than younger children. The second assumption was that intelligence is needed for success in school. Thus, those judged to be bright by their teachers possess more intelligence than those whom teachers judged to be dull.

These two factors formed the empirical basis upon which Binet built the 1905 scale and its successors. In order to be included in the scale, a test had to show a pattern of decreasing difficulty for older children and a higher success rate for those children labeled bright by their teachers. Each test that passed these criteria was placed in the age group or level corresponding to the chronological age when about 60 to 70 percent of an unselected group of children could pass it. When a new child was tested, his or her mental level was said to be equivalent to that of the highest age group wherein he or she could pass all (or all but one) of the tests for that group. These two criteria continue to form the basis for the development and selection of test items for the measurement of intelligence today.



Edward L. Thorndike (1874–1949)

2

Early Efforts in the United States and England

RESEARCH IN THE UNITED STATES

Binet and Simon were not working in a vacuum at the turn of the century. American psychologists such as Goddard and Terman were seeking usable tests for clinical diagnosis of mental retardation and other psychologists were pursuing tests for use in the schools. I have already noted the work of Boas and Gilbert in developing age norms for some anthropometric and simple psychophysical measures, and of Sharp and Wissler in correlating test scores with indices of scholastic performance. I will now examine some of the other studies that were being conducted prior to World War I.

The educational objectives of the testing movement were identified by Kirkpatrick in a speech to the American Psychological Association in 1899. Noting that much prior research, such as that of Cattell and Farrand (1896) and Sharp (1898–1899), had used college students as subjects, Kirkpatrick (1900) called for the development of ability tests for schoolchildren as well:

> I wish to emphasize to this Association the importance of testing persons of different ages and seeking for the normal standard at each age before we can intelligently interpret individual records. The psychological problem of tests of general mental ability cannot be solved till the psychogenetic [i.e., developmental] problem of the stages of improve-

> ment in various lines of mental development has been solved. . . . It is desirable to have tests of such a nature that they can be taken by children as well as adults, that they shall be such that all persons tested will have had about equal opportunity for the exercise of the power tested, and that in the interest of economy of time the tests so far as possible shall be so planned that they can be given to a whole class or school at once, instead of to each individual separately. (pp. 279–280)

This proposal sounds very much like a formula for many of today's testing programs.

Other authors echoed Kirkpatrick's call. Kelly (1903) studied the distribution of psychophysical test performances in the hope that "some ready and simple method might be determined of differentiating the normal from the abnormal child" (p. 346). He stated that research should develop "norms in terms of which a child can readily be scientifically classed for pedagogical purposes" (p. 371). Whipple (1904) argued that psychophysical tests such as those Kelly had used were not useful as measures of intelligence, even in children. What correlation was observed, Whipple suggested, was due to individual differences in the ability to understand the directions of the testing situation, an ability that is clearly a complex mental process such as that called for by Binet and Henri.

E. L. Thorndike and Educational Measurement

Criticisms of educational practice. Perhaps the most strident critic of educational practice at the time was Edward L. Thorndike, who was already well known for his studies of animal learning. Thorndike entered the field of mental testing in 1902 (Aikens, Thorndike, & Hubbell, 1902) with a paper that criticized current testing practices and proposed a theory of intelligence that explicitly rejected any general intellectual factor, postulating instead a very large number of neural bonds. By the following year his thinking was sufficiently well defined that his book *Educational Psychology* (1903) read like a manifesto for the progress of education and mental testing.

In 1903 the prevalent theory of education held that there was transfer of ability from one scholastic area to another and that proper instruction involved mental discipline.

> Schemes for individual instruction and for different rates of promotion are undertaken largely because of certain beliefs concerning the prevalence and amount of differences in mental capacity; the conduct of at least two classes out of every three is determined in great measure by the teachers' faith that mental abilities are so little spe-

> cialized that improvement in any one of them will help all the rest; manual training is often introduced into schools on the strength of somebody's confidence that skill in movement is intimately connected with efficiency in thinking. (Thorndike, 1903, pp. 1–2)

Thorndike asserted that learning was specific to the context and he rejected the commonly held belief that instruction in one subject, for example, Latin, would lead to improvement in another, arithmetic: "Since those who succeed in the study of Latin are better in general discrimination and judgment than those who fail, we conclude [erroneously] that learning Latin vastly improves general discrimination and judgment" (p. 93). It is the general power of the brain to form neural bonds that differentiates the better learner from the poorer, and this power increases with age. "It suits the vanity of educational theory to fancy that the changes [that result from study] are wholly due to discipline. But it is almost certain that maturity alone would cause a fair gain in efficiency" (p. 93).

The nature of intelligence. Thorndike believed in a biological explanation of intelligence—the neural bonds postulated in his connectionist theory of learning. He also believed that the structure and efficiency of the nervous system were largely determined by inheritance.

> There is no reason to suppose that the brain is less influenced by [immediate ancestry (Thorndike's term for genetic influence)] than the tissues that cause height or the shape of the skull bones that causes the cephalic index, or the deposits of pigment that cause eye color. Immediate ancestry is thus a probable cause for original mental nature. (1903, p. 51)

> What ancestry does is to reduce the variability of the offspring and determine the point about which they will vary.... Immediate ancestry will then, when influential, cause children to deviate from the general average toward the condition of their parents and to vary less among themselves than would the same number of unrelated individuals. (pp. 48–49)

Although the Lamarckian theory of inheritance of acquired characteristics was still very much touted at the time as an explanation for evolution, Thorndike rejected it totally: "We must deny the mental acquisitions of one generation any considerable share in the original natures of the next. Original nature springs from original nature. Its improvement depends on the elimination of the worse, not their reformation" (p. 65).

The role of education. In this book, Thorndike also spelled out the mission of education in the United States:

> 1. To supply the needs of the brain's healthy growth and to remove physiological impediments to it.
>
> 2. To provide stimuli to desirable mental variations and to withhold stimuli from the undesirable.
>
> 3. To make the outcome of desirable activities pleasurable and to inhibit their opposites by discomfort.
>
> The three chief practical problems of education would thus be those of hygiene, of opportunity and of incentives and deterrents. (1903, p. 79)

Furthermore, people differed in mental ability, and the schools had a duty to identify and to capitalize upon those differences for the welfare of society:

> All environmental agencies and especially our educational agencies are a great system of means not only of making men good and intelligent and efficient, but also of picking out and labeling those who for any reason are good and intelligent and efficient. . . . They help society in general tremendously by providing it not with better men, but with the knowledge of which men are good. . . . The schools [of the United States] always have and always will work to create a caste, to emphasize inequalities. Our care should be that they emphasize inequalities, not of adventitious circumstances, but of intellect, energy, idealism and achievement. (pp. 94–96)

Thorndike noted the "growing demand for institutions and separate classes for the feebleminded" (p. 122) and felt that mental measurement or the assessment of intelligence was essential for the proper care and treatment of these individuals. He believed that ability was distributed in the population in a way that followed the normal distribution, and he considered the very bright and the very dull to be extremes in this distribution, using the very modern–sounding term "exceptional" to refer to both. In relating these beliefs to the practice of education and the character that a measure of intelligence should have, he observed:

> English writers agree in using the terms idiots, imbeciles, and feebleminded to refer in order to the three lowest conditions of intellect. . . . but it is certain from our knowledge of the distribution of mental traits that any effort to sharply separate idiocy from imbecility, or the latter from feebleness of mind must fail. The words are but names used roughly for sections of a continuous surface of frequency. The obvious thing to do is to arrange a scale of intellect and describe that of each individual by his precise station on that scale, not by a vague name. (p. 128)

Thorndike also suggested, very much in line with Binet's position, that feeble–mindedness as it might be measured by such a scale was not necessarily a permanent condition. He explained that, given the necessary scale:

> It seems desirable further to separate children who are feeble–minded and are destined to remain so from those who are simply backward in mental growth and may eventually reach a fair station. We know that in physical growth some children who from six to twelve or thereabouts are far below average in stature for their age, in later years make up part or all of the deficiency, and there are many reasons for believing the same to be the case with mental growth. The essentially dull should never be confused in theory or in actual treatment with those who are temporarily deficient. (Thorndike, 1903, p. 129)

The problem, for which Thorndike offered no solution, was how to tell the difference.

Norsworthy's study. Thorndike's students and colleagues at Teachers College, Columbia University, were actively searching for such a measure of mental ability. Cattell's (Cattell & Farrand, 1896; Wissler, 1901) program on the main campus of Columbia had not yielded much that would be useful, but Thorndike and his students continued the quest. By 1903 Naomi Norsworthy, working under Thorndike's direction, had already collected the data for her doctoral study (Norsworthy, 1906) relating mental ability to the scholastic performance of schoolchildren. A group of 30 girls selected by their teachers as being unable to profit from regular school work was compared to a group of agemates from the same New York City school. The two groups of girls had been tested using physiological measures (height, weight, and so forth), psychophysical and perceptual measures (mazes and memory tests of varying complexity), and tests of abstraction such as writing opposites and giving examples of named classes of objects. Few differences were found in the simpler tasks, but in summarizing the results of the complex tasks, Thorndike (1903) related a finding that could have been written by Binet:

> The one chief and essential characteristic of these children is thus their inability to think in symbols or with relationships or in such a way as to let a number of processes combine to decide what a given thought or reaction should be. Concrete facts they can think and respond to one by one, but they can not think in symbols that stand for groups of facts or elements in facts, nor can they think facts together in causal or other series or respond correctly to related groups . . . and they give us some reason to believe that as they grow older they will develop continually

> less and less rapidly than ordinary children and so fall farther and farther behind. (pp. 134–135)

Norsworthy (1906) concluded that the adequate description of a person's abilities required scores for several traits. She recommended that these be presented as a profile that would identify the person's strengths and weaknesses.

Thorndike continued his attack on the problems of measurement when, in 1904, he published the first book devoted to measurement theory, *An Introduction to the Theory of Mental and Social Measurements.* Here he took note of some of the technical problems confronting those who would measure mental phenomena, including problems of equality of units, of locating a true zero, and of the instability of the phenomena themselves:

> With human affairs not only do our measurements give varying results; the thing itself is not the same from time to time, and the individual things of a common group are not identical with each other. . . . Even a very simple mental trait, say arithmetical ability or superstition or respect for law, is, compared with physical things, exceedingly complex. (p. 6)

In a manner reminiscent of Binet, he also warned that control of the conditions under which human ability was measured was critical: "Every extrinsic condition influencing that ability should be alike for all. Otherwise we are led into errors, which may be called errors of inferring an ability *in abstracto* from its manifestation under particular conditions" (p. 160).

I have covered in some detail Thorndike's pre–1905 position on the nature of intelligence and the way to measure it for two reasons. First, it gives one a good idea of then–current American thought on the topic. The measurement of intelligence or human mental capacity was a major objective of research in American psychology before the publication of Binet's scale. Thorndike, like most members of the English–speaking psychological community, believed that genetic factors played a large part in determining human characteristics. Therefore, a measure of intelligence, as distinct from scholastic achievement, must measure a largely inherited quality. Second, the program of testing and measurement at Teachers College was already large and influential and would become more so. For the next 50 years, Thorndike, who was rapidly becoming a central figure at the College, would profoundly influence American thinking about the conduct of science with respect to education. While few of the key figures in test development were direct products of Columbia, there is little doubt that Thorndike's ideas influenced many aspects of the new mental testing movement in America.

Lewis Terman's Early Work

Another major figure in American mental testing also made a significant contribution to the field before the publication of Binet's first scale. Lewis Terman, who acknowledged that his lifelong interest in intellectually deviant individuals began when he was a child, had become familiar with Binet's work on measures of intelligence, as well as that of other Continental psychologists, while he was an undergraduate at Indiana University. In the course of graduate study at Clark University he undertook an investigation of mental abilities which was directly influenced by Binet's work. A classmate, E. B. Huey, had visited Binet during 1903 and 1904 and had brought back some of Binet's most recent ideas. For his doctoral dissertation Terman undertook to apply Binet's tests, along with several of his own, to seven "bright" and seven "stupid" boys (Terman's terms).

The study, published in the 1906 volume of *Pedagogical Seminary*, led him to the following conclusions:

1. The "bright" boys were superior in intellectual but not motor tasks.

2. "Intelligence in these subjects does not tend to develop along special lines" (p. 372). Intelligence is a general characteristic of the individual.

3. There were no differences in persistence.

4. "Bright" boys preferred reading; "stupid" boys preferred games.

5. The results should be interpreted as favoring an endowment hypothesis rather than one related to training.

(It is interesting, perhaps prophetic, that the next item in the journal was an annotated bibliography of child study for the year 1905 which included references to six works by Binet.) Terman sent a copy of his study to Binet. Although there is no reference to Terman's work in Binet's paper presenting the 1908 scale, it is highly likely that Binet read the paper because some of the new tests he used were very similar to ones that Terman had introduced in his study.

RESEARCH IN ENGLAND

Prior to the publication of the 1905 scale, English psychologists were also searching for ways to measure intelligence. I have already mentioned that Johnston and Burt each applied Binet's tests shortly after their publication

to English children. However, another Englishman was also working on the problem of the measurement of intelligence and, quite independently, taking a course of action similar in some ways to Binet's. Initially introduced to psychology in Wundt's lab at Leipzig, Charles Spearman seems to have been an argumentative sort who was in constant conflict with other theorists almost from the start. His initial publications in the areas of intelligence (1904a) and psychometrics (1904b) set the stage for a debate that raged for 35 years.

In the first paper Spearman (1904a) proposed a theory of intelligence known as the two–factor theory which held that the level of any person's intellectual functioning was determined by a general factor of intelligence that pervaded all aspects of the person's behavior. He observed that all tests of cognitive functioning correlated positively with each other and he argued that general intelligence, his famous *g*, was the reason. Any particular intelligence test, according to the theory, measured two "factors": the general factor, *g*, and a specific factor unique to that test, *s*. The correct way to measure an individual's level of *g* was to take the mean performance on a "hotchpotch" of different tests. In this way the common factor would be measured in all of them and the effects of the specific factors would tend to cancel each other out.

Spearman's second 1904 paper (it actually appears earlier in the journal) was purely statistical, but served an important function in support of the two–factor theory. In this paper Spearman pointed out that mental test scores include a component of random error of measurement, and he proposed a formula by which it was possible, given certain fairly reasonable assumptions, to estimate the correlation that would exist between two tests if they were measured without error. This was the correction for attenuation (a correlation would be smaller or attenuated if the scores on which it was computed contained errors of measurement), and it related to the two–factor theory because the theory applied precisely only to variables that were measured without error.

Spearman spent much of the remainder of his career elaborating on and defending his theory. He considered Binet and Simon's 1905 scale a variant of the method he had proposed, and he vigorously claimed priority. (Binet also expressed frequent complaints about the failure of others to recognize his priority in one or another area, but not in the matter of the intelligence scale.) Of Binet, Spearman wrote in 1930: "Typical among its [faculty psychology's] champions, Binet had long been busily measuring such faculties as 'imagination,' 'memory,' 'attention,' and the like. Altogether remote from him lay any such idea as that of measurable 'general' intelligence" (p. 324).

Spearman implied that Binet and Simon had appropriated his idea about how intelligence should be measured. They "incorporated this hotchpotch procedure in their celebrated scale of tests published in 1905; this was composed of a great many promiscuous tests and was said to discover the subject's 'level,' which is only another name for his mean result at the different tests" or his level of *g* (p. 325). For his part, Binet, in reviewing Spearman's 1904 papers for *L'Année*, complained that the tests Spearman used were inappropriate because they did not tap higher–level processes. There is no evidence to suggest that Spearman's papers influenced Binet's thinking about the most effective way to construct a scale, but the scale Binet eventually published was completely consistent with Spearman's proposal.

THE TRANSATLANTIC CONNECTION

Out of all this ferment it was ultimately Terman who emerged as the leader in developing a practical measure of intelligence. Looking back on the events of 1900 to 1905, Terman (1932) would later comment that Spearman's "logic appeared to be waterproof, but the conclusion to which it led [*g*] . . . seemed to me as absurd then as it does now" (p. 319). Terman also could not accept Thorndike's proposal of a nearly infinite number of specific neural bonds. In addition, he seemed quite put out by the tone in Thorndike's writing: "For a youth still in his twenties, he seemed to me shockingly lacking in a 'decent respect for the opinions of mankind!' " (p. 319). Yet Terman conceded that "it would have been an untold boon to me if I could have spent a year [studying statistics] with Thorndike immediately upon leaving Clark" (p. 320). (It is interesting that Thorndike considered mathematics and statistics to be his greatest weaknesses.)

The Binet–Simon Scale in America

We have already seen that Goddard brought the first version of the Binet–Simon scale to America in 1908 for use in identifying the mentally retarded and that he subsequently published a description of the 1908 scale. To say that the 1908 scale had a major impact on American clinical and educational psychology would be one of the understatements of the century. To say that the scale took the United States by storm would be more accurate. It would seem that almost everybody who had an interest in clinical psychology and could read French published an English translation or adaptation of the scale. Huey (1910), Kuhlmann (1911), and Wallin (1911) each published a version, as did Clara

Town (Binet & Simon, 1911/1915). Others reported various uses of the scales. Terman, after several years of inactivity due to tuberculosis, reentered the arena with a large study using items from the 1908 scale (Terman & Childs, 1912). The extent of the barrage is revealed in an annotated bibliography of references to the Binet scales compiled by Kohs (1914), in which he cited 254 studies and commentaries on one or another version of the Binet scales, including 20 written by Goddard. Three years later Kohs (1917) reported an additional 457 references! Reviewing this period, Peterson wrote in 1925:

> The [testing] movement had got so well under way, and several investigators had begun so early to collect data for their own modifications of this scale . . . before Binet's final revision appeared in available form, that the [1911 version] did not readily, or indeed ever, replace the 1908 scale. . . . Those investigators who have worked out revisions of their own since 1911 have drawn freely on all three of Binet's scales and have added new tests as necessity demanded. (pp. 240–241)

Many of the American adaptations included changes similar to the ones Binet offered in 1908 and 1911, such as having an equal number of tests at each age, correcting the age placement of some of the tests, and extending the range of the scale up or down. Huey (1910), for example, included tests from the 1905 scale in his version in order to extend the scale down to the first and second years of life. (The 1908 scale began with Test 7 from the 1905 scale, which was assigned to the three–year level.) Several authors recommended using fractions of a year to better express mental level.

A major area of debate concerned who should use the scale. In his *Les idées modernes sur les enfants*, Binet (1909) had listed the test titles of the 1908 scale and recommended that school teachers and parents use the scale to determine the level of intelligence of their pupils or offspring, a suggestion that surprised his biographer because of Binet's concern for clinical interpretation in the testing setting (Wolf, 1973, p. 317). However, by 1911 Town noted a much different tone in his writing. She quoted Binet as writing that "the idea that a method of examination can be made precise enough to be trusted to everyone must be abandoned" (p. 240). She then went on to warn that "unfortunately the American public has not read these paragraphs, and the result which is threatening is a wholesale use of the Scale in an unscientific manner" (p. 240). Town's warning fell on deaf ears; American psychologists were more interested in using the scales to solve problems than in regulating their use.

Other psychologists were hard at work on an American standardization of the Binet–Simon scales. While Binet and Simon had used a relatively small number of cases at each age, American investigators were gathering samples

that were huge by comparison. Wallin's 1911 scale, which, like Huey's, contained items from the 1905 as well as the 1908 scales, was administered to several hundred cases, and Kuhlmann tested his version on more than 1,300 subjects. Because Goddard, Huey, and Kuhlmann all worked as clinical psychologists in institutions for the mentally retarded, they were particularly interested in the low end of the distribution and had access to large numbers of retarded subjects. To determine the expected range of human performance, Goddard (1911) tested 2,000 "normal children" for his standardization.

Terman, who had gone to California for health reasons and had been helped and encouraged by Huey to obtain a position at Stanford, used his base of operations there during 1910 to test 400 children with his adaptation of the 1908 scale. His report of that study (Terman, 1911) provided a list of the tests he was using. (He had added interpretation of fables, sentence completion, vocabulary, and others to the basic Binet set.) Terman also included an example of the record blank he was using and offered 50 free record booklets to any qualified person (what "qualified" meant was not specified) who wanted to try out the tests on the one condition that he receive a copy of the results. On the basis of his 400 cases, he concluded, consistent with other authors, that (a) the tests at the bottom of the Binet–Simon scale were too easy for their age level, (b) the tests at the top were too hard, and (c) it was necessary to standardize directions and administrative procedures if test results were to be comparable. Terman recommended his own set of tests and procedures as the best to use.

Binet was aware of the attention that his work had attracted in the United States. Terman corresponded with him directly, as did others. However, there is no mention in his late works of this correspondence or of the use of his scales in America. While it was not uncommon for authors of this period to omit references to studies by others (or to refer to a study by simply giving the author's name and the date of publication, assuming that everyone had read it and knew where to find it), Binet seems to have been particularly prone to the habit. One possible reason may be that Binet was distressed by Americans revising his scale without asking his permission. Cyril Burt, who corresponded with Binet and did ask permission in 1908 from both Binet and Simon to use the tests in Britain, stated in a letter to Binet's biographer, Wolf, that this was the case. Burt's revision of the scale was published in 1921. (For some reason, Burt [1952] failed to make *any* mention of Binet in his autobiographical sketch, although he did mention his own work with Binet–type scales that he published in 1921.)

The status of mental testing was summarized in manuals of mental tests by Guy Montrose Whipple in 1910, and again in 1914–15. At both times he noted the need for standardized procedures and gave comprehensive descrip-

tions of the tests then in use. He covered everything from the correct way to measure height (either standing or sitting) and various forms of reaction time, to Ebbinghaus's tests of combinatory ability, and to memory and analogies tests, the latter having been first proposed by Burt in 1911. The Binet scales were included in the 1910 volume, but were omitted in 1915 because they were widely available in other sources. Whipple also sounded a warning that, like so many others, went unheeded:

> Now that interest is directed so much toward the question of "types," it seems particularly necessary to caution against the fallacy of taking the result of a single test as a positive indication that *S* falls into this or that type—because, of course, belonging to a type really implies possession of a persistent tendency.... When the function is of a "higher" and more complex order, *more than one test* must be used.... We can [not] regard any mental function as so clean cut, distinct and open to isolation that any single mental test can fully and finally map its dimensions.... To try to concoct a single and final test of such a comprehensive capacity as general intelligence becomes doubly absurd. (1914, p. 12)

Whipple described the Binet–Simon scales in the 1910 volume, but omitted any full discussion of them in his next edition because two fairly complete versions of Binet's works were about to be published. Goddard, long a champion of the Binet method, had commissioned Elizabeth Kite, a member of his staff at Vineland, to prepare a translation of the papers from *L'Année* which related directly to the three versions of the scale (Binet & Simon, 1916). Goddard's efforts were in part a reaction to the proliferation of versions of the scales, and in part an attempt to make Binet's original work available to a wider audience of American scholars and educators. At about the same time Town brought out the third edition of her version of Binet's 1911 scale (Binet & Simon, 1911/1915) . This little book was a translation of Binet and Simon's last complete statement of their 1911 version of the scale. Thus, American readers had two fairly complete versions of Binet's work available to them in English, and Whipple covered the Binet scales by referring to these sources.

TECHNICAL CONSIDERATIONS

The Age Scale vs. the Point Scale

Not everyone was completely satisfied with the form of the scale that Binet had produced. In Italy, Treves and Saffiotti suggested as early as 1912 that the age–scale concept that Binet used was not the best vehicle to express intellectual status. They argued that the tests should be arranged by type and

difficulty and adopted a "faculty psychology" position on the nature of intelligence (cited in Young, 1924, p. 22). Similar arguments were voiced by Huey (1912). By 1914 Robert Yerkes and his coworkers had provided just such an alternative in what they called a point scale (Yerkes & Anderson, 1915; Yerkes, Bridges, & Hardwick, 1915).

Yerkes and his associates described three basic differences between an age scale such as the 1908 Binet–Simon scale and their Point Scale. First, the item arrangement was different. Binet had used a variety of item types at each age. No two tests were the same at any single age, although a particular type of test might be repeated at several age levels. Age–scale advocates claimed that this variety helped maintain the subject's motivation and interest. But critics argued that it introduced a chance element into an individual's score because the same intellectual functions were not required at every age and some people were better at, or more familiar with some tasks than others:

> The age arrangement of the Binet Scale assumes that the mental development of all normal individuals proceeds by similar stages, that the correlation between different functions is the same for all individuals at a given stage, and that each stage of mental development corresponds, in turn, to a certain physical age. . . . These assumptions are not yet justified. On the contrary, the evidence thus far is against them. (Yerkes et al., 1915, pp. 31–32)

The Point Scale consisted of twenty subscales, each containing items of a particular type. Within each subscale, the items were ordered by difficulty.

The second major difference was in the amount of credit given for a correct answer. In the age scale a single credit was given for reaching criterion performance on any test. For example, in the 1908 Binet–Simon scale, there was a set of five comprehension questions that formed part of the tests for 10–year–olds. In order to receive credit for passing this test, the subject had to give three or more correct responses. No distinction was made in scoring the test between three correct and five correct answers, but the difference between two and three correct was the difference between passing and failing. In a point scale, the score for a response may differ, depending on its quality or, in cases such as the one above, its quantity. Yerkes and Anderson (1915) argued that this approach yielded more information for each unit of testing time.

The third difference between the two methods of scale construction was the manner in which norms were developed. While Binet was very much the clinician and concerned with each individual case, his way of measuring intelligence was explicitly normative. A child's mental level was determined by comparing test performance with his or her age–mates. The means by which

this normative comparison was accomplished was by the placement of the tests. Each test was assigned to a particular age level, the level at which the average child could pass it. The definition of this point was a matter of some debate. Binet was never willing to state a percent of success that was required for the placement of a test. Others put the figure at anywhere from 50 percent (Otis, 1916) to 87 percent correct. Truman Kelley (1916) provided an analysis of the various percentage options that were then being considered.

Whatever percentage was selected, the placement of tests at various age levels was a difficult and time–consuming task that had to be carried out by trial and error. The objective was to so place the items that the average chronological age of children assigned to a particular level was equal to the mental age assigned to that level. Yerkes et al. (1915) argued that this trial–and–error method of test placement was inefficient and that such a standardization was too rigid. It was much simpler, they claimed, to determine the average number of points earned by children of a given age and to use that value as the normative standard for the age. A child's mental level was then determined by finding the group whose mean score was most like his or hers.

In addition to simplifying the placement of tests, Yerkes argued that the point scale approach made it possible to develop specialized norms for specific purposes or populations. In particular, he asserted that adult performance should not be expressed in terms of children's norms. One could also develop local norms for particular populations, something that was not possible with an age scale. He argued that if a point–scale format were used, it would then be simple to collect data from a new normative sample.

Yerkes' approach was adopted in its general outline by David Wechsler in the scales that he developed in the 1930s and later. The most recent revision of the Stanford–Binet is also much closer to Yerkes' format than to Binet's or Terman's original age–scale organization.

Normative Expressions—IQ and CI

In his 1908 and 1911 scales, Binet used an age–scale format to determine a child's level of mental functioning. Intended to indicate what the child could do, the mental level or mental age did not answer the critical question of whether the child was retarded or advanced. Binet recommended that this conclusion be reached by comparing the subject's performance with that of other children of the same age. For children younger than nine, a mental level two years below the chronological age indicated retardation; above the ninth year, three years. Binet felt that such a rough classification was about what the data

warranted, and that the diagnosis of retardation should only be made in light of his three–pronged examination—psychological, pedagogical, and medical.

IQ defined. Others believed that the normative meaning of mental levels could be more adequately and accurately specified. Stern, in his 1912 paper, proposed an alternative by which the relationship between mental age and chronological age could be expressed as a single number. By forming a ratio of mental age to chronological age one obtained an "Intelligenzquotient,"

$$\frac{\text{Mental Age (MA)}}{\text{Chronological Age (CA)}} = \text{Intelligence Quotient (IQ)}$$

which would be 1.00 for all children who tested "at age," greater than 1.00 for those who were advanced, and less than 1.00 for those who were retarded. This would be true regardless of the child's age.

This last fact may be responsible for a notion that has caused untold havoc in mental testing ever since because it can be misinterpreted to mean that an individual's intelligence is constant. That is, whereas the mental level (or mental age, the term that gained currency at about this time) of any individual would continue to grow throughout the period of development, the average Intelligence Quotient (or IQ, as it soon came to be called) for people of any age is, by definition of the way the scale was constructed, 1.00. (It wasn't until the 1920s that the index was multipled by 100 to give the range of values familiar today.) Because, on the average, those who were below the mean in test performance at any given age tended also to be below the mean at all other ages and those who were above the mean tended to be above at other ages, IQ values tended to show stability over time. This tendency of individuals to maintain the same relative position in a group came to be exaggerated and distorted by some test users, and certainly by critics of testing, as a claim that the intelligence of any individual was constant and could not be altered by environmental changes. Injudicious statements by proponents of testing had the unfortunate consequence of fostering this impression, and the result was a debate that has gone on with greater or lesser intensity ever since.

Alternatives to IQ. The IQ was not the only measure suggested to express the relative standings of individuals. The year before Stern's paper, Weiss (1911) had pointed out that because psychological and educational scales have arbitrary origins and units, it would be desirable to develop a common method of scaling so that measures would be comparable and could be treated statistically. He proposed that the mean of each distribution be set at 50 and that the unit for any given scale be the observed mean divided by

50. He claimed that if each instrument were then standardized on an appropriate reference group, comparable scales would result.

Robert Woodworth, a neighbor, friend, and Columbia colleague of E. L. Thorndike, can probably be given credit for the most direct contribution to what is now standard practice for reporting relative performance. In 1912 he proposed that test scores be expressed as *standardized* deviations from the mean of the distribution as a first step in finding the average of several tests that were not comparably scaled. Each individual's deviation from the group mean could be divided by some measure of the variability of the distribution. The standard deviation, the average deviation, and the interquartile range were examples he used. He also showed that such a scoring system simplified computation of the correlation coefficient.

Truman Kelley (1914) reviewed the proposals of both Weiss and Woodworth and pointed out that Weiss's method would not work. Kelley advocated what he called "standard measures," which were Woodworth's standardized deviations using the standard deviation. These, of course, are what we call *z*–scores today. It remained for Otis (1917) to show that standard scores could be transformed to yield a distribution with any desired mean and standard deviation and to argue that IQs should use this approach with a mean of 100 and a standard deviation of about 15. All major intelligence tests in use today express test results as some minor variant of these transformed *z*–scores.

The Coefficient of Intelligence. Yerkes and Wood (1916) proposed an alternative to Stern's IQ which they believed was superior for use with the Point Scale. After reviewing five possibilities (mental age, the age difference proposed by Binet, percentile ranks, deviations from the mean or median, and the IQ) and finding problems with each, they offered the Coefficient of Intelligence or CI. This they defined as the individual's observed score divided by the mean of the age group to which the individual belonged. Like the IQ, it would yield a value of 1.00 for someone who scored at the group mean, values below 1.00 for those below the mean, and values above 1.00 for those above the mean. Because Yerkes advocated using different norms for different social status groups and for different ages (Yerkes & Anderson, 1915), the mean of the norm group was seen as a more appropriate denominator than was chronological age. By dividing the individual's score by the mean of the norm group, performance could be expressed relative to different groups rather than to just the one group on which the item placements had been determined.

As part of her SOMPA procedure, Jane Mercer (1979) adopted Yerkes' recommendation that special norms based on socioeconomic status be used. It is an interesting irony that while Mercer's use of Yerkes' idea has been

widely hailed as sensitive to cultural and social class differences, Yerkes himself has been depicted by some modern critics of the testing movement as the archfiend of testing in the 1920s.

ADVENT OF THE STANFORD REVISION

Terman had served notice as early as 1911 that he was developing a revision of the Binet–Simon scales for use in the United States. Terman and Childs (1912) had begun their work in 1910 using some of the tests that Terman had developed for his doctoral study as well as tests used by Binet and others. The process of moving from those early attempts to the final form of the scale that he published in 1916 (Terman, 1916) is chronicled by Terman et al. (1915, 1917). (Terman [1932] later noted that "it was Thorndike whose writings stimulated me most at this time, perhaps because I found myself in almost perpetual disagreement with him" [p. 325]. The basis of this disagreement was largely, of course, about the nature of intelligence and the best way to measure it. Terman's testing practices were almost perfectly aligned with Spearman's theory, and Thorndike was at the opposite theoretical pole.)

On the basis of a pretest with 310 California public school students from 1910 to 1912, items for the scale were refined, the age placement of some items was adjusted, and some new tests were added. The refined scale was administered to 982 white American–born schoolchildren, also from California. An effort was made to test only children who were within two months of their birthdays. This sampling yielded adequate results for ages up to 14, but in the first quarter of the century, school attendance patterns were such that older children who were still in school could not be considered representative of their age groups. In order to fill out the top of the age range, Terman (1916) and his associates tested 32 high school students, 30 owners of small businesses in Palo Alto, 150 "migrating unemployed" living in a "hobo hotel," and 15 juvenile delinquents. By modern standards this sampling of older subjects falls far short of the minimum for representativeness, but at the time it was the best that had ever been obtained. All responses were recorded verbatim, and the testing took about 50–60 minutes for each subject.

One of Terman's major objectives in his revision was to standardize procedures. With at least half a dozen forms of the Binet–Simon scale in use around the country, many in the hands of testers with little training, Terman felt that it was important for comparability of results that this variability be reduced. Already there was concern that unqualified people were making

unfounded claims for mental tests. To assure standard procedures for his testing, Terman used only examiners who had received at least six months of training in the use of the tests from him. Uniform scoring standards were guaranteed because all scoring was done by Terman himself. This scoring procedure violated one of Binet's original rules for testing—that all responses be scored at the time that they were made—but the variation is perhaps justified because these were not clinical appraisals.

Levels of performance were expressed as IQs. (This [Terman et al., 1915] is the earliest general application of Stern's suggestion that I have found. It was probably Terman's acceptance and vigorous defense of the IQ concept that was chiefly responsible for bringing the term into common usage.) Three times the tests were scored, item placements adjusted, and criteria checked before the final item arrangement and scoring criteria were determined. Item placements and validities were checked by three criteria:

> 1. Each item should show an increasing passing rate with increase in chronological age of the examinees.
>
> 2. Each item should agree with teachers' judgments of intellect or scholastic ability. That is, children who are judged bright by their teachers should pass the item more often than those judged dull.
>
> 3. The whole scale should be internally consistent. The percent passing an item should be higher for students with IQs above 110 than for those with IQs in the 90–110 range, and this group should have a higher passing rate than those with IQs below 90. "A test which satisfies this criterion must be accepted as valid or the entire scale must be rejected. Henceforth it stands or falls with the scale as a whole." (Terman et al., 1915, p. 553)

The publication of the Stanford Revision of the Binet–Simon Scale, as it was then called, marked a point of major importance in the history of mental measurement. With its formal appearance in 1916 the Stanford–Binet, as it soon came to be called, rapidly took over as the standard for intelligence measurement. With all its faults, it was considered so far superior to anything else then available that it quickly routed the alternatives. For the next 20 years any test that claimed to measure intelligence was judged largely by the degree to which its results agreed with those obtained from the Stanford–Binet.

Lewis M. Terman (1877–1956)

3

The Army Testing Program and Its Legacy

Although there is certainly no causal connection, the publication of the Stanford–Binet in 1916 was closely followed by America's entry into World War I. This was the first war in which psychology as a scientific discipline played any significant role. The tests that were developed during the war and the findings that were later based on them had an impact on pressing social problems in education and immigration, and on scientific questions about the nature and measurement of mental ability.

PSYCHOLOGY IN THE WAR EFFORT

As early as 1908 Binet had advocated the use of intelligence tests by the French army to discharge the mentally unfit. He noted that the Germans were already applying mental measurement to military matters and he strongly urged the French authorities to do likewise. Binet and Simon actually administered some preliminary tests to 11 soldiers, but the tests had to be carried out in the presence of a military psychiatrist. There was a strong resentment among French psychiatrists of the period against Binet and his methods, and Binet did not hold these colleagues in particularly high regard either. The upshot was that the military psychiatrist, a Dr. Simonin, presented a highly critical (and in a number of ways apparently inaccurate) report of the testing, and the power of the psychiatric establishment was sufficient

to kill any interest in Binet's proposal at the Ministry of War. As Peterson (1925) observed:

> This opposition, together with the death of Binet in 1911, left to the American psychologists in the World War the carrying out for the first time and on a big scale of a [testing] plan which in many of its details as to procedure and method was worked out by Binet, but which was also independently developed in America. (pp. 293–94)

American psychology went all out in behalf of the war effort. While some authors (e.g., Gould, 1981; Kamin, 1974) have implied that the effort was restricted to the development of mental tests (and it is undeniable that this activity had the greatest long–term and public impact), there were actually 13 separate committees devoted to applying psychological knowledge to military matters (see Yerkes, 1919). Each committee was chaired by a well–known expert in the relevant aspect of psychology: Raymond Dodge chaired the committee on perception and signal detection; Woodworth, the one on problems of emotional stability, fear, and self–control; Thorndike, the one on aviation psychology. (The reason for Thorndike's appointment is not clear, but one might suspect that his background in education and the army's need to develop pilot–training programs played a role.) Other committees were chaired by the likes of Watson, Terman, and Thurstone. Yerkes, who was then president of the American Psychological Association and chairman of the Committee for Psychology of the National Research Council, assigned himself the task of chairing the Committee on the Psychological Examination of Recruits, a role which his authorship of the Point Scale qualified him to play.

Prior to American involvement in the war there had been some developments in the area of group testing of intelligence. Norsworthy (1906) and Bonser (1910) had tested children in groups and related the results to scholastic performance. Colleagues and students of Thorndike at Teachers College had developed group measures of scholastic achievement, as had Vaney, under Binet's direction, in Paris. Workers at other institutions were also developing achievement or ability measures that could be used with groups. C. A. Scott (1913), in a study that seems to anticipate future developments, had given group tests of sentence completion, information, reading comprehension, meaningful memory, and "quickness" to students and found the simple sum of scores on these tests to correlate .70 with teachers' judgments of "brightness." Prior to the opening of hostilities, Pressey and Pressey (1918) had completed work on a group point scale of intelligence, including norms based on 1,100 students, but their test was intended for use

with schoolchildren and its publication was delayed by the war until 1918. Yerkes was able to convince the army that it needed a test for adults, and he and his associates took on the task.

Advent of the Committee

For many years the standard texts in psychological measurement have described the Committee on the Psychological Examination of Recruits in a way that implies that it was a united effort of the psychometric community. A recent article by von Mayrhauser (1987) provides an interesting alternative to this picture. Working primarily from diaries, minutes of meetings, and records of correspondence, von Mayrhauser paints a rather different scenario. He describes Yerkes' vigorous push for a testing program in the military as an attempt to sell the new technology of applied psychology and to advance his own personal career, which had suffered because his greatest interests lay in the area of primate behavior, while the emphasis at Harvard, where he held an assistant professorship, was "in a more human–focused psychology" (p. 131).

In his paper, von Mayrhauser documented the several meetings between members of the Executive Committee of the APA during the spring of 1917. The key issue was a disagreement between Yerkes and Walter Dill Scott over the appropriate wartime contribution that psychology should make. Yerkes strongly advocated developing mental (intelligence) tests that could be administered to groups. His background in a mental health clinic in Boston disposed him to propose that tests be used to eliminate the mentally unfit. Scott, on the other hand, had a strong reputation in the business community by virtue of his role in developing the personnel classification procedures of early industrial psychology. Thus, Yerkes' model was one of selection, while Scott's was one of classification or placement of each recruit into the job for which he was best suited and where his skills could be of greatest use. The disagreement between the two was never resolved, with the result that Yerkes chaired the Committee on the Psychological Examination of Recruits, while the Committee on the Classification of Personnel was headed by Scott. (See von Mayrhauser, 1987, for a description of the course and outcome of the debate.) There are substantial indications that the immediate impact of Scott's committee was greater than that of Yerkes', but there is no doubt that the long–term impact of the Committee on the Psychological Examination of Recruits has been more profound (which is not necessarily to say that it has been more beneficial).

Structure and Function of the Wartime Testing Effort

A description of the Committee on the Psychological Examination of Recruits and its work is given by Yoakum and Yerkes (1920):

> The committee consisted of R. M. Yerkes, Chairman; W. V. Bingham, Secretary; H. H. Goddard, T. H. Haines, L. M. Terman, G. M. Whipple, and F. L. Wells. Each of these men brought to the work of the committee a large amount of material which was sifted to produce the group and individual examining materials of the first "Examiner's Guide." Hundreds of tests already published were also available.... A complete group test, the work of A. S. Otis of Leland Stanford University, quite similar in form to that finally adopted by the Army was in manuscript. It was also drawn upon in making the army test. (p. 2)

DuBois (1970) gave a detailed description of the activities of the committee and of the sources of the various materials that eventually found their way into the final versions of the tests. The committee met for the first time in May of 1917, and by August of that year the tests were ready for a large scale tryout. Terman (1932) noted that during much of this period of frantic activity he lived with Yerkes in his home in Washington, D.C., and, although they continued to differ on the merits of the IQ and the CI, they became good friends. Yoakum and Yerkes described the army's initial reaction to the test tryouts:

> During October and November [the tests] were applied in four cantonments under conditions which could scarcely have been more unfavorable but with results which led the official medical inspector to formulate the following recommendations:
>
> "The purposes of psychological testing are (a) to aid in segregating the mentally incompetent, (b) to classify men according to their mental capacity, (c) to assist in selecting competent men for responsible positions.
>
> "In the opinion of this office these reports (accompanying recommendation) indicate very definitely that the desired results have been achieved.
>
> "The success of this work in a large series of observations, some five thousand officers and eighty thousand men, makes it reasonably certain that similar results may be expected if the system be extended to include the entire enlisted and drafted personnel and all newly appointed officers.

> "In view of these considerations, I recommend that all company officers, all candidates for officers' training camps and all drafted and enlisted men be required to take the prescribed psychological tests." (Yoakum & Yerkes, 1920, p. xi)

The work of the committee ultimately resulted in five equivalent forms of an examination for literate recruits (Form Alpha), which came to be called the Army Alpha. For those recruits who did not speak English, could not read, or got low scores on Form Alpha, the committee provided a "performance" test known as Form Beta (the Army Beta). Form Beta was modeled on a test developed by Rudolph Pintner and Donald Paterson (1915, 1917) for use with deaf subjects. It employed a variety of form boards and mazes, including some of those developed in Australia by Porteus (1915). Administration of the test required no use of language. Instructions were given in pantomime. For those who still appeared mentally unfit to serve in the army, the plan was to examine them with the Stanford–Binet.

Impact of Testing on the Army

Gould (1981) gave a vivid and colorful account of the problems that the testing program encountered. His book, in which he chastises others for letting their own preconceptions color their interpretations of facts, contains such statements as "Yerkes brought together all the major hereditarians of American psychometrics to write the army mental tests" (p. 194) and "As camp followers themselves, Yerkes corps decided to test a more traditional category of colleagues: the local prostitutes" (p. 197). While the first statement is demonstrably false (as are a number of others in the book) and the second is at least partially true, Gould, functioning as a historian, is guilty of the sins he finds in others.

Moreover, the tone of Gould's statements is very much out of keeping with the scientific thinking prevalent during the first quarter of this century, and the intellectual climate in which the army testing program was conducted. Watson's radical behaviorism and his assertions about the modifiability of behavior notwithstanding (remember that he was also acting as a psychologist in the service of the army, but on a different committee), the majority of scientists of the time believed that science was on the verge of answering all of the questions about the nature of the universe and that many aspects of human behavior were profoundly affected by heredity. That they were wrong in some of their beliefs and theories can hardly be used as a criticism of their intent within the context of their own times. Rare indeed is the prophet who accurately perceives the zeitgeist of a future era.

Wolf (1973) noted that "like so many of his fellow scientists at the turn of the century, Binet was sure that [the scientific] spirit would lead mankind out of the darkness" (p. 294) and that scientific answers to many problems of great social import were just around the corner. The literature of the time is rife with expressions of such a belief, so it is little wonder that the army psychologists were of a similar opinion regarding the future of mental tests.

While the real political and social consequences of the army testing program can never be known, it is interesting to contrast the following two assessments, the first from Gould (1981), the second from Dodge (1920):

> I do not think that the army ever made much use of the tests. One can well imagine how professional officers felt about smart–assed young psychologists who arrived without invitation, often assumed officer's rank without undergoing basic training, commandeered a building to give the tests (if they could), saw each recruit for an hour in a large group, and then proceeded to usurp the officer's traditional role in judging the worthiness of men for various military tasks. (Gould, 1981, p. 194)

> In an address at the Personnel Officers' School . . . Major–General Hutchinson, C.B.D.S.O., Director of Organization of the British Army, spoke very frankly of the serious mistake of Great Britain in recruiting her skilled labor indiscriminantly into fighting units. They made good soldiers, but the plan seriously interfered with the development of technical units and the "output of many vital things."
>
> . . . If it had not been for the great American reservoir of skilled labor it would have probably cost the war. That the United States did not make a similar, and with the exhaustion of the reservoir, a disastrous mistake in the military distribution of our skilled labor is due primarily to the Committee on the Classification of Personnel in the Army [the committee working under Scott]. . . .
>
> The work of the Committee on the Psychological Examination of Recruits was another of the notable mental engineering achievements of the war. Its original purpose was to help to eliminate from the army at the earliest possible moment those recruits whose defective intelligence would make them a menace to the military organization. But the military value of an early and reliable estimate of the general intelligence of each recruit proved enormously greater than had been anticipated. (Dodge, cited in Yoakum & Yerkes, 1920, pp. 184–185)

For better or for worse, the army testing program in the First World War profoundly changed the mental testing movement in the United States. Before the war, testing had been going on at a rapid pace, but it was primarily individual testing and confined to clinics or a few research applications in schools. However, with the development of easily administered group tests it became possible to test large numbers of people, both students and workers. The army psychologists were well aware of the potential misuses that could be made of test information:

> The ease with which the army group test can be given and scored makes it a dangerous method in the hands of the inexpert. It was not prepared for civilian use and is applicable only within certain limits to other uses than that for which it was prepared. (Yoakum & Yerkes, 1920, p. 2)

The warnings of the army psychologists notwithstanding, mental testing moved rapidly into the public domain in the years following the war. This may be attributed in part to the nation's entrepreneurial spirit. There was a need, a market for tests, and if American psychology did not move to fill it some other source would. (Occasionally even the most devoted academics and clinicians are not above writing a book, or a test, if they think it will sell. The psychometricians of 1920 were not the first, and the practice has been repeated with increasing frequency since.) There was also a perceived national need for tests that would provide answers to pressing social problems and scientific questions.

CONCERNS AND CONFUSION CAUSED BY TESTING

Concerns in the Educational Community

Binet's position. Ever since the rise of elective courses and vocational education in high schools, which began around the turn of the century, educators had been looking for tests to aid in the selection of those who would be encouraged to pursue higher education and to provide vocational guidance. Education reform was seen by many as a national priority, and intelligence tests were hailed as an important tool of the reform. As far back as 1909, Binet, in *Les idées modernes sur les enfants*, had "prescribed the determination of the global intelligence of pupils by using the Binet–Simon metric scale . . . recommending it to parents as a means of estimating their children's intelligence" (Wolf, 1973, p. 317).

Binet also gave his criteria for a good school. "There should be objective evaluations of everything possible, primarily to test the teachers'

competence. Pupils in each grade could be tested by means of measures of their achievement in all school subjects" (Wolf, p. 316). Binet was vitally concerned about providing each child with an education focused at his or her level. He believed that the curriculum should be arranged so that school subjects were taught only after the necessary mental apparatus had developed, and he advocated a multi–track system. While recognizing the educational needs of the mentally retarded, Binet "may have been one of the first to urge the organization of special classes for the 'above average.' He argued that it is 'through the elite, and not through the efforts of the average that humanity invents and makes progress,' and therefore, children with superior intelligence 'should receive the education that they need' " (Wolf, p. 318). Peterson (1925) observed that "one of the chief post–Binet developments of mental testing has been a more explicit recognition of the problem of selecting and training super–normal children according to their greater possibilities" (p. 287).

Terman and educational reform. Terman's fluent French and his devotion to Binet's approach to intelligence measurement virtually guarantee that he had read the Frenchman's ideas on the application of mental tests to problems of the schools and was in sympathy with them. In fact, like many other American psychologists with public school contacts, Terman believed that testing students and putting them into their proper academic tracks was economically necessary in order to get the most for the educational dollar and was socially appropriate as the way to assure equal educational opportunity for all.

A warning bell was sounded in 1910 by Ayres. His book, *Laggards in Our Schools*, brought the problem of low school achievement to national prominence, but he, like many others of the period, attributed the problem to poor school management, bad teaching, and low motivation. Terman (1919) demurred:

> It will be shown that these innate differences in intelligence are chiefly responsible for the problem of the school laggard. (p. 24)

He added,

> Our tests show that 90 percent, at least, of school retardation is without doubt due to mental inferiority. . . . One of the results of placing children of the same mental age together has been the cutting down of failures by fully 50 percent. (p. 303)

If one ignores the reference to the still hotly debated topic of innateness, Terman's tone is definitely that of an educational progressive, at least for 1920.

Terman had great plans for mental testing in the schools. He believed that a mental age of seven was necessary for a child to do average first–grade work, and that "the greatest usefulness [of intelligence tests] will be found in their universal application to school children. . . . 'A mental test for every child' is no longer an unreasonable slogan" (1920, p. 20). As a start, he proposed that:

> All the pupils in the fourth grade and beyond should be given a test by the group method every year, and those whose scores are either very high or very low in the group examination should be given a Binet test. . . . It is [also] highly desirable that every pupil be given a mental test within the first half–year of his school life. (1919, pp. 15–16)

The use of intelligence tests in the schools spread like wildfire. "Probably a million children in the schools of the United States were given a group mental test during the year 1919–1920. In 1920–1921 the number was probably not less than two millions. We may expect the number to exceed five millions within a few years" (Terman et al., 1922, p. 3). However, Terman also sounded a mild note of caution: Intelligence testing could not bring about the educational millennium, but it could certainly be used to make schools more efficient. Terman believed students could learn more in the instructional time allotted.

Reform–minded educators, although they harbored some quaint ideas ("It is very generally believed that adenoids seriously retard mental development and that their removal is nearly always followed by a marked intellectual awakening" [Terman, 1919, p. 151]), were actively using mental tests to evaluate their educational reforms. Noting that "there is general agreement among teachers that the difference in mental level is the chief cause of trouble in the average classroom" (Dickson, 1922, p. 33), the entire city of Oakland, California, moved gradually onto a three–track system of advanced, normal, and limited classes, starting in 1918. Berkeley followed two years later. Up to this time, the trend had been to keep every student in school to a later age, and failure had become so common that it was not unusual to find students 3 or 4 years overage for their grade. It was a widespread fear among educators that standards would be lowered so that the less able could be promoted. The Oakland Plan was based on the strangely modern–sounding premise that "the high school must classify according to brightness and must offer modified courses of study, or the present standards for academic work will fall" (Dickson, 1922, p. 50). To avert such action, school districts all across the country were trying out variants of this plan.

While educational reform using tested mental ability was a major practical product of the wartime testing program, two other issues shared the arena of national attention. These, both with roots that extended back well before the war, were the developmental question of when intellectual growth reached maturity, and the social policy question of whether there were "racial" (read "national") differences in intelligence. The former is a valid scientific question to which we are still seeking answers, and the latter, while also a legitimate area of research, provides in this instance an illustration of one of the most flagrant misuses and misinterpretations of psychological data in the history of American psychology. They both reached a climax with the publication of the reports and interpretations of the army data.

The Attack on Mental Age

One of the features of human development that allowed Binet to establish his scale in the first place was the relatively regular progression of *average* development. Binet, Thorndike, and others had shown that although there was considerable variability around the average for any age, the mean (or median) for older children uniformly exceeded that for younger children on any given intellectual task. The whole concept of an age scale depended on this regularity. Once development stopped or substantially changed in rate, the meaning of a mental age was lost.

The pattern of mental growth. Numerous studies had been conducted on the development of physical attributes, and the growth patterns showed a certain regularity. There was a natural tendency to assume that some similar growth pattern would characterize mental ability. In his early studies with children, Binet found a fairly regular decline in the difficulty level of a particular task with an increase in age. His samples were too small and too restricted to allow him to draw very definite or general conclusions, but he felt that a person's mental level increased until the mature level was reached, at which time the growth of the intellect stopped. The 1908 scale was standardized "on the laboring and small merchant classes. In this population . . . a level of twelve years appeared to be the normal one [for adults]" (Wolf, 1973, p. 256).

Since most uses of the 1908 scale and its American successors had been with the mentally retarded or with school–aged children, relatively little was known about the upper levels of mental ability until the army studies of World War I. School attendance patterns were such in the first quarter of the century that it was not possible to obtain anything like a representative sample of children over age fourteen from the schools, which is why Terman was forced to use the bizarre collection of hobos, businessmen, and so forth for his upper

age levels. From this sample he concluded that the growth curve for intelligence leveled off at about 16 years of age. That is, average performance of adults on the Stanford–Binet was at the same level as that of 16–year–olds.

Both the concept of mental age (MA) and that of IQ required that one know the chronological age (CA) at which development stopped. In Terman's scheme of things, the tests were so arranged that the median mental age of children at a given chronological age was equal to their chronological age, making the median IQ for each age 1.00. Since the IQ was computed as MA divided by CA, and since CA continued to increase at a constant rate until death, the meaning of the IQ changed once mental age stopped increasing. Many objectors to the IQ concept, including Yerkes, made this point early and often. Terman believed that by using the age of 16 as the denominator for all IQs of people aged 16 or older he could avoid the problem and retain the IQ as a useful index.

Setting the mental age of the average American adult at 16 presented little difficulty so long as testing was restricted to the schools. However, with the publication of the army data, and with the appearance of *The Revolt Against Civilization* by Stoddard in 1921, a full–fledged attack on the mental testing movement (including the IQ index) was mounted. Most famous of the critics was a young commentator for the *New Republic* named Walter Lippmann. In a series of articles in 1922 he attacked several interpretations of the army data (while acknowledging that the testing program itself had been of significant use for its original military purposes). One main object of his attack was the conclusion that the average mental age of adult American males, as represented by the army draft, was 14. He suggested, with some justification, that this finding should call into question the standardization of the Stanford–Binet.

The army grading system. In order to cast the army tests into a metric that could be related to other mental tests already in use, a study had been conducted early in the army testing program in which 653 men were given both Form Alpha and the Stanford–Binet. On the basis of these test scores, a table of equivalents and a regression equation were prepared from which mental ages could be estimated from Army Alpha scores. By converting the frequency distribution of the Army Alpha scores into one of mental ages, the army psychologists concluded (Yerkes, 1921) that the average mental age of the white draft was about 14 years. When blacks were included, the average was somewhat lower. Lippmann argued that using mental ages from the Stanford–Binet standardization was inappropriate and that the army results should be taken as more representative, which, of course, they were.

The army program tested a large number of recruits (roughly 1.7 million). The draft was structured so that men in critical civilian jobs would be exempted,

but otherwise it drew from a broad cross–section of young American manhood. Since this was a time of heavy immigration from southern and eastern Europe, there was a large number of adult males living in the United States whose command of English was far from perfect. Also, educational opportunities for blacks and immigrants were often slight or nonexistent, resulting in many illiterate draftees. Form Beta was designed to meet the need for a way to measure the intelligence of such men. There was a table of equivalents by which these scores could be cast into mental ages as well. The claim that the mental age of the average American adult was 14 years, which is how the results were portrayed, was based on these tables of equivalents.

In order to translate the test scores or mental ages into terms that could be used by the army for decision making (and to simplify later character assessments), a rough letter–grading system was employed. Those at the top, roughly 5 percent, were classified as "A men"; about the next 10 percent as "B men"; roughly 15 percent as "C+"; 25 percent as "C"; 20 percent as "C–"; and the rest as "D" or "E." "A men" were considered first–rate officer material, and lower grades were assigned duties requiring successively less independence and judgment. The grades were determined by the *assumed* properties of intelligence; a normal distribution and a high relationship to efficiency in performing military tasks. Critics such as Lippmann attacked this grading system as arbitrary, which it was, but they could not deny that those who scored higher on the tests generally received higher ratings from their commanding officers.

Lippmann's attack. The point that drew the most strident criticism from both in– and outside the psychological community was the assertion by several writers that the army results reflected genetically determined and immutable levels of intelligence. While Lippmann (1922) acknowledged that "a fair reading of the evidence will, I think, convince anyone that as a *system of grading* the intelligence tests may prove superior in the end to the system now prevailing in the public schools" (p. 277), he also asserted that "the whole claim of intelligence testers to have found a reliable measure of human capacity rests on the assumption, imported into the argument, that education is essentially impotent because intelligence is hereditary and unchangeable" (p. 277).

Lippmann's argument at this point struck directly at the issue of constancy of IQ. It was a common belief at that time that a child could not be educated beyond his or her mental age. Thus, a child with a mental age of six could not be taught subjects beyond those of the first or second grade. (Binet himself gave indirect support to this idea when he argued that schools should not attempt to teach subjects to those whose mental abilities had not developed sufficiently, but he did not carry out the argument to a terminal level.) Furthermore,

since many children seemed to top out in mental growth by about age 14, there was no reason to send such children to high school. The more able students (those with mental ages above 14) could receive a better education if the less able were expelled. Bagley (1922a, 1922b) joined Lippmann and others in strongly rejecting the idea that the less able should be forced out of school, arguing that education could be effective at all levels of ability. Terman (1922) replied that "far from encouraging teachers to neglect the dull as unworthy of their efforts, the psychologist of individual differences believes that the one purpose of intelligence tests in the schools is to aid us in making the most of every child, the dull as well as the bright" (p. 62).

The crux of this aspect of the debate over the army test data seems to have been two–fold. First, Terman and other proponents of mental tests advocated ability grouping in the schools using test results. The idea of ability groups implied permanent differences in ability, an idea which many critics rejected. Lippmann wrote of predestination and infant damnation as logical outgrowths of ability grouping. Second, many educators were beginning to argue for a program of vocational education to replace academic subjects in the high schools for students of low measured intelligence. Terman urged both ability grouping and vocational education:

> Instead of being undemocratic, as some have argued, such differentiation of courses and enlargement of opportunities for vocational training of the humbler sort is a necessary corollary of the truly democratic ideal. (Terman, 1919, p. 91)

In one of his less accurate prognostications he even suggested that:

> The evolution of modern industrial organization together with the mechanization of processing by machinery is making possible a larger and larger utilization of inferior mentality. . . . It is even suggested that our chief difficulty may soon be to provide enough suitable jobs for those of higher intellectual capacity. (p. 276)

As is often the case when a technological advance becomes widely available, there were soon all sorts of fantastic claims being made for intelligence tests. Wallin (1923) complained that "one writer states that . . . any first grade teacher can on the first day of school after a 20–minute examination classify her pupils in regard to their intellectual ability, and section them accurately for the purpose of instruction" (p. 232). Statements like this could be considered frivolous were it not that some people still believe them. (See Rist, 1970, for a detailed description of an even worse methodology in action half a century later. His study found a teacher forming ability groups in her class without any objective evidence.)

Much of the debate about the average mental age of adults, as well as that about the constancy of the IQ, resulted from problems of definition. Binet's use of mental levels was hailed at the time as a breakthrough because it made the metric for intelligence measurement easy to understand. However, there were alternatives even in 1908, and by the time Terman released the Stanford–Binet, mental age and its related IQ had probably outlived their usefulness as an easily understood scale. Certainly, by the time the army data were published there was really no excuse other than habit for expressing performance in the age–scale metric. Nevertheless, it was true that the mean performance of army recruits on the test was about equal to that of 14–year–olds. Had the results been expressed in that way, this phase of the scaling argument, with its attendant national breast–beating, could possibly have been avoided. Likewise, the controversy about constancy of the IQ is considerably defused if the issue is expressed as an observation that children, on the average, tend to retain the same relative position in a group of their peers over a period of years, rather than that their IQs remain constant.

Mental Testing and Immigration Policy

The other use to which the army data were put, and which constitutes a low point for such research, was in the service of the debate over immigration policy. The tide of immigrants had been flooding for years, and there was widespread concern, strongly voiced by Stoddard in the book that drew Lippmann's fire, that the intellectual quality of recent arrivals was below that of their predecessors. As early as 1913 the American Medical Association, in response to concern that the mental quality of immigrants was dropping, editorialized that "inspection by experts at the port of entry will result in a much larger percentage of defective immigration being detained" (AMA, 1913, p. 209). The experts were to be medical men who would administer a battery of physical and psychological tests to screen out the mentally unfit. As part of this program, Knox published a nonverbal test in 1914 to be used with immigrants (cited in DuBois, 1970). Mullan (1917) reported on a small–scale study that was conducted under the auspices of the U.S. Surgeon General to determine if various tests could detect immigrants who were mentally unfit. A wide variety of tests, many of school–related subjects, was given to 293 immigrants on their second day at Ellis Island. Mullan, in a tone reminiscent of Binet, observed that "an immigrant's reply cannot be accurately analyzed or graded by a rigid standard. It must be considered and weighed in conjunction with many other factors, and for this the judgment of a trained and experienced diagnostician is needed" (p. 23). The concern with denying admission to those of low mental ability was definitely an item of governmental atten-

tion, and research had started on a test–oriented response to the problem before the research program that led to the Army Alpha had begun.

After the war the rate of new arrivals remained high, and there was pressure in Congress to pass legislation to restrict immigration. Carl Brigham (1923) analyzed the army data from the point of view of the national origin of the recruits. His analysis, which bears all the marks of Binet's defense of animal magnetism, is a study in finding what you are looking for. Starting with the generally held belief in innate intelligence, he found substantial differences between groups of different national origin. In general, more recent immigrants, who tended to be from southern or eastern Europe, earned lower test scores. Brigham interpreted these results to mean that people with southern or eastern European backgrounds were genetically inferior to those from northern or western Europe, a conclusion that was generally welcomed by the American political power structure of the day.

Brigham tried in several ways to test the hypothesis that environmental factors affected performance, but his preferred explanation always came back to "racial" differences. "Mediterraneans," "Slavs," and "Alpines" were found to be inferior to "Nordics." The later pages of Brigham's book contain some quotes and assertions guaranteed to insult just about everyone who is not of English, German, or Scandinavian descent.

In 1924 Congress passed the Immigration Act, which limited the number of immigrants from Europe and denied Asians even the opportunity to apply for citizenship. Neither Brigham nor the testing movement, in which many leading figures strongly believed in a large hereditary component of intelligence, can be held responsible for the Immigration Act. The legislation would have become law with its provisions intact, even if psychologists had argued the opposite side of the issue, because it was a politically popular solution to what was widely perceived to be a national problem. Fear of inundation of the country by great "masses of the unwashed" was rampant. Discrimination on the basis of national origin, with access being limited to those nationality groups already here, was at the core of the program, and its proponents were glad to have the army data to support them.

Many of the leading psychologists of the time agreed with the objectives of the proposed legislation, but to assert, as Kamin (1974) did, that the psychologists or their data were substantially responsible for the fact or form of the act is to ascribe to them far more impact than they had. Samelson (1975, 1982) has provided a thorough refutation of Kamin's claims. (By contrast, most of the other "Western democracies" had then and still have today much more restrictive policies on immigration than were contained in the 1924 act, policies that were based on nationality and often on financial status. In some

parts of the world racist practices are explicitly a part of official policy. This does not make the Immigration Act of 1924 right; it merely demonstrates that the policy embodied in the act is probably more the rule than the exception in human affairs.) It is unfortunate that psychology became involved in this matter, and it is doubly unfortunate that the data, whatever their quality, should have been so badly misinterpreted. Even Brigham's retraction in 1930 of his 1923 position was really done for the wrong reason, but at the time the methodology for the necessary analyses was not widely known, and the data were taken to show something that they did not show.

TESTING DEVELOPMENTS IN THE 1920s

The 1920s was a period of exploration and expansion for psychological testing, despite the occasional criticism it received both from the public and from its practitioners. The success of the army testing program and the needs of an educational system that was expanding to keep students for a longer period and was concerned about placing them in ability tracks created an atmosphere favorable to the further development of group tests of intelligence. By 1931 Pintner was able to list 37 fairly prominent group tests measuring a variety of functions relating to general intelligence. Most of the tests were similar in many ways to the Alpha, but other item types and test structures were being tried out.

One of the best and most widely used group tests, the National Intelligence Tests, was developed by Haggerty, Terman, Thorndike, Whipple, and Yerkes (Whipple, 1921). It consisted of five equivalent forms and, following Thorndike's strongly held belief in the need to equate examinees for testwiseness, included a large number of practice items. Considering that all of its authors had worked fairly directly with the development of the Army Alpha, it is little wonder that there was considerable similarity between them.

The CAVD—A Test Ahead of Its Time

One of the most ambitious test development programs of the period was going on at Teachers College, Columbia University, although it is interesting that the resulting test, the CAVD, never had much public impact. Its development, however, did add new understanding in several areas such as the definition of intelligence, selection of appropriate test items, and item scaling. While some of the scaling features of the resulting test series were far from satisfactory, the basic design of the instrument is very similar to ones now being developed, including the Fourth Edition of the Stanford–Binet.

The role of value. In 1904 Thorndike had called attention to the desirable properties that a mental measuring device should possess. Twenty–two years later he offered an instrument designed to fulfill those conditions. Most test development up to that time had been empirical, being based on items that followed the expected pattern of decreasing difficulty with increasing age. Binet's scales were without a theoretical basis, as were those of Terman and most others. However, Thorndike (Thorndike, Bregman, Cobb, & Woodyard, 1926) observed that this atheoretical definition of intelligence in terms of the operations that were used to measure it (see Boring, 1923) actually involved some a priori definitions about what was valuable as intellectual behavior:

> Valuation came in from the start because Binet tried only abilities which he valued as intellectual. He did not take *all* the psychological features of five–, six–, and seven–year–olds and choose as his series of tests those which separated the ages most distinctly. In revising Binet's series Terman and others have paid less and less attention to lateness of development and more and more to significance as valued symptoms of intelligence in their choice. (Thorndike et al., 1926, p. 16)

This matter of valuation, Thorndike pointed out, made the definition of the tasks used to measure intelligence depend upon the cultural context:

> What abilities and tasks shall be treated as intellectual is essentially a matter of arbitrary assumption or choice at the outset, either directly, of the abilities or tasks themselves, or indirectly, of the consensus which provides the criterion. After the first choice is made, tasks not included in it, and even not known, may be found to correlate perfectly with the adopted total, and so be "intellectual"; but their intellectualness is tested by and depends on the first arbitrary choice. Had a different first choice been made, they might not be intellectual. (p. 61)

Dimensions of intellect. Thorndike's theory of *intellect* (the term he preferred) held that there were four general dimensions to intelligent behavior: altitude, width, area, and speed. *Altitude*, whose limits he felt were genetically determined, referred to the complexity or difficulty of tasks that an individual was capable of performing. *Width* referred to the number of tasks at a given level of difficulty that the person could do and, at a specific altitude, was largely dependent on experience. *Area* was seen as a function of altitude and width, but the possible width of intellect varied with altitude, being greater at higher altitudes. Finally, *speed* or rate was the number of tasks of a given kind and complexity that the person could perform in a given unit of time. The complete description of a person's intellect required assessment of all four dimensions, but they were not seen as being of equal importance. "Common

sense considers extent and quickness [of intellect] as unimportant in comparison with reaching a level far above the average" (Thorndike et al., 1926, p. 35).

In addition to the four general properties of intellect, Thorndike's theory held that there was an intellect relatively specific to each type of task. Consequently, the test that he and his associates constructed was really a set of four parallel test series, each designed to measure one type of intellect. The four intellects that the authors chose—Completion, Arithmetic, Vocabulary, Directions—gave the instrument its name, CAVD. In discussing the test and what it measured, Thorndike always referred to "intellect CAVD" to emphasize the fact that general intelligence was not being measured, but that the test tapped a particular subset of the vast array of intellectual abilities that one might choose to measure. He selected these in part for demonstration and in part because they seemed, from empirical study, to be particularly useful in predicting academic performance.

The various dimensions of intellect were positively correlated: "Our experiments . . . indicate that intellect has a rather high degree of unity and consistency and independence of non–intellective factors" (p. 63). This represents a shift from the extreme position he had held 15 years earlier, and it opened the door for an eventual reconciliation with Spearman through the work of others, notably Thurstone.

Although he viewed altitude of intellect, and possibly some aspects of its width as well, as being limited by genetic factors, and while he harbored eugenic sentiments, Thorndike was an empiricist in his science. He did not believe that intelligence tests did, or could, measure intelligence directly, and he argued that it was not possible to construct a useful test that was independent of environmental factors:

> We are measuring available power of intellectual achievement without any specification as to its genesis. A person who has acquired the intellectual tool, reading, probably has a considerable advantage over one of equal original capacity who has not acquired that tool. . . . There is also danger that, if we include in a series of intellectual tasks only those in whose accomplishment differences in education can make little or no difference, we shall have a collection of freakish puzzles, irrelevant to the actual operations of intellect by persons twelve years or older in the United States to–day—or possibly have nothing at all. (pp. 95–96)

Structure of the test. The resulting test, the CAVD, had an interesting structure. It was composed of four separate sequences of tasks, one for each dimension. Each sequence was made up of 170 items. The items were grouped into 17 levels—10 items at each level. All items at a level were of

equal difficulty, and the levels had been scaled so that the increment in difficulty from one level to the next was uniform. The 17 levels of the test purportedly covered 20 equal units on the scale of intellect.

The purpose behind the CAVD was to construct what in modern terms is called a ratio scale for the measurement of the altitude of intelligence. Once Thorndike had developed the interval scale of item sets, it remained for him to determine the point at which intelligence reached absolute zero. This he did by using a group of judges who rated the amount of intelligence needed to perform various tasks. The tasks ranged from very difficult to very simple responses, and the scaling resulted in a zero point slightly below the behavior of earthworms (i.e., making simple responses to simple stimuli). The final scale had 23 units between zero and the lowest level on the test, for a total of 43 steps of intellect from zero to the highest level measured.

Another innovation in the CAVD was the way in which a person's level of intellect was determined. Applying logic from psychophysics, Thorndike reasoned that a person should have a 50 percent chance of success on those items that were exactly at his or her level of ability when guessing was not a factor. This would only be possible when the items were scaled independently of the group to which they were being administered. With 10 items on each dimension, there were a total of 40 items at each level of the test, and it was possible, using information from several test levels, to estimate the scale value at which a person would have a 50 percent success rate even if that point fell between two test levels. The resulting score was a value on a ratio scale that ran from zero to 430, indicating the altitude of intellect on each dimension.

The CAVD never caught on with the testing establishment, probably because it was a complex and cumbersome instrument to administer and score and its scores were not expressed in the familiar metric of mental ages and IQs. The ill fate of the CAVD was unfortunate because, although its scaling method was crude, the test was almost 50 years ahead of its time.

In 1925, Thurstone had presented a method to determine scale units that was similar to the CAVD's, but better. Unfortunately, widespread acceptance of this approach to psychological scaling did not come until Rasch's work many years later (Rasch, 1960). In 1928 Thurstone also published a better method of determining the absolute zero of intelligence, a method which was based on the fact that the variability of performance increased with mental age. By extrapolating back from existing data, he was able to estimate the age of zero intelligence at or shortly before birth!

Although it did not incorporate Thurstone's scaling improvements, the basic design of the CAVD clearly foreshadowed elements of the adaptive testing movement of the 1970s, even of the Fourth Edition of the Stanford–Binet, the test which, in its first incarnation and with Terman's strongly held belief in the value of mental age and the ratio IQ, was probably most responsible for the inability of the CAVD to win the acceptance of the psychological community.

Louis L. Thurstone (1887–1955)

4

New Intelligence Studies and Tests

EVOLVING THEORIES OF INTELLIGENCE

The period of the late twenties and thirties saw some major progress in the theoretical basis for intelligence tests. One of the features of intelligence testing that had drawn considerable criticism was the lack of a sound theoretical base for the tests. Binet's approach was theoretically agnostic and avowedly empirical. If tests differentiated in the expected way, he used them. Other authors adopted Binet's tests, but not his empiricism, and, as Tuddenham (1963) noted, "the basic difficulty stems from the attempts by most workers over the years to substitute Spearman's theory of intelligence for Binet's, while continuing to use tests founded on Binet's pragmatic measuring instrument" (p. 502). The two major theories of the twenties and thirties were Spearman's two–factor theory, postulating general intelligence (*g*), and Thorndike's theory of multiple neural bonds, which rejected any general intellective factor.

Binet was ambiguous in his definition of intelligence. At some points he treated it like Spearman's *g*, while at other times he was a "faculty psychologist." Spearman (1931), in exasperation while defending himself from a charge that he borrowed the idea of *g* from Binet, quoted Binet as saying: "Almost all phenomena with which psychology is concerned are phenomena of intelligence" (p. 403). Spearman went on to say:

> So far as theorizing is concerned, his [Binet's] adoption of our general factor was only half-hearted. He was torn in two opposite directions. On the one hand, he naturally endeavored to bring his utterances into harmony with his new metric scale. But on the other hand he continually regressed back to the old faculty doctrine; for upon this had been built up, and now stood irrevocably founded, his whole general psychological outlook. (p. 404)

The Spearman–Thorndike Debate

The fundamental contest over the nature of intelligence was fought out across the Atlantic between Spearman and his students in England, and Thorndike and his students in the United States. Binet did not really leave any theoretical heirs in Europe. His colleagues (such as Simon) were practitioners and did not make further advances from the frontiers Binet had defined. Although Terman and the other adapters of the Binet–Simon scales in America remained true to the idea of a single entity of intelligence, they did not enter the controversy. Of course, each practicing psychologist held his or her own particular position, but in general the division broke down into the two camps.

***Thorndike rejects* g.** In 1909, Thorndike, Lay, and Dean had tested the hypothesis of *g* on a set of measures similar to those Spearman (1904a) had used in his original study and found no support for the two–factor theory. However, at the time, the criteria for testing the theory were poorly specified. Spearman simply held that the matrix of correlations should show an order; it should be possible to arrange the variables so that the matrix of their correlations would show the highest values in the upper left–hand corner and the magnitudes of the correlations would decrease as one went down any column or across any row.

Both studies (Spearman 1904a, and Thorndike 1909) used similar tasks to measure intelligence—some Binet–type tasks and certain sensorimotor tasks; however, Spearman used a sample that included both boys and girls and that was heterogeneous with respect to age. Thorndike seems always to have had access to large numbers of subjects, and in this case he was able to apply the tests to two homogeneous samples, one of college women and one of high school boys. Rather than finding *g*, Thorndike concluded:

> In general there is evidence of a complex set of bonds between the psychological equivalents of both what we call the formal side of thought and what we call its content, so that one is almost tempted to replace Spearman's statement by the equally extravagant one that there is *nothing whatever common to all mental functions, or to any half of them.* (Thorndike, Lay, & Dean, 1909, p. 368)

At about the same time that Thorndike and his associates had determined that Spearman's theory was wrong, Burt (1909) conducted a study in which he contrasted the performance of two groups of English schoolboys on a series of mental tests. He concluded from his results that there was a single common factor in his measures and that it was largely of hereditary origin.

The tetrad equation. Spearman (Hart & Spearman, 1912) responded to Thorndike's criticism by pointing out that Thorndike had not satisfied one of the conditions for testing the two–factor theory—that the test battery not contain more than one test of a particular type. In his paper Spearman also presented the derivation of the mathematical criterion for the theory, his famous tetrad equation. Given a set of correlated variables (*a, b, c,* and *d*), Spearman argued that the theory of two factors (*g* and *s*) was supported if all equations of the following form were satisfied:

$$r_{ab}r_{cd} - r_{ac}r_{bd} = 0$$

(r_{ab} is the correlation between variable *a* and variable *b*, and so forth.) Of course, in any set of real data the tetrad criterion would be satisfied only approximately. Spearman asserted that Thorndike's data satisfied the criterion when redundant variables (those measuring functions that were too similar) were eliminated.

As part of a larger study, one of Thorndike's students (Simpson, 1912) conducted a further test of Spearman's hypothesis. He gave 15 tests of various sorts (which he had grouped a priori into six categories) to 17 "good men" (bright college students and Columbia faculty) and 20 "poor men" (unemployed men and laborers). In what amounted to an early factor analysis, Simpson found that the tests in his a priori groupings "held together" in logically meaningful groups. (His "factor analysis" was entirely visual rather than mathematical.) From this analysis he concluded that "it is quite evident that Spearman's theory is not in harmony with the facts we have secured" (p. 91), but he added:

> There is a close inter–relation among certain mental abilities, and consequently a something that may be called 'general mental ability' or 'general intelligence'; and that on the other hand certain capacities are relatively specialized, and do not necessarily imply other abilities except to a very limited extent. (p. 109)

Simpson's conclusions in 1912 anticipated Holzinger's (1938) 25 years later.

Simpson's test of Spearman's theory was based on the criterion of a hierarchy or order of the correlations, and it did not take Spearman long to point out that he had provided a better criterion. In 1914, Spearman responded with a

review of his tetrad equation and he extended it to larger data sets. He then applied the method, adding a few restrictions of his own (such as deleting tests that were too similar) to Simpson's data and found, much to no one's surprise, that Simpson's data fit the two–factor theory.

About this time war in Europe began to interfere with the battle of *g*, and little more was said by the major combatants until 1920. However, in 1916, a Scottish psychologist named Godfrey Thomson entered the lists in opposition to Spearman and showed with a set of artificial data that it was possible to produce correlations that fit Spearman's requirements without resorting to *g*. Thomson argued that the Hart–Spearman proof of the sufficiency of the tetrad criterion was valid only when it was satisfied exactly, which never occurred in actual practice. He also produced a correlation matrix that was based on dice (which few would argue possess *g*) and that fit the model. Thomson had designed his tests to contain correlated "group factors" that were clearly separable, but the fact that the factors were positively correlated gave the impression that *g* was present. (Thirty years later, Thurstone, 1947, used the same type of approach in his well–known box problem, in which he analyzed the measurements of sets of boxes, to make a similar point about the theory of multiple factors.)

Spearman claims victory. The end of World War I saw a resumption of the argument over the nature of intelligence. In America the Stanford–Binet was considered by many to be a measure of general intelligence, and its widespread acceptance as the standard against which other instruments were judged helped Spearman (1920) to conclude that "among the most unexpected events in the psychology of the last dozen years has been the sudden spring of 'general intelligence' from an almost universal incredulity to no less universal investment with the highest importance" (p. 159). With a level of understatement that is fairly common in his writings, he further asserted, "As regards the fundamental theory [of two factors], I venture to maintain that this has now been demonstrated with finality . . . it becomes a bed of Procrustes, into which all our doctrines must somehow or other be made to fit" (p. 172).

Spearman's claim of total victory was a little premature. At the same time (the next article in the same journal), Thomson (1920) reviewed recent tests of the two–factor theory and found that several of them produced evidence of at least one additional factor, variously called "persistence of motives," "cleverness," and "purpose" by their authors. On further statistical grounds, Thomson rejected Spearman's claim for the distribution of the tetrad criterion, saying that his own finding "proves the invalidity of Professor Spearman's mathematical argument . . . [and] that [the] theory returns to the status of a possible, but unproven, theory" (p. 180).

Thomson was basically in the Thorndike theoretical camp. He suggested that intelligence was made up of a very large number of specific abilities and that any test merely sampled from this universe of abilities. Two tests correlated to the extent that they drew samples that overlapped in the tasks they required. Thomson's sampling theory predicted that group factors would appear and that *g* was possible but not essential. As evidence in favor of more specific abilities, Thomson drew attention to the research on transfer of training which, at that time, indicated that very little transfer took place between different kinds of learning. He concluded that this evidence showed that the abilities involved in the different kinds of learning were relatively discrete and independent.

Thorndike's definition of intelligence. As always, Thorndike was far from silent, although his writings covered a much wider range than the debate over the theory of two factors. In 1919 he severely criticized Binet–type tests because of their coachability. Noting that coaching seriously impaired the validity of intelligence tests, he advocated creating a very large number of alternate forms of tests and making test items public in order to equate knowledge about the content of the tests. He and his students were working on a set of intelligence tests that would meet these criteria and that would be used for admission to Columbia (Thorndike, 1920b).

While Spearman claimed that the nature of intelligence was now known and other writers—even Terman to a certain extent—were suggesting that the testing millennium was just around the corner, Thorndike saw the problem of defining and measuring intelligence as increasingly complex and the solution as receding farther into the distance. In a series of papers (1919, 1920a, 1920c) he proposed that the separate functions measured by different types of intellectual tasks could not be explained with a single construct: "The primary fact is that intelligence is not one thing but many. The abilities measured by a speed test with language and mathematics are not identical with, or even very similar to, those measured by a test with pictures and less exacting in speed" (1920c, p. 287). He argued that intelligence tests really measured only a limited aspect of intelligent behavior, which he labeled *abstract intelligence*. In addition, he said, intelligence comprises at least two other major kinds, *social intelligence*, the ability to understand and work successfully with people, and *mechanical intelligence*, the ability to understand and deal with concrete things. (While mechanical abilities have been studied in attempts to predict vocational success, relatively little attention has been paid until quite recently to social intelligence.)

In some ways, Thorndike's theory of intelligence was becoming more like Binet's, an integration of many aspects of the person. In response to a request for his definition of intelligence ("Intelligence," 1921), he stated:

> It is probably unwise to spend much time in attempts to separate off sharply certain qualities of man, as his intelligence, from such emotional and vocational qualities as his interest in mental activity, carefulness, determination . . , persistence . . ; or from his amount of knowledge; or from his moral or esthetic tastes. (p. 124)

He then responded to the point of the symposium: "We may define intellect in general as *the power of good responses from the point of view of truth or fact. . .* " (p. 124). His view of the essential unity of the personality in intellectual functioning reads almost as though it could have been written by Binet: "In human nature good traits go together. To him who hath a superior intellect is given also on the average a superior character. . . . There is no principle of compensation whereby a weak intellect is offset by a strong will" (1920a, pp. 233–234).

Thorndike's response. Turning his attention back to issues raised by Spearman, Thorndike (1921) performed the analyses prescribed by Spearman on three sets of army data from the war. As usual, the samples were large and relatively homogeneous. In one sample the set of variables included the combined tests of the Alpha and Beta examinations for 800 subjects. Different tests were analyzed in the other two samples. Thorndike concluded that there were separate dimensions of ability that were themselves correlated, such as "numbers as content" and "spatial relations as content," a position that foreshadowed Thurstone's results 15 years later. In addition, he took a direct shot at *g:*

> Everybody will agree that many complexities of individual differences are superadded by likenesses and differences in training. I fear, however, that even if we did dissect out all the consequences of nurture, leaving only a skeleton of inborn capacities, the organization of these would still be much more complex than that required by Spearman's theory. (1921, p. 151)

In reviewing Thorndike's results, which were based on much larger and more adequate samples than any he had ever used, Spearman (1922) pointed out that the selection of tests did not meet some of his restrictions and that when the offending tests were dropped, *g* appeared as expected. He then issued a challenge to Thorndike to agree to a set of conditions governing an experiment that would settle their ongoing differences. Although the ground rules were more or less agreed upon (Thorndike, 1924; Spearman, 1925), the study was never conducted.

In what he considered to be his most important work, Spearman presented a revision and expansion of his theory in 1923. Here he defined *g* as a fund of mental energy that a person could bring to a task. General intelligence, or *g*, reflected differences in people's abilities to apprehend experiences, educe relations among these experiences, and educe correlates. Apprehension of relations was essentially equivalent to encoding and remembering information; eduction of relations was inferential reasoning at the level of particular events; and eduction of correlates involved tasks in which the examinee was given an element and a relationship and was called upon to produce another element which had the specified relationship to the first.

The Spearman–Thorndike debate was never firmly resolved. Strasheim (1926) constructed a set of test exercises to measure the mental functions postulated by Spearman and found that such a test yielded evidence of a single dimension of intellect. This test and the CAVD, which Thorndike had designed to measure intelligence as he conceived it (see chapter 3), were the only two instruments specifically derived from the two competing theories. Each of the antagonists continued to believe in his own theory, but in the late twenties the controversy shifted to a higher mathematical level and Thorndike left others—notably Truman Kelley and L. L. Thurstone—to attack the two–factor theory with the theory of multiple factors.

Factor Analysis and the Multiple Factor Theory

Spearman's two–factor theory is generally credited with being the starting point for factor analysis as a logical method, although, as Harman (1976) pointed out, Pearson had published the method of principal axes in 1901, and Pearson's mathematics has been used in the solution of factor problems. The critical difference between the two men's contributions is that Spearman postulated that a psychological construct, a dimension of the mind, underlay the mathematical result. Pearson's procedure, on the other hand, was solely a data–reduction method.

The tetrad equation put Spearman's notion of a general intellective factor on a firm mathematical basis. Once this was accomplished, the two–factor theory could be attacked on mathematical grounds, and, of course, it was. Thomson (1916, 1920) pointed out various weaknesses, including that the criterion could only be relied upon when it was exactly satisfied and that in real data the tetrad equation yielded a distribution of values that frequently differed from that predicted by Spearman. Various authors, in particular Kelley (1928), found cases where the distribution of tetrad values indicated the presence of additional factors, thus using the theory's own criterion to attack it. In what amounts to

one of the first actual factor analyses, Kelley used partial correlation techniques to remove *g* from a set of data and then analyzed the residuals. He found evidence for several dimensions of individual differences, which were later verified by methods that Thurstone was just beginning to develop.

In the early 1930s two important advances were made by Thurstone (1931) and Hotelling (1933) that would profoundly affect the future of intelligence measurement. The less important was Hotelling's (1933) publication of the generalization of Pearson's method of principal components analysis. Pearson had shown how to find the single largest component in a set of data, and Hotelling developed the mathematics by which it became possible to decompose a set of variables into a set of uncorrelated factors. Hotelling's contribution was less important at the time than Thurstone's because it was not related to any psychological theory, although it has since replaced Thurstone's method as the procedure for analyzing data on modern computers. (For the purposes of this book, principal components and principal axes are considered as equivalent because they use the same method of factor extraction, which is what Hotelling discovered, and differ only in the diagonal elements of the correlation matrix that is analyzed. See Harman, 1976, or Gorsuch, 1983, for details.)

The introduction of multiple factor analysis. The revolutionary event was Thurstone's publication of the method of multiple factor analysis. In his paper Thurstone (1931) pointed out that the tetrad equation was a special case of a more general concept, that the number of dimensions required to account for the correlations among any set of tests could be defined mathematically by procedures of matrix algebra. His generalization of the concepts underlying the two–factor theory and the relatively simple methods of analysis that he developed to accompany it opened up a new line of attack on Spearman's work.

Thurstone's method of analysis came to be called the *centroid method* of factor analysis because it involved identifying successive "centers" of covariation in the data. Although it was a direct development of the line of research that Kelley had been taking, Thurstone's insight regarding the relationship between factors and matrices of correlations was a major reformulation of the problem which resulted in a much more comprehensible structure. A complete exposition of the method, including many concepts that are still in common use today, such as rotation of the factors, estimation of common variance, and selection of the number of factors to retain, was published in 1935 under the title *The Vectors of Mind.* A more generally available revision and expansion of this landmark work was published in 1947 as *Multiple Factor Analysis.*

Thurstone had long been active in the mental testing movement. He had served as an army psychologist during World War I and had written more than one intelligence test. His Cycle–Omnibus Test (1921) presented six different kinds of items in a spiral format in an effort to resolve one of the differences between the Stanford–Binet and the Yerkes–Bridges Point Scale. One item of each type was given at each level of difficulty. A subject attempted all six items at one level before going on to the next level. This ensured that each subject was tested equally in all areas. Thus, Thurstone's test bypassed the criticism against the Stanford–Binet that different functions were tested at different ages, while it retained the constant variety of items that Terman claimed was necessary to maintain interest in the tasks, particularly among young subjects.

Thurstone was an active and severe critic of the age–scale method employed with Binet–type instruments. He (Thurstone, 1926) pointed out that there were different conceptions of mental age and that the correct one stipulated that mental age was the expected test score for people of a given chronological age. He noted that "Binet may still be given credit for having introduced certain types of objectivity in mental measurement but his invention of the mental age concept was an awkward and unfortunate one" (p. 278). Throughout the twenties, Thurstone presented alternative scaling methods for mental measurement, culminating in his study of attitude measurement (Thurstone & Chave, 1929).

The Primary Mental Abilities. Having introduced multiple factor analysis, Thurstone set out to demonstrate its value and to propose a description of the domain of abilities. In what remains a monumental effort, he conducted a study of the Primary Mental Abilities (PMA), the first study of its kind, to reveal the structure of intelligence. Data were collected on 56 measures—by far the largest number attempted up to that time. The measures were administered to 240 college students who had volunteered. Each subject spent 15 hours during the summer of 1934 completing the battery, and the resulting scores were analyzed by centroid factor analysis. Sketchy results were published in 1936 (Thurstone, 1936a, 1936b), with the full study presented in 1938.

The PMA study could be considered one of psychology's wonders of the world, given the computational equipment then available. Ledyard Tucker was an assistant of Thurstone's at the time and was responsible for overseeing the computational activities required for the analysis. One day in 1972, while standing within sight of where the work was done at the University of Chicago, he described the process (Tucker, personal communication).

Thurstone's laboratory had, in addition to a matrix multiplying machine, a staff of about 20 WPA (Work Projects Administration) workers—this was the depression. With the work divided among these 20 people, each of whom was working 8 hours a day, and with the most modern equipment, the computations for the study took approximately 6 months. (In 1972, I had recently completed a reanalysis of the data, using 1970's equipment, in less than 6 seconds.) Thurstone expressed to Tucker the wish that each of his students could have had access to such a staff of computing assistants so that he or she could have conducted a factorial study of meaningful size as a doctoral dissertation.

One of the innovations that Thurstone presented in this study was the rotation of factors. He realized that once the decision as to the number of factors had been made, the placement of the factors within the space of the set of variables was mathematically arbitrary. Therefore, he advocated rotating the factors to positions where they would have the greatest psychological meaning. To accomplish this he proposed three criteria: psychological meaning itself (the results should make psychological sense); positive manifold (since the observed variables were all positively correlated, the factor coefficients should also be positive); and simple structure (each variable should be explainable in terms of a small number of contributing sources, i.e., the factors). Contrary to the impression given by Gould (1981), the first publication of the data (Thurstone, 1936b) reported a rotation in which the factors were allowed to be correlated. Thurstone (1952) attributed the fact that the complete results (Thurstone, 1938) reported an orthogonal rotation to a suggestion by Thorndike that the study's impact would be reduced if too many innovations were introduced in one paper. From the very beginning Thurstone was an advocate of oblique rotations in factor analysis. (For those not familiar with the logic and vocabulary of factor analysis, books by Gorsuch, 1983; Harman, 1976; or R. M. Thorndike, 1978, provide a useful introduction.)

The PMA factor analysis yielded 12 factors, seven of which were defined with sufficient clarity for Thurstone to name them. Thurstone (1936b) acknowledged that "most of the factors that we have identified have appeared in previous studies by Kelley, Spearman, and others, in the form of group factors. . . . The factors . . . are nearly uncorrelated . . . with one conspicuous exception . . . visualizing and number. This correlation is about 0.40" (p. 133). Because he believed that the multidimensional nature of intelligence as he had described it precluded the existence of *g*, he recommended that "each individual should be described in terms of a profile of mental abilities instead of a single index of intelligence" (p. 133). Norsworthy's 1906 recommendation that intel-

lectual abilities be presented as a profile was thus reappearing with a much more sophisticated methodology.

Like Spearman and others who had done factor–like studies, Thurstone believed that his analyses would lead to the discovery of mental structures, perhaps with physical analogs. For Spearman, *g* was real, a "thing" in the mind, and to a certain extent Thurstone seems to have subscribed to the same notion. Factors were seen as underlying causes of test performance rather than merely descriptions of dimensions of covariation. It was Thurstone's eventual hope to measure these variables directly.

As a preliminary approximation, Thurstone constructed a battery of tests that would function much like a single test but would provide relatively pure measures of each of the primary factors. Many of the tests were modified versions of ones in the original PMA study and were widely used as parts of other tests. This approach—the collection of several tests into a battery that would yield a set of scores measuring separate mental functions—has dominated many of the theoretical developments in group testing ever since. (It is interesting that although the multifactor batteries have received a great deal of theoretical attention, most test practitioners have continued to use individual and group tests that yield one score, or at most a small number of scores for interpretation.)

Spearman's reaction to multiple factor analysis. Of course, Spearman (1939) did not agree entirely with Thurstone's analysis. In addition to questioning a number of technical points, he argued that Thurstone had extracted too many factors. To prove his point, he reanalyzed the data after arranging the variables in ten groups and found that he obtained *g* as the major factor, along with small group factors for verbal, spatial, number, and memory abilities. This was consistent with his previous results (as they had been gradually modified to account for successive problems) and he claimed victory. The dimensions of disagreement between Spearman and Thurstone set the agenda for a continuing controversy, primarily between British and American factorists, which formed much of the core of the intelligence debate until the 1970s.

For several decades the argument over the nature of intelligence centered on how to perform factor analyses. The way that a factor analysis is conducted and the number and type of variables selected for inclusion largely determine what the analysis reveals. It is not possible to extract from a factor analysis something that is not there in the first place. For various mathematical reasons, it is necessary for at least three variables that measure a factor to be included in the analysis in order to definitely confirm the presence of the factor. Spearman guaranteed by his grouping of the variables that the kinds of multiple factors Thurstone found in the PMA data would not surface in his reanalyses but that *g* would. The reason is

simple and applies to all factor analyses: The mathematical procedure arbitrarily selects the largest dimension of covariation first. Since ability variables are almost always positively correlated, the largest dimension is always one that shows moderate positive relationships with all or most of the variables, hence *g*. Without rotation, almost any factor analysis of ability variables will produce a general factor of the kind that Spearman's theory predicted, but this is particularly likely when the set of variables is small and heterogeneous.

Group factors occur when there are variables that have correlations larger than the general factor can explain. This happens when there is more than one variable of a particular kind in the analysis. For example, if we conduct an analysis of three variables, one test of verbal ability, one of mathematical ability, and one of rote memory, we will get a single factor that is positively related to each of the tests and two small factors that may contrast different aspects of the performance, but it will not be possible to find factors that represent verbal ability, mathematical ability, and memory respectively because *g* is the only factor represented by three tests. However, had we included three tests of each type, we would have gotten the same *g* as the first factor in our set of nine variables. This factor would have been fairly large, but it is quite likely that there would have been at least two other factors that would also be of modest size. By rotation we could find one factor representing the cluster of three verbal tests, a second factor representing the mathematical tests, and a third factor for memory. The rotation would combine *g* with the other factors and relocate them to reflect the kinds of tests that had been included in the battery.

The consequences of factor theory. One important but often overlooked by–product of Thurstone's method of rotation was that it changed the type of test that was included in subsequent investigations. Thurstone's factors tended to be defined by homogeneous, highly speeded tests composed of relatively simple tasks. Complex tests that emphasized the same content as one of his primaries would load on the primary, but would also contain other factors. Thus, tests selected as markers for particular abilities in later studies were increasingly speeded and specific. The criteria for factor rotation, primarily simple structure, came to define a new universe of tasks different from the altitude and power tasks of earlier years. The fractionalization of abilities that followed would not have happened had this shift in the tasks used to define intelligence not also occurred.

During the 1940s, 1950s, and 1960s factor analysis served primarily as a descriptive procedure. Spearmanites could always find *g* by including only a few variables, as they usually did, and by refusing to rotate the factors. Thurstonians could always find multiple factors by including several variables

of each type in the analysis and by rotating the solution to simple structure. Because each side played the game by different rules, neither could win and, in a sense, both were right. Each method provided a useful description of the data set that an author might choose to analyze, but neither side could produce their desired result from the other's data because of the fundamental disagreement over the appropriate nature of the data. (This same fact had emerged, but less clearly, in the Spearman–Thorndike debate that was described previously.)

Thurstone actually provided the seeds of a solution to the problem with the idea of rotation. As noted earlier, he believed from the start that the factors of mental ability had modest positive correlations with each other. A rotation to simple structure resulted in factors that were correlated, and these factors could in turn be factor analyzed. When a second–order factor analysis (a factor analysis of the matrix of correlations among the primary factors) was performed, a general ability factor similar to *g* emerged as the reason that the primary factors were correlated. While neither Spearman nor Thurstone was particularly happy with this resolution, by the 1950s the debate about *g* versus multiple factors had come to be viewed as a methodological issue. Certainly, the debate had not been completely resolved, and continues to the present day, but some of the urgency was gone. Attention came to be focused more on the configuration of the factorial solution—the relationships among the factors—rather than on the suitability of a single factor to explain all of intelligence.

NEW AND REVISED TESTS

The Stanford–Binet Revision

Terman remained aloof from the debate over the factor structure of intelligence. He believed that his methods provided the best available way to measure general intelligence and that the problem was to refine these methods. (If forced to take sides, he would definitely have come down on Spearman's side of the controversy.) While Thorndike, Thurstone, and Spearman argued, Terman devoted much of his effort to further development of the Stanford–Binet. He ("Intelligence," 1921) had already started work on a second edition.

By 1921, only five years after its introduction, many users had pointed out one or another shortcoming of the Stanford revision of the Binet–Simon scale. The objects of criticism ranged from the overly verbal and academically oriented character of some of the items to the age placement of particular tests. Critics claimed that items were too easy at the bottom and too hard at the top. In addition, there was insufficient range at the top to assess the upper

levels of adolescent and adult performance. A particularly important criticism concerned the nature of the standardization sample. About 1,000 white American–born children from California and a small, unrepresentative sample of older individuals had been included in the standardization. Possible bias in this group was one of the points of attack Lippmann selected for his objections to using the Stanford–Binet in interpreting the army data.

Although the age–scale concept had been under increasingly vigorous attack for many years, Terman was completely committed to it and retained it in his 1937 revision despite the logical and statistical arguments against an age scale. In fact, as had been pointed out repeatedly (see Otis, 1917; Thurstone in "Intelligence," 1921), there was no necessary difference between an age scale and a point scale in terms of the type of items used. Terman himself acknowledged this point. And he was well aware that the age placement of items created problems absent in a point scale, which ordered items by level of difficulty. Probably the real reason Terman insisted on the age scale was his belief in the value of the ratio IQ. In presenting the new form and arguing for retaining the age scale, he claimed that educators and other users were accustomed to the MAs and IQs and that it would take 20 years for them to adjust to a new metric. He also asserted that his ratio IQs were almost standard scores because the standard deviations at most ages were nearly equal to 17. In fact, the standard deviations ranged from 12.5 to 20.7, but Terman held firmly to his position.

For 15 years Terman and his colleagues at Stanford, in particular Maud Merrill, worked with varying degrees of intensity on the revision. The greatest amount of work was carried out in the early 1930s. In 1937 Terman and Merrill published their revised scale in two forms (Forms *L* for Lewis and *M* for Maud).

A trial set of 408 items (209 for Form *L* and 199 for Form *M*) was ready for testing by late 1930. Sensitive to the criticisms about his 1916 standardization sample, Terman made a serious attempt to draw a sample that was representative of the national population in 1930–32: "Seventeen different communities in eleven states [California, Nevada, New York, Colorado, Kansas, Virginia, Vermont, Texas, Minnesota, Indiana, and Kentucky] were sampled to secure 3,184 subjects upon whom [the] final standardization was based" (Terman & Merrill, 1937, p. 12). A team of seven examiners, all carefully trained by Stanford University personnel, did the testing. An attempt was made, both in drawing the original sample and in weighting the final results, to control for socioeconomic level, but all subjects were native–born white Americans. From ages 1.5 to 5.5 years, approximately 50 girls and 50 boys at each 6–month interval were tested; from ages 6 to 14, 100 boys and

100 girls at each year of age; and from ages 15 to 18, approximately 50 of each sex at each year.

Several item validity criteria were used in the selection of items for the final scale. "The first principle of sifting was to give preference, other things equal, to types of items that experience had shown to yield high correlations with acceptable criteria of intelligence" (p. 7). Then, "curves of percents passing at successive mental ages were plotted and the steepness of these curves afforded a graphic indication of the validity of the tests . . . each test was given a provisional location at the level where the proportion of passes was approximately 50 percent" (pp. 8–9). Next, the correlation of each item with total score on the test was computed. The final item selection criteria "in order of importance . . . were: (1) validity, (2) ease and objectivity of scoring, and (3) various practical considerations such as time economy, interest to the subject, the need for variety, etc." (p. 9).

The standardization process was a long one:

> The "correct" standardization of an age scale depends, of course, upon the age location of the separate tests and upon the amount of credit (months of mental age) allowed for passing them. . . . It is at present not possible to lay down, in advance, rules which if followed will cause the scale to yield a mean I.Q. of 100 at each level. In the present revision, as in the original Stanford–Binet, it has been necessary to work empirically by revising and re–revising until an arrangement of the tests was formed which achieved the desired goal. (pp. 22–23)

The end result was a major improvement over its predecessor. Where the 1916 scale had been composed of 90 items in one form, the revision had 129 items in each form. The range of the scale had been increased by adding some items and changing the age placement of others at both ends. It could be used with individuals whose mental age was as low as two years, and at the top there were now three levels of superior adult, yielding MAs as high as 22 years–10 months. The higher MAs, of course, represented an extrapolation of the scale upward for the very bright teenagers in the standardization sample.

Although data had been accumulating suggesting that intellectual development continued well beyond the age of 16 (e. g., Thorndike, 1923), Terman continued to use 16 as the highest chronological age for the purpose of determining IQ. However, because the decline in the rate of increase in MA was gradual rather than abrupt (the developmental curve seemed to start leveling off at about age 13), Terman introduced a sliding CA for the denominator of the ratio. "From thirteen to sixteen we cumulatively drop one out of every three additional months of chronological age and all of it after sixteen" (Ter-

man & Merrill, 1937, p. 30). This had the effect of making 15 the highest CA used in the denominator of the IQ ratio. The norms tables provided for IQs up to about 170 for most ages, but for age 14 and above the maximum value gradually declined, dropping to 152 for those 16 or older. At the other end, a child had to be 6.5 years old before the minimum value of 30 could be assigned because the lowest MA was two years.

Terman was still a strong believer in *g*, and he claimed that his revised test measured general intelligence. Quinn McNemar, one of his colleagues at Stanford, took the standardization data and performed an extensive statistical analysis (McNemar, 1942) to substantiate Terman's claim. In addition to preparing distributions of scores subdivided into various groupings of the subjects, McNemar performed factor analyses of the items at several age levels. The factor analyses, which were performed more according to Spearman's rules than Thurstone's, did in fact confirm the presence of a single general factor, with the possible presence of a small group factor contrasting memory and verbal ability occasionally in evidence. McNemar was able to show that the items of the new Stanford–Binet permitted description by a single factor, but as Jones (1949) later demonstrated, the data could also support a multifactor interpretation.

Development of the Wechsler–Bellevue

Terman was correct when he wrote in 1937 (Terman & Merrill, 1937) that Binet–type tests in general, and the Stanford–Binet in particular, had no serious rivals among measures of intelligence. There were many group tests available for both children and adults as a result of testing programs developed during World War I and later for the schools. In 1918 the Thorndike Test of Mental Alertness, replaced in 1924 by the Intelligence Examination for High School Graduates, was being used for college admissions decisions. The first Scholastic Aptitude Test for college admissions was produced for the College Entrance Examination Board by Brigham in 1926, and a wide variety of tests of various aptitudes was available by 1935. Virtually all of these tests made at least part of their claim to validity by showing high positive correlations with the Stanford–Binet.

Tests needed for adults. However, the Stanford–Binet did not satisfy every clinician's needs. In particular, David Wechsler, who, among his other duties, was staff psychologist at New York's Bellevue Hospital, felt the need for an individually administered test that could be used with adults: "The continued use of children's scales for adults has no scientific justification. The scales now in use fail to meet some of the most elementary

requirements" for standardization (Wechsler, 1939, p. 15). He enumerated four major shortcomings of the individual mental tests then available:

1. The lack of standardization on adult groups made the norms inappropriate.

2. The material used in the tests was not appropriate for adults. Tasks that interested children seemed silly to adults.

3. There was too much emphasis on speed of performance. Adults could not see a reason to work fast just because an examiner said they should, and speed seemed to be one of the first functions to deteriorate with age.

4. "The concept of mental age, fundamental as it is to the definition of juvenile intelligence, may be grossly misleading when applied to the definition of adult mental capacity." (p. 18)

Wechsler had tested adults during his service as an army psychologist during World War I. After completing his doctorate with Woodworth at Columbia in 1925, he worked in various psychological capacities in the New York area, including Cattell's test publishing company, The Psychological Corporation, where he participated in the development of automobile driving simulators and driving tests. By this time, he may already have begun his plans for an intelligence test for adults (Kuder, 1986).

Interest in intelligence measurement for older individuals had been growing for years. As Weisenburg, Roe, and McBride (1936) pointed out, there had been a number of studies using college students in the early years (such as those by Sharp and Wissler) but the army testing program really brought the problem of measuring adult intelligence to the fore. As the difficulties of measuring child and adolescent intelligence gradually were solved, the issue of defining the sequence of intellectual development in the adult years came up for examination. Research had generally indicated that mental ability continued to increase until about age 20 (Terman's beliefs notwithstanding). Weisenburg et al. (1936) concluded that "the greatest development in 'test intelligence' occurs before the age of twenty, and . . . the gains or losses which appear from that age to sixty are comparatively very slight" (p. 109). Others claimed a rate of change ranging from fairly rapid decline to actual continued growth, depending on the type of test used and the function tested (e.g., see Garrett, Bryan, & Perl, 1935; Lorge, 1936).

The deviation IQ. With vigor and insight, Wechsler attacked the use of mental age for adults (although most of his points had previously been made by others whom he failed to cite). He pointed out that chronological age was actually an *expected* mental age; that is, the mental age that people who had lived for

a given number of years should, on the average, attain. Therefore, the IQ was a ratio of observed score to expected score—both of which were expressed in years—and was a measure of relative performance, as were standard scores and percentile ranks—both of which had been advocated by others for years. "The essence of the IQ concept is that part of its definition which asserts that for a valid evaluation of an individual's brightness, one must compare his mental ability to that of the average individual of his own age" (Wechsler, 1939, p. 31). Therefore, any IQs based on extrapolation of results from children or adolescents to adults, such as Terman proposed for the 1937 Stanford–Binet, were invalid and did not constitute IQs at all:

> To speak of an individual having an M.A. of 20 years is both practically and scientifically meaningless. . . . The M.A. method of defining intelligence cannot logically be used to define levels of intelligence higher than that obtained by that age group beyond which M.A. scores cease to increase with chronological age. (p. 23)

It was with his 1939 book that Wechsler presented what has become the standard way of expressing the relative performance of individuals on measures of mental ability, the *deviation IQ*. While retaining the term *Intelligence Quotient* to refer to the concept, he completely changed the way it was computed, making it instead a transformed standard score, as Otis had suggested in 1917. He set the mean equal to 100 for continuity with conventional practice, but he set the *probable error* (a statistic approximately equal to two–thirds of the standard deviation, but defined so that 50 percent of the distribution is contained within the interval of the mean plus or minus 1 PE) equal to 10. This was done so the middle 50 percent of subjects would earn IQs in the range of 90 to 110. The values yielded by this system were similar to those given by the ratio IQs from the Stanford–Binet, but the standard deviation tended to be slightly smaller. More importantly, Wechsler's method resulted in a scale where a particular IQ value retained the same meaning in terms of relative position in the norm group, regardless of age level. Differences in standard deviations from one age to another, with the resulting change in the meaning of the ratio IQ, had been one of the major criticisms of the Stanford–Binet, even in its revised form. The deviation IQ answered this criticism.

The Wechsler–Bellevue. New measures were needed to explore the later stages of development because the tests designed for children, such as the various revisions of the Binet–Simon scales, contained items that were not suitable for use with adults—even if the metric of mental age was avoided. Wechsler (1939) provided the following picture of the procedure by which items for his scale for adults were selected:

> (1) A careful analysis was made of the various standardized tests of intelligence now in use. These were studied with special attention to authors' comments with reference to the type of functions measured, the character of the population on which the scales were originally standardized, and the evidence of the test's reliability. (2) An attempt was made to evaluate each test's claim to validity as evidenced by its degree of correlation (a) with other recognized tests and, (b) more important still, with subjective ratings of intelligence. The last named included teachers' estimates, ratings by army officers. . . and estimates of business executives.... (3) An attempt was made to rate the tests on the basis both of our own clinical experience and of that of others. (4) Some two years were devoted to the preliminary experimental work of trying out various likely tests on several groups of known intelligence level. (p. 78)

As a result of this process, 11 tests were identified for inclusion. Five tests—Information, Comprehension, Memory Span for Digits, Similarities, and Arithmetical Reasoning (with Vocabulary as an alternate)—made up the Verbal Scale; Picture Arrangement, Picture Completion, Block Design, Object Assembly, and Digit Symbol composed the Performance Scale. The Verbal Scale and the Performance Scale were combined to yield a Full Scale score.

The standardization of the Wechsler–Bellevue scale was accomplished in a manner very different from that employed by Terman and Merrill. Wechsler had been using most of the tests for several years in his clinical practice and in his work at Bellevue Hospital. By June 1938, he had accumulated 3,499 cases: 1,639 children between ages 6 and 16, and 1,860 adults ranging in age from 17 to 70. The norming samples—670 children (ages 7 to 16) and 1,081 adults (ages 17 to 70)—were selected from these two groups. An attempt was made to have the samples reflect the educational and employment characteristics of the American population as revealed in the 1930 census. However, because almost all of the cases were from the metropolitan New York area, it was necessary to do some creative substitution to represent the agricultural classification—barbers, bakers, and teamsters were used.

The individual tests were each separately standardized and scaled to have a mean of 10 and a standard deviation of 3. Most tests had a maximum possible score of about 20, and a table of equivalents was provided to transform the raw scores into "equivalent weighted scores" that ranged from 0 to 17. The five weighted scores for the Performance tests were summed, and the total used to enter a table of Performance Scale IQs. The same procedure was used with the other five tests to find the Verbal Scale IQ. The total of all ten weighted scores provided the table entry to obtain the Full Scale IQ that was essentially equiva-

lent to a Stanford–Binet IQ. Two sets of norms tables were provided, one for use with subjects between 10 and 16 years; the other, with individuals from 16 to 60.

The Wechsler–Bellevue was designed to measure adult intelligence as Wechsler (1939) conceived it. It was here that he produced his widely quoted definition: "*Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment*" (p. 3). But despite the apparent clarity of this pronouncement, Wechsler was a little fuzzy about what he meant. He stated that "Professor Spearman's generalized proof of the two–factor theory of human abilities constitutes one of the great discoveries of psychology," that "the combining of a variety of tests into a single measure of intelligence, *ipso facto*, presupposes a certain functional unity between them" and that " '*g*' is a psychomathematical quantity which measures the mind's capacity to do work" (pp. 6–8). Yet, he concluded that, "intelligence tests measure more than mere learning ability or reasoning ability or even general intellectual ability; in addition, they inevitably measure a number of other capacities which cannot be defined as either purely cognitive or intellective" (p. 11). The last statement shows the clear influence of E. L. Thorndike's vision of a multifaceted intelligence, which is hardly surprising in light of Wechsler's Columbia doctorate.

Wechsler's theory of intelligence was also profoundly influenced by a study by Alexander (1935) in which the author found *g*, several group intellective factors such as verbal and practical abilities, and two non–ability factors. Alexander concluded that these last two factors, which he labeled *X* and *Z*, affected all tests. They were interpreted to represent persistence, interest in the test tasks, desire to succeed, and other personality dimensions that were not directly represented by manifest variables in the analysis. Wechsler (1939) asserted that a test could not be a measure of general intelligence without measuring these nonintellective factors:

> That is why, for example, the Stanford–Binet scale becomes less and less a successful measure of intelligence at the upper levels where it is overburdened with items such as vocabulary and abstract definitions, items which even when loaded with "*g*" and "*v*" are poor in the "*X*" and "*Z*" factors. (p. 11)

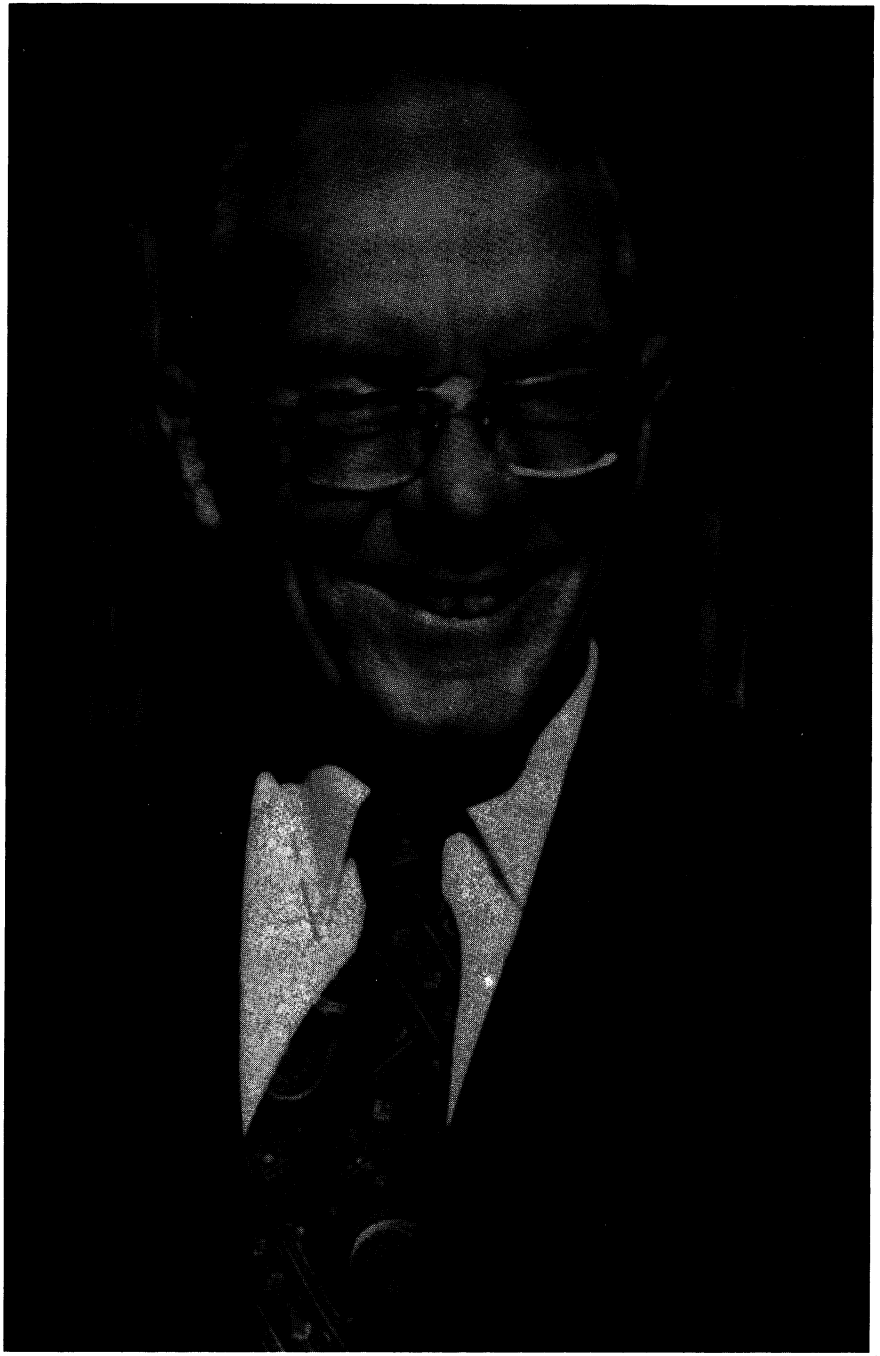
In his test these non–ability factors would be measured by the Performance Scale, adding this vital element to the Full Scale IQ as an estimate of general intelligence.

A number of the features of Wechsler's first scale, and also of the series of scales that he later developed, are strikingly similar to the Yerkes–Bridges Point Scale of 1915, although Wechsler gave no indication that he was aware of any of

Yerkes' work. (Indeed, it wasn't until Matarazzo's 1972 revision of Wechsler's book that Yerkes' name was even mentioned, and then only in connection with the army testing program.) The tasks were grouped by type so that each subject attempted only comprehension items in one set, only digit span items in another set, and so on through all ten tests. Although Wechsler did not address the issue, it is likely that this aspect of the point scale works better with adults than with young children. The effect of variation in tasks on the interest and motivation of examinees was given as one argument for retaining the item heterogeneity of the age scale in the revised Stanford–Binet, but this would probably have had a smaller effect with adults and might have even been disconcerting to them.

A second similarity between the Yerkes–Bridges scale and Wechsler's was the assignment of variable credit to items depending on the quality of the response. Both Binet and Terman had argued that items should be scored as correct or incorrect, with no further differentiation as to the quality of response. On some items Yerkes had awarded 0, 1, or 2 points depending on the completeness or maturity of the response, and Wechsler adopted this practice.

The advent of the Wechsler–Bellevue rounded out the collection of tests available to measure human intellectual abilities. The Binet–type scales for individual assessment of children had led the way, followed by the short–lived Point Scale. (Yerkes had also suggested and done some work on separate point scales for children, adolescents, and adults, a task that Wechsler would complete many years later.) Group tests for both children and adults had appeared in response to the needs of the educational community and the military. In 1939, the Wechsler–Bellevue was introduced, providing an individually administered clinical instrument for use with adults. While other tests have been introduced and the old standbys have been revised periodically, there have been relatively few important innovations in ability measurement introduced in the years since 1939.



Robert L. Thorndike (1910–)

5

Testing in the Second Half-Century

THE RATE OF PROGRESS SLOWS

The period from 1918 to 1939 had been one of dramatic change in the testing movement: The forerunners of most of the types of tests and inventories in use at the present time had been developed. Psychological testing had spread throughout the school system and had been applied in industry and civil service as well as in psychological clinics and research. Testing as a method had survived the strong (and often justified) criticism that resulted from occasionally wildly extravagant claims and absurd interpretations of test results. An interesting perspective on the period was provided by Oscar K. Buros (1977) as he reflected on a 50–year career in testing:

> In many ways, the year 1927 was a banner year in testing. I like to think of it as the approximate year in which the testing movement reached maturity. The unreasonably high expectations of earlier years were being replaced by more modest expectations of the usefulness of tests. The limitations were being widely recognized. (p. 9)

A somewhat more humbling comment is Buros's subsequent observation that "except for the tremendous advances in electronic scoring, analysis, and reporting of test results, we don't have a great deal to show for fifty years of work" (p. 10) since then.

Buros was essentially correct, but the date could probably be moved ahead a few years to about 1940. The first fifty years of the testing movement, the period up to about 1940, produced a staggering increase in the number and variety of methods available to psychology for observing human behavior. There have been few real additions in the second 50 years, and most of those have been quite recent.

The *Mental Measurements Yearbooks*

In 1925, G. M. Ruch called for the inclusion of information about test validity, reliability, administration, scoring, and adequacy of norms in every published test manual: "The test buyer is surely entitled to the same protection as the buyer of food products, namely, *the true ingredients printed on the outside of each package*" (p. 358). Buros himself, responding to Ruch's call, was responsible during the late 1930s for an innovation that would affect testing in many ways, the *Mental Measurements Yearbooks.* This series, which currently (1989) comprises eleven volumes (in nine editions) and publishes critical reviews of tests and a bibliography of most references that mention tests, began in 1935 as a modest 44–page bibliography of the 250 tests published in the years 1933 and 1934. Neither the 1935 pamphlet nor its 83–page successor (503 tests) published in 1936 contained any reviews of tests. By 1937 Buros had produced a 141–page bibliography that included reprints of journal reviews of most of the recent books on tests and testing.

The 1938 *Mental Measurements Yearbook* (*MMY*), which was the first volume to contain a collection of commissioned reviews of the tests themselves and really marked the inception of the modern series, was positively received by the professional journals of the day, but, as one might guess, the reaction from test authors and publishers was mixed, depending largely on how favorably their tests had been reviewed. Buros's description of his mail and his excerpts therefrom make interesting reading (Buros, 1941).

Buros's objective in founding the *MMY* was to provide an impetus for psychology to improve its tests and testing practices by exposing those instruments that did not meet acceptable standards for construction, standardization, and support, and by rewarding through positive publicity those that were particularly effective or well done. His plan required the active cooperation of the affected professionals, and, in general, it has been a success. The 1940 *MMY* included contributions from 250 individual reviewers. There were 55 different tests listed in the section on intelligence measures, although not all of them were reviewed. For example, the 1937 Stanford–Binet was not reviewed in this volume because it had been covered in the 1938 issue. However, the entry for it

contained references to 134 articles relating to Stanford revisions of the Binet scales, including Terman and Childs's 1912 seminal paper. There were also edited excerpts from journal reviews of the revised Stanford–Binet by Burt (*Eugenics Review*), Kent (*Psychological Record*), and Krugman (*Journal of Educational Psychology*). The 1940 (or Second) *MMY* thus began the pattern that has characterized the series, namely, that a test would generally not be reviewed a second time unless there were some significant change that called for particular attention. In the fifty years since its inception, the *MMY* has come to play the role of conscience of the testing profession, although Buros's hope that it would foster substantial improvements in testing practice has probably not been realized. Since his death in 1978, the effort has been carried on by the Buros Institute for Mental Measurements at the University of Nebraska.

Reviews of the revised Stanford–Binet were generally very positive, although Krugman (1939), for example, after noting that the revision was a major improvement over the 1916 version, still listed 12 particular reservations. To illustrate that it is impossible to remove cultural context from a test of intelligence, no matter how homogeneous the group, it is interesting to note that two of Krugman's objections dealt with material (relating to farm animals) that he felt was not appropriate for his clients, namely children from New York City. Correcting what he saw as faults would have made the tests equally inappropriate for other groups.

The new Wechsler–Bellevue scale was immediately hailed as the answer to some clinical problems. Writing in the *MMY*, Wells (1940) stated, "This series is by a considerable margin the best available procedure for adults, in a clinical setting" (p. 264). Kent (1940), in reviewing Wechsler's *Measurement of Adult Intelligence*, compared the scale favorably with Binet–type scales such as the Stanford–Binet and the Kuhlmann and Herring revisions:

> Clinical examiners who are relatively satisfied with some form of the Binet–Simon scale for examining children will welcome a test which is intrinsically better adapted to adult interests and which is adequately standardized for adults. To this extent, the Bellevue scale will meet a need that has been keenly felt for more than twenty years. (p. 253)

While noting a strong preference for the deviation form of IQ that Wechsler had introduced, Kent further complained, "It does not, however, show any such advance over the Binet scale as might be expected as a result of our collective experience in the use of tests for a full generation" (p. 253).

Military Testing in World War II

The beginning of mental testing's second half–century also saw the start of another major military conflict. The American psychological community was much better prepared to assist in the war effort in 1941 than it had been in 1917. Many useful testing procedures had been developed, and a fairly large number of psychologists had been trained in methods of test construction and analysis. The list of those who participated during this war reads like a Who's Who of psychometrics for the next 40 years, as the young psychologists who served in the testing programs continued their professional careers.

Psychological testing played a major role in the U.S. war effort in World War II. Each branch of the service—the army, the army air forces (AAF), and the navy—had a testing program to identify those who were or were not fit for military duty. The ones found to be fit for service were then classified by the specialities where their talents would be of greatest value to the war effort. During the course of the war more than 9 million men were given one battery of tests or another.

The military testing programs of World War II saw the first widespread application of batteries of tests designed to assess different functions—an innovation introduced by Thurstone with the publication of the Primary Mental Abilities tests. Initially the army testing program used an overall screening device called the Army General Classification Test (AGCT) that was similar to the Army Alpha of World War I. Eventually this test evolved into a battery of tests yielding four partial scores (arithmetic computation, arithmetic reasoning, reading and vocabulary, and spatial relations) as well as a total score (DuBois, 1970). The navy and the army air forces had similar testing programs, but each service focused on the type of testing that met its particular needs. There is no way to estimate the value of these programs in terms of dollars or lives saved, but some indication of their effectiveness can be gained from two examples: one from the army air forces described by DuBois (1970) and R. L. Thorndike (1949), and the other from the navy by Frederiksen (1984). Jensen (1982a) has estimated that the successors of these World War II batteries currently save the modern military over $400 million per year through improved training efficiency.

To determine how much the test battery actually improved the selection of potential pilots, the army air forces sent more than 1,000 cadets who had taken the usual pilot selection battery into training without applying the ordinary selection cutoff. In the entire unselected group, the correlation between the selection index (a weighted composite of several tests) and a pass/fail criterion in initial training was 0.64. In the subgroup of candidates whose test performance exceeded the cutoff score for admission (that is, those who

would ordinarily have been admitted and upon whom the index of validity would ordinarily have been determined) the correlation was only 0.18. Use of the test battery to select those who would be admitted to training dramatically increased the success rate among candidates, thus saving time, money, and probably lives in the war effort.

The AAF study offered insight into the effect of talent range restriction on the correlation of a test with a criterion, which has always been a problem for psychological testing. That is, when tests are used in selection of applicants for some position, and if the test has any validity at all, the range of criterion performance of those selected will be less than the range found in the total applicant pool because the lowest scoring individuals on the test will have been eliminated by the selection procedure itself. This will necessarily reduce the computed validity of the selection procedure because its value is to some extent related to variability on the predictor and criterion. Thus, the validity coefficient of the AAF pilot selection battery was known to be an underestimate of the actual validity of the battery for selecting potentially successful aviation cadets. The study showed the extent of that bias.

The navy benefited from psychological testing and the insights it provided in a different way (Frederiksen, 1984). Tests were used both as personnel selection tools and as criterion measures in the naval gunnery training program. The tasks performed by naval gunners involved largely psychomotor skills such as those required to disassemble and repair a naval gun. The tests that should predict this skill should involve psychomotor rather than verbal tasks, but the psychologists noted that the tests that correlated most highly with success in training were verbal tests. The criteria that had been used to appraise success in the program were verbal tests of knowledge about guns and gunnery. However, those trainees who could answer questions about repairing guns often could not fix them. When the naval examiners changed the criterion to involve primarily performance of the physical tasks, the most valid predictors soon became the psychomotor and spatial relations tests. Frederiksen noted that changing the criterion also resulted in improved methods of instruction.

Focus on Aptitude Testing Batteries

After the war there was a period of relative quiet in mental testing that lasted about 15 years. During this period, the major advances in intelligence measurement focused on the development of test batteries for use in counseling and placement. In 1947, the U.S. Employment Service released the General Aptitude Test Battery (GATB), a multifactor test of occupational

abilities (Dvorak, 1947). The GATB reflected Thurstone's work on the Primary Mental Abilities test battery and an extensive program of wartime research that had included government-sponsored studies on the most effective use of the civilian workforce, conducted for the War Manpower Commission. Over the years, the GATB has been the subject of continuing development and research in an ongoing attempt to predict occupational success and satisfaction. There are now validity studies involving more than 2,000 specific occupations.

Two other test batteries that appeared during this period for use in career counseling in secondary schools were the Differential Aptitude Tests or DAT, published by The Psychological Corporation in 1947, and the Flanagan Aptitude Classification Tests (FACT), published by Science Research Associates in 1953. Both of these batteries were intended to provide a set of relatively independent scores that could be used to identify an individual's most suitable course of study or occupation. The DAT has been revised several times and is still available as one of about a dozen such instruments, but the FACT is no longer available, probably due in part to its length and complexity.

The multifactor aptitude batteries that came out of World War II and that were highly influential during the 1950s appear to have reached a dead end. Over the years a number of test batteries have been offered. A few, like the GATB and the DAT, continue to have an important place in psychological testing, but many others have followed the FACT out of production. The reason for this is probably twofold. First, the needs of schools for this type of information have largely been filled by achievement tests and by scholastic aptitude tests that give scores for broad areas such as verbal and quantitative abilities. These test scores seem more direct and easier for general test users to interpret, and they avoid some of the implications of inherited traits that have plagued aptitude/intelligence tests. The second reason seems to be that the degree of detail provided by aptitude batteries is not useful for making predictions and decisions. There is growing evidence (see Jensen, 1982b; Schmidt & Hunter, 1981; R. L. Thorndike, 1985) that the information provided by such batteries is too specific to be useful for practical purposes in today's civilian world. This line of research suggests that the first unrotated factor from a test battery carries most of the useful predictive information about future performance, and this factor is very similar to a measure of general intelligence. In other words, Spearman's g is alive and well and present in the GATB, the DAT, and other such batteries, but it is hidden when multifactor scores are used.

INTELLIGENCE TESTS UNDERGO REVISION

Revisions of the Stanford–Binet

Neither the Stanford–Binet nor the Wechsler–Bellevue played a major role in or was affected by the psychometrics of World War II. An Army Wechsler–Bellevue was prepared, but by this time group tests had been developed to the point that individual tests were used only in very special circumstances.

In fact, the 1937 Stanford–Binet remained essentially unchanged for almost 50 years, even though a third edition, Form L–M, was published in 1960 and renormed in 1972. This 1960 version of the Stanford–Binet used the same items and the same parameters for the norms group that had provided the mental ages for the 1937 edition. Some item placements were changed, and the directions and scoring were clarified on the basis of intervening experience, but the organization and most of the content remained unaltered.

Form L–M combined the best 142 items from Forms L and M of the 1937 revision. This third edition had six regular items at every level except adult (where there were eight) and an alternate item for each level, to be used if one of the regular items seemed inappropriate for a particular examinee, or if some event interfered with its proper administration (see Terman & Merrill, 1960). The items for the revision were selected on the basis of results with 4,498 examinees who had been tested under various conditions during the period 1950–54, but except for correcting minor deficiencies, no items were significantly changed, and no new items were added:

> Changes in difficulty of subtests were determined by comparing percents passing the individual tests in the 1950's with the percents passing in the 1930's constituting the original standardization group. Criteria for selection of test items were: (1) increase in percent passing with age (or mental age); and (2) validity determined by biserial correlation of test with total score. Changes consisted in the elimination or relocation of tests which have been found to have changed significantly in difficulty since the original standardization; the elimination or substitution of tests which are no longer suitable by reason of cultural changes; further clarification of ambiguities of scoring principles and test administration; and the correction of structural inadequacies of the 1937 scale, first by introducing adjustments to make the average mental age that the scale gives more nearly equal to the average chronological age at each age level and second, by providing revised and extended IQ tables that incorporate

> built–in adjustments for atypical variability of IQs at certain age levels so that the standard score IQs provided are comparable at all age levels. (pp. 39–40)

Perhaps the biggest single revision in the 1960 Stanford–Binet was the method by which IQs were expressed. This edition adopted the deviation IQ, introduced by Wechsler in 1939, to describe relative performance. The scoring of the test still yielded a mental age that was based on the performance of the 1937 norm group, but the ratio IQ was gone, replaced by a transformed standard score such as Woodworth and Otis had suggested for Binet–type scales some 40 years earlier.

By 1970 it had become clear that the norms for the Stanford–Binet were obsolete. Terman had died in 1956 and Merrill had been retired for many years, so the publisher, Houghton Mifflin Company, selected R. L. Thorndike to head up the renorming study. Thorndike was the co–author (first with Irving Lorge and later with Elizabeth Hagen) of the Lorge–Thorndike Intelligence Tests and (again with Hagen) the recently released Cognitive Abilities Test, published by the same company. Houghton Mifflin had just completed the standardization of the Cognitive Abilities Test (then known as the CAT but now known as the CogAT).

A unique opportunity was thus made available to relate the results of the individually administered Stanford–Binet to a group abilities test. The norming sample for the 1972 edition of the Stanford–Binet could be selected from students tested on the CogAT (where there were children of appropriate age) and from the siblings of CogAT examinees. Thus, the norming sample was stratified not only in terms of the usual variables of age, geographic region, socioeconomic status, and so forth, but also on the basis of intelligence as measured by the CogAT. In fact, this last variable was the primary basis for stratification in the development of the 1972 Stanford–Binet norms. The usual demographic characteristics were included only by virtue of the CogAT standardization. Norming sites were selected on the basis of available CogAT samples.

In addition, another major change in the 1972 norms was the inclusion of black and Hispanic children. The 1937 norm group had included only white American–born English–speaking subjects:

> In the 1972 Binet testing, black and Spanish surnamed youngsters *were* tested, but only if the primary language spoken in the home was English. . . . Thus, with this qualification the 1972 norms are designed to be inclusive of the United States population without regard to racial or national origin. (Terman & Merrill, 1973, p. 360)

The decision to revise again. The 40–year–old norms had concealed how out–of–date the Stanford–Binet had become by the 1970s. With the new norms the weaknesses in the test, such as out–of–date items, became apparent, and in 1978 the publisher decided to undertake a major revision. The Riverside Publishing Company (a subsidiary of Houghton Mifflin) already had a contract with R. L. Thorndike and Elizabeth Hagen to produce a multi–score individual abilities test, and that project became redirected to the Stanford–Binet revision. Both of these authors were thoroughly experienced in group testing methods, but the revision of the Stanford–Binet included clinical aspects, so the publisher sought the collaboration of Jerome Sattler, a recognized authority on the assessment of children's intelligence.

The result was the most radical revision of the Binet–Simon scale since its inception. In fact, in some respects the Stanford–Binet Intelligence Scale: Fourth Edition is so different from its predecessors that it is hard not to call this revision a completely new instrument. The 1986 revision of the Stanford–Binet has abandoned the age scale entirely and makes use of recent advances in psychometric theory. There are now 15 separate tests, and any examinee may take as few as 8 or as many as 13 tests depending on his or her age and ability. Six tests are common to all age and ability levels: Vocabulary, Bead Memory, Quantitative, Memory for Sentences, Pattern Analysis, and Comprehension. Examinees at the lowest ability levels also take the Absurdities and Copying tests, while older and more able children will add the Memory for Digits, Memory for Objects, Matrices, and Number Series tests instead. At the highest ability levels the Paper Folding and Cutting, Verbal Relations, and Equation Building tests are included in the battery. A complete description of the tests and the order in which they are administered is given in Thorndike, Hagen, and Sattler (1986a).

All earlier forms of the Stanford–Binet had yielded a single score that was called a mental age and yielded an IQ. In the Fourth Edition the authors decided, on the basis of correlational evidence and a hierarchical theory of intelligence, to arrange the tests in four areas: Verbal Reasoning, Abstract/Visual Reasoning, Quantitative Reasoning, and Short–Term Memory. Area scores as well as a total or composite score are provided. All performances are expressed in what the authors call Standard Age Scores (SAS's) in an attempt to avoid some of the connotations of the term *IQ*. SAS's are normalized standard scores that reflect a person's relative position in the distribution, and as such they are essentially the same as deviation IQs. Perhaps the change in terminology will avoid some of the overinterpretation that has become attached to the IQ.

Administering the new Stanford–Binet. An examinee begins the Fourth Edition by taking the Vocabulary Test, which acts as a routing test. Age is

used to determine the starting point on the Vocabulary Test. Age and Vocabulary Test score are then used to determine which additional tests a given individual will take and what entry level will be used with each test. A person starts each subsequent test at the same entry level.

Use of a routing test is unique to the Fourth Edition of the Stanford–Binet and has some interesting advantages and consequences. For example, two examinees of the same age but differing markedly in ability may attempt a few vocabulary items in common, but obtain very dissimilar scores and, thus, disparate entry levels. These diverse entry levels would send them to widely separated places in the other tests (and quite possibly to different tests entirely). The result of this routing procedure is that each examinee is administered those items closest to his or her ability level, thus yielding more information while presumably saving time.

The items in each test are arranged by difficulty, with two items at each difficulty level. Using the entry level, a pair of items is administered. If either item is failed, the two items at the next lower difficulty level are given. Successively easier items are administered until the examinee has passed four items (both items at each of two consecutive difficulty levels). This constitutes the examinee's basal level. When a basal level has been determined (or if the examinee passes both initial items), items of higher difficulty are presented. Successively more difficult items are given until the examinee fails three out of four or all four items at two consecutive levels.

Item selection and norming. As is the case with most modern tests, the new Stanford–Binet underwent careful screening of the items for cultural, linguistic, and gender bias. Prior to initial tryout in 1979, all verbal items (vocabulary, comprehension, and so forth) were screened by a panel of eight reviewers. After the first field trial a second panel of 17 reviewers analyzed the surviving items for bias. All major American ethnic groups were represented on these panels, as described by Thorndike, Hagen, and Sattler (1986b).

More than 5,000 individuals from 47 states and the District of Columbia served as subjects in the standardization of the Fourth Edition. The sample was carefully stratified by geographic region, community size, ethnic group, age, and gender. Socioeconomic status was monitored to assure a representative sample on that basis. More than 600 experienced school psychologists, professors of psychology, and others who had administered either the Stanford–Binet or the Wechsler scales participated in the standardization.

In a number of ways, the Fourth Edition of the Stanford–Binet is the first really new intelligence test that has been produced in many years. (The other new individually administered intelligence test for children, the Kaufman

Assessment Battery for Children, K–ABC [Kaufman & Kaufman, 1983], claims to measure simultaneous and sequential mental processing, but it uses a selection of items similar to those found in the Binet and Wechsler scales, and the general approach is that of a standard point scale like the Wechsler.) The Fourth Edition uses the adaptive testing methods that were developed in the 1970s. The 15 tests, which yield scores for Verbal Reasoning, Abstract/Visual Reasoning, Quantitative Reasoning, and Short–Term Memory as well as a composite score, provide a great deal of information in the testing time used because each examinee spends most of the testing period on items that are neither too easy nor too hard.

On the other hand, few new item types have been found that work as well as the ones developed between 1900 and 1925, so the new Stanford–Binet mainly uses tuned–up versions of old friends. The authors tried several item types beyond those that made it into the final test, but had to select those that were easiest to administer and score, provided the most information within the model of intelligence that they had selected, and fitted into a practical testing period. The Fourth Edition of the Stanford–Binet is complex to administer and relatively time–consuming, but the user receives a commensurately large amount of information in return.

Revisions of the Wechsler Scales

The Wechsler scales have also undergone considerable modification and expansion since the original Wechsler–Bellevue was published in 1939. In 1949 Wechsler published the Wechsler Intelligence Scale for Children (WISC), applying the point scale method and the deviation IQ to the measurement of intelligence in individuals between the ages of 5 and 15. A further downward extension of the Wechsler scales to age 4, the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) was published in 1967.

This invasion of the age range that had been the territory of the Stanford–Binet brought healthy pressure to bear on the older test. Over the years the balance gradually shifted to the point that in the late 1970s the 1974 revision of the WISC (WISC–R) had replaced the Stanford–Binet as the most widely used instrument for measuring children's intelligence. As the reviewers for the Eighth *MMY* noted in 1978, the WISC–R remained very similar to the 1949 version and to the other Wechsler scales, containing the same subtests and many of the same items. However, the point–scale format, the up–to–date norms, and the availability of Verbal and Performance Scale IQ scores gave the test an appeal that clinicians did not find with the Stanford–Binet. The standardization and the test materials are first–rate, as is the case with the new Stanford–Binet and the K–ABC.

The original Wechsler–Bellevue was replaced by the Wechsler Adult Intelligence Scale (WAIS) in 1955 (Wechsler, 1958), and a revised WAIS was released in 1981. Neither the WAIS nor its revision, the WAIS–R, incorporated any major changes; they were updates of the existing tests with some but not all of their predecessors' problems corrected. In his review of the WAIS–R in the *Ninth Mental Measurements Yearbook*, Kaufman (1985) observed that 80 percent of the items were unchanged from the 1955 edition and most of the changes were limited to removing ethnic and gender bias: "The WAIS–R is *the* criterion of adult intelligence, and no other instrument even comes close. . . . Yet the substance of the WAIS–R manual could just as well have been written 20 years ago" (p. 1703). Nevertheless, the WAIS–R is the most widely used individual test for measuring adult intelligence and is second in overall use among psychological instruments to the Minnesota Multiphasic Personality Inventory (Lubin, Larsen, Matarazzo, & Seever, 1985). Matarazzo (Ninth *MMY*) even suggests that it is the best psychological test ever produced, although, as Kaufman's comments above would indicate, not every test user would share this opinion.

NEW THEORIES OF INTELLIGENCE

Thurstone had developed the basic structure of factor analysis and applied it to a large number of tests before World War II. This led, as we have seen, to the multifactor batteries that appeared shortly after the war. In the late 1940s and 1950s psychologists were factor analyzing everything in sight. Particularly with the advent of computers, it became a simple matter to perform analyses of quite large data sets. The number of factors of mental ability increased so rapidly that the field soon approached anarchy, leading McNemar (1964) to wonder what ever happened to the notion of intelligence.

Guilford's Structure of Intellect model. One attempt to bring order out of chaos was provided by J. P. Guilford (1959, 1967, 1982), whose research began as part of the military personnel program of the war and has continued into the 1980s. Guilford posits a model of intellectual organization which he calls the Structure of Intellect or SI. Developed by factor analytic methods, the theory postulates three basic dimensions of intellectual functioning: operations, products, and contents. The most recent version of the theory postulates five kinds of operations: cognition, memory, divergent production, convergent production, and evaluation. There are also five kinds of contents in the theory: visual, auditory, symbolic, semantic, and behavioral. Lastly, there are six kinds of products: units, classes, relations, systems, transformations, and implications. Guilford asserts that there is an ability that corresponds to every

three–way combination of a level from each of the dimensions. For example, there is an ability to remember (memory) visual relations and there is a separate ability to evaluate semantic transformations. All possible three–way combinations lead to 150 distinct abilities.

Guilford and his associates have been able to provide correlational evidence in support of more than 100 of the predicted factors in a very large number of studies over many years using a factor rotation procedure known as Procrustes rotation. In this procedure the investigator seeks to find a prespecified set of factors by rotating the original factors to positions that most closely approximate the factors that are hypothesized to exist. However, others, notably Horn (Horn, 1976; Undheim & Horn, 1977), have shown that Guilford's methodology can confirm random hypotheses just as well as those generated by the theory. Horn concludes "that much of the evidence presented in support of SI theory is not compelling" (1976, p. 442).

Hierarchical models. A second way to organize the large number of factors that had been identified was exemplified in the work of Humphreys (1962) and Vernon (1961), who proposed that mental abilities were organized in a hierarchical structure. At the lowest level were highly specific abilities relating to narrow fields of information or specific skills. These narrow factors were all positively correlated (in the domain of abilities) and could be grouped into clusters or factors that represented broader abilities. The grouping could be continued, yielding more and more general factors, until very broad factors such as Spearman's *g* would appear. These models applied Thurstone's notion of second– and higher–order factors.

The main question with hierarchical models is the problem of how general or specific to leave the factors that form the cornerstones of the theory. Vernon (1961) advocated two broad group factors as the basic dimensions of human ability, a verbal–educational factor (v:ed) and a practical or performance factor (k:m). Below these factors, he claimed, there are a larger number of minor group factors that correspond in many ways to the factors that Thurstone found, and above them is *g*. Humphreys's proposal was an elaboration of a five–level hierarchy originally suggested by Burt (1940).

R. B. Cattell (Cattell, 1963; Horn & Cattell, 1966) also proposed a hierarchical theory that divided *g* into two major subfactors. According to Cattell, these major factors are fluid intelligence (*Gf*) and crystallized intelligence (*Gc*), but unlike Vernon's factors, they differ both in the material with which they deal and in terms of the degree to which they are a product of experience. Because of the importance of Cattell's theory in the development of the new Stanford–Binet and current research on intelligence, it is discussed in detail in chapter 6.

Other current theoretical models. Factorial and hierarchical models of intelligence have dominated much of intelligence theory since World War II, but in recent years an old alternative has reappeared—research on the intellectual correlates of reaction time. Galton had argued that reaction time and sensory sensitivity measures were good indicators of intelligence. As we have seen, this line of research, represented by studies like those of Cattell and Farrand (1896), soon fell under the onslaught of Binet–type tests. So complete was the victory of Binet's method over measures of simple neurological functions that, as Eysenck (1986) points out, even well–conducted studies that supported reaction time measures as indices of intelligence (such as that of Peak and Boring in 1926) went relatively unnoticed. However, in recent years interest in studies of the efficiency of neural processing as evidenced by measures of *complex* reaction time have been drawing attention. Eysenck (1986) has even gone so far as to claim that "it may be stated with some confidence that the 'true' correlation between RT [reaction time] and intelligence, when both RT and IQ are measured without chance errors, would be between –0.7 or –0.8" (p. 605). He suggests that reaction time measures may eventually provide culture–free measures of biological intelligence.

A second promising area of research and theory about mental functioning focuses on cognitive processes. These theories, which are covered in chapter 6, emphasize interpretation of the environment and adaptation to it. One of the theories that attempts to integrate earlier conceptualizations of intelligence with cognitive theory is that of Sternberg (1985), who has proposed what he calls a triarchic theory. This theory builds on Cattell's distinction between fluid and crystallized abilities and emphasizes the roles of problem solving and reaction to novel stimuli. Work such as that of Jensen (1982b) and Eysenck, and theories such as Sternberg's may soon open up possibilities for the first significant advances in mental measurements in many years.

TESTING ON TRIAL

During the war years no one seriously questioned the use of ability tests to make personnel decisions. After the war there was such a rapid expansion of educational institutions and an increased demand for a college education, spurred on by the large influx of returning soldiers whose educational expenses would be paid by the GI Bill, that there was a serious need for educational selection methods. The successful use of tests by the military for personnel selection and placement also increased the attention that industry paid to testing. Two concomitances (or consequences) of this rapid expansion

of testing in education and industry, as Matarazzo (1972) pointed out, were a dramatic increase in the number of tests being used (100 million per year by the early 1960s) and a rapid growth in the number of psychologists providing a variety of services to the public.

Standards within the profession. The decade of the 1950s bore some resemblance to the period just after World War I in that tests were being oversold. They were being touted as capable of doing things that they could not do and of providing simple, quick answers to complex problems. Access to psychological tests was largely unrestricted. Although most of the better test publishers urged caution in interpreting the results of their tests, they did little to control distribution and could do nothing about improper interpretation by unqualified or unprincipled individuals.

As noted earlier, Buros's work to promote improved tests and information about tests had started before the war, but it was an unofficial effort. In 1954 the American Psychological Association (APA) published a series of recommendations designed to set minimum standards for the development and use of commercially produced educational and psychological tests. The standards were revised in 1955, 1966, 1974, and 1985. The latest revision of the *Standards for Educational and Psychological Testing* was a joint effort of the APA, the American Educational Research Association, and the National Council on Measurement in Education. It gave detailed recommendations on what minimum information concerning test construction procedures, reliability and validity of scores, and character of normative samples should be included in test manuals. In recent years the emphasis in the guidelines has shifted toward a concern with the types of interpretations that are appropriate from any given test. Caution and the need to validate any test score interpretation with independent information from other sources are stressed.

The dramatic increase in the number of psychologists and the move of many of these individuals out of clinics and universities into private practice and public schools sometimes made it more difficult to maintain standards of professional conduct. In the 1950s, there were no licensing laws for psychologists. However, within the profession, the APA published a statement on ethical standards for psychologists in 1953, and has revised it periodically since. Over the years, several states have enacted licensing or certification procedures to control who can claim the title of psychologist, and this has also served to control the misuse of tests.

In an effort to ensure, as far as possible, the proper use and interpretation of test scores, most test publishers have also developed minimum educational and experiential standards that individuals must meet before they may purchase

copies of tests that yield potentially sensitive information such as scores on aptitude and personality measures. Test sponsoring agencies such as the College Board also publish information on the correct uses of their products and warn against specific common misuses. There is, of course, a limit to how much control they can exercise in a free society, but the testing industry has made a substantial effort at self–regulation.

Reaction sets in. Efforts at self–policing did not prevent a ground swell of antitest sentiment. By the early 1960s the predictable reaction to the uncontrolled use of tests had set in, and it has continued in one guise or another to the present day. The reaction had a number of identifiable causes. Testing had become a pervasive part of American society in the schools and in employment. Government agencies and private employers were using personality and ability measures in personnel selection without proper validation. Some critics saw tests as an invasion of privacy. Others saw them as an infringement on individual freedom. Still others, as the movement for racial equality got started, saw them as an instrument for racial oppression.

Best–seller lists of the early 1960s included titles such as *The Tyranny of Testing*, *They Shall Not Pass,* and *The Brain Watchers*, all of which warned that tests were being used for nefarious purposes by ill–intentioned psychologists and educators, a Big Brother government, or corporate executives. The uproar eventually precipitated a series of congressional hearings in June 1965 on tests as an invasion of privacy. The hearings were reported in detail in a full (oversize) issue of *The American Psychologist* in November of that year. (I had just started graduate school at the University of Minnesota, where this volume was referred to as the "green giant" because of its distinctive color and size.)

While there was some truth to the criticisms raised in the popular books, they represented "a class of criticism of the assessment field which professional psychologists find objectionable—objectionable because the criticisms are not tempered by careful objectivity or balanced presentation" (Carter, 1965, p. 123). The years since the onset of the attack on testing resumed have seen several spates of such destructive criticism. Jensen (1980) has complained that most critics "are indiscriminate in their criticisms," (p. 18) failing to differentiate those tests to which their criticisms apply from those to which they do not. They also fail to give any empirical basis for their criticisms, fail to offer any useful alternatives, and, worst of all, accuse tests of causing social injustices which they merely reveal.

One of the major reasons that tests have come under such concerted and continuing attack for more than 25 years is that there has been a rapid change in social values during the period, perhaps more rapid than at any other time in

American history. In the opinion of some, it has become socially unacceptable to acknowledge racial, ethnic, gender, or class differences, and any instrument by which such differences are found must, by definition, be biased, and therefore bad. This line of thinking reaches an extreme position in statements such as those of Jackson (1975), in which he argued that any characteristic of the society or of a job or school curriculum on which the performance of one racial group is not equal to that of others is biased, and therefore improper. Positions such as this, which take a debatable conclusion as a basic premise, do not contribute to the solution of practical social problems. They also, if taken at face value, identify a number of roles in current American society where the distribution of talent and income is unbalanced in the opposite direction. The proper function of psychological tests, and of any other form of information that is used for decision making, is to identify *relevant* individual differences without regard to irrelevant demographic characteristics.

There is no question that early intelligence tests required some information specific to the majority culture, and for this they were rightly criticized. Most of the better test publishers have now taken steps to remove cultural bias from their tests *as far as possible*, such as those steps described earlier for the Fourth Edition of the Stanford–Binet, but the score differences between racial and socioeconomic groups have persisted. (There is some question whether it is possible or meaningful, even in theory, to measure intelligence in a manner that is independent of the cultural context. Certainly E. L. Thorndike's position in 1926 suggested this impossibility. See page 57.) A great deal of the effort in intelligence test development at this time is directed to the issue of defining the nature of the differences that have been identified and determining the degree to which tests can be constructed that do not reveal irrelevant differences. Page (1984) has given a vigorous defense of the current state of intelligence measurement. Other articles in the same volume (Plake, 1984) illuminate various legal and technical issues in the measurement of human abilities.

Jensen and the modifiability of IQ. Perhaps the greatest storm in the history of mental testing was the one created by the publication in 1969 of Arthur Jensen's *Harvard Educational Review* monograph, "How much can we boost IQ and scholastic achievement?" In this paper Jensen reviewed in detail the evidence then available for the modifiability of intelligence by environmental manipulations and concluded that the behavioral differences revealed by intelligence measures were largely the result of genetic endowment. This, of course, was exactly consistent with the position taken by Terman, Goddard, Yerkes, and other early testers, quite independently of the Army Alpha test results that had been so improperly interpreted by Brigham. But because Jensen included the results of some research that indicated there might be

differences in the type of intelligence possessed by blacks and whites, the study was seen as primarily a statement of inherited racial differences in intelligence, and it drew a predictable reaction. Jensen was denounced as a racist in the public press and his work was given close and unsympathetic scrutiny by his professional colleagues, who found several weak points in his argument.

What Jensen had actually done was review the evidence on why Head Start programs had not been effective in reducing the intellectual gap between children of different socioeconomic status (SES) levels. Since SES is related to race (there are more black, Hispanic, and Native American children in the lower SES groups), many of the children in the programs were from these minority groups. The fact that Head Start had not been able to change the measured intelligence of children in these groups might be taken as evidence that one's level of intelligence is inherited. Jensen suggested that genetic explanations should not be ruled out a priori because there was some evidence that differences in genotype were partially responsible for differences among individuals in measured intelligence. This, coupled with his suggestion that lower SES children may learn most effectively in a different way than higher SES children, was taken by critics as an assertion of genetically based racial differences in intelligence.

A more recent attempt to raise measured IQ by environmental enrichment, known as the Milwaukee Project (see Garber, 1988; Sommer & Sommer, 1983, 1984), obtained what appeared at first to be large gains in intelligence. Subsequent follow-up showed that although the children who had received the enrichment had higher initial IQs than matched controls, the advantage decreased between early childhood and school entry. Perhaps the most surprising finding of this 10-year study was that environmental enrichment seemed to have no lasting effect on academic performance or attitudes toward school. Educators and psychologists have yet to find a reliable method to raise tested intelligence.

Herrnstein, IQ, and SES. Two years after Jensen's monograph, Richard Herrnstein (1971) published an article in the *Atlantic Monthly* magazine in which he suggested that there was a genetic basis for social class. His argument, which drew on many of the same sources Jensen used, was that:

> Against this background [of American equalitarianism] the main significance of intelligence testing is what it says about a society built around human inequalities. The message is so clear it can be made in the form of a syllogism:

1. If differences in mental abilities are inherited, and
2. if success requires those abilities, and
3. if earnings and prestige depend on success,
4. then social standing (which reflects earnings and prestige) will be based to some extent on inherited differences among people. (pp. 58–63)

Needless to say, this argument was also met with something less than widespread enthusiasm. However, the article itself, which was solicited by the magazine, made no direct racial claims. In fact, as Cronbach (1975) observes, it was primarily the editor's introduction that made what racial comments were to be found in relation to the article. Herrnstein's 1973 book covering his experiences following the publication of this article provides an interesting commentary on the tactics of the extreme liberal movement that was present on American college campuses at the time.

Both Jensen and Herrnstein were excoriated in the press, and Herrnstein was threatened with physical harm and prevented from speaking at various locations around the country. Cronbach (1975), in reviewing the major attacks on testing that have been mounted since 1920, noted the critical role of the zeitgeist in determining how successful the attacks were. The Lippmann–Bagley foray of the 1920s encountered an audience that was not very interested. Tests were still relatively new and, although there had been a rapid expansion in their use, they were not the pervasive influence they had become by 1960. There was also a brief assault by Davis and associates shortly after World War II (see Eells, Davis, Havighurst, Herrick, & Tyler, 1951), but again the climate was not ready. However, by the 1960s both the public and law makers were prepared to join in the attack on tests and their uses. Racial differences in test scores provided a major focus:

> In the America of 1969, to make a statement about race differences even at the level of hypothesis was to offend blacks and threaten their political interests. Many laymen and scholars condemned Jensen [and later Herrnstein] not for false impressions [they] might have given but for making *any* statement about race. (Cronbach, 1975, p. 6)

Legislation and the courts. The first statute containing a clause aimed at eliminating possible discriminatory uses of psychological tests was the Civil Rights Act of 1964. It said in part:

> It shall not be an unlawful employment practice for an employer to give and act upon the results of any professionally developed ability test, provided that such test, its administration or action upon the results thereof, is not designed, intended, or used to discriminate because of race, color, religion, sex, or national origin. (Title VII, Section 703[h])

By 1975 the California legislature had twice passed legislation prohibiting "group mental testing in schools, on the grounds that their effect is to limit the education black children receive" (Cronbach, 1975, p. 1), but both times the bill was vetoed by the governor. Since then, there has been judicial action in California severely restricting the testing of minority children with any form of standardized intelligence test for placement purposes in classes for learning disabled or mentally retarded students. The spate of lawsuits, commission hearings, and so forth is going to take years to resolve itself into a coherent set of social principles. (Most current texts on psychological testing and assessment provide an overview of recent litigation and regulations relating to the use of tests. Anastasi, 1988, and Kaplan and Saccuzzo, 1989, are examples.)

The actions in California are only part of an increasing trend over the past 20 years for the courts to become involved in the use of psychological tests. We have no way to predict how long this trend will continue and how much effect it will have on the measurement of human intelligence. It is clear that the public attention tests have received has encouraged professional psychologists in general and those involved in assessment in particular to regulate their work more adequately. The message from government is unambiguous: either you do it or the legislatures and the courts will. There has been progress in making certain that the inferences drawn from test scores are justifiable, and the problem of bias in testing has been given close scrutiny.

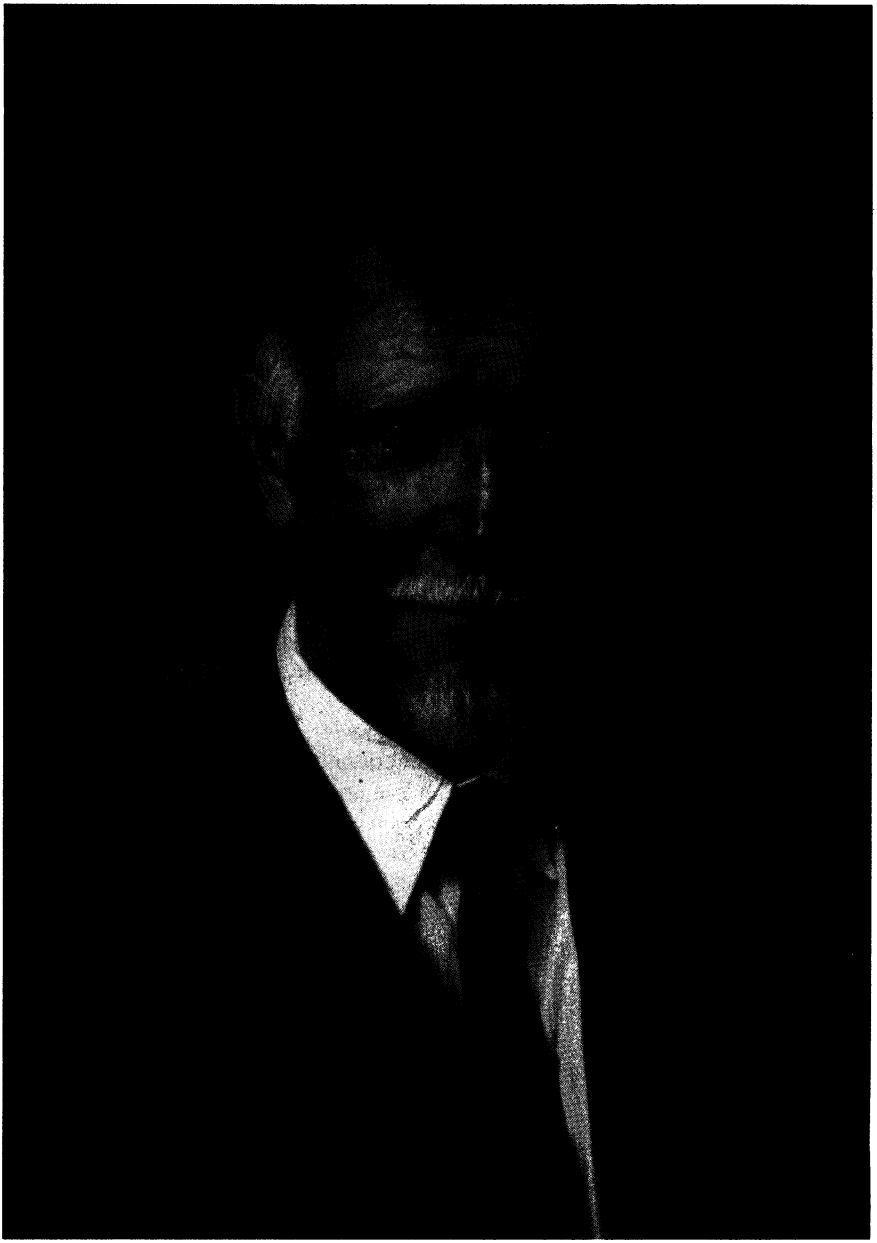
In general, however, a detached observer would have to conclude that the empirical evidence favors the tests. Most attempts to find bias in the well-constructed instruments from the major test publishers have failed, and most of the tests offered as alternatives for use with minority groups have been shown to be either invalid for practical criteria or more biased than the tests they were designed to replace. Progress toward elimination of subjective bias in schooling and the workplace and reward of talent without regard to racial, social, or ethnic background have been two of the major accomplishments of testing. If there is an unequal distribution of developed talent, the solution is to seek to eliminate the disparity by social programs, not to claim that it does not exist. Should well-meaning courts and legislatures substitute their beliefs for the empirical evidence that has been collected over a period spanning more than 25 years, the beneficial effects of tests in providing objective alternatives to the subjective

judgments of teachers and employers will be lost and the possibility of bias in our social institutions will increase.

The last 50 years have been a period of consolidation for psychological measurement. It has moved through a period during which psychologists experimented with alternatives to measures of general intelligence. Factorial and multi–score tests were developed and, for predicting quite specific kinds of training performances, they have been successful. But for predicting broader types of performance such as general scholastic achievement, job training outcomes, and even job performance over a variety of tasks, the evidence is leading back to measures of general ability. Cronbach (1978) believed that "the term 'intelligence test' is a vestige from an obsolete view of human development, but this decade still has to live with it" (p. 250). Given the evidence that has accumulated in recent years, the term may be around for a long time to come. Anastasi (1985) pointed out that our conception of intelligence is broadening and that the traditional view of intelligence as measured by presently available tests is:

> A kind of intelligent behavior that is both developed by formal schooling and required for progress within the academic system.... The particular combination of cognitive skills and knowledge sampled by these tests plays a significant part in much of what goes on in modern, technologically advanced societies. (p. xxviii)

So long as this remains true, such tests will continue to be used to make important decisions because they are better than any alternative that has been offered.



Raymond B. Cattell (1905–)

6

Recent Research on the Nature of Intelligence

COGNITIVE SCIENCE AND INTELLIGENCE

The Rise of Cognitive Science

Several developments converged in the 1970s to give intelligence testing a new lease on life. First, there was the growing realization that the ability profiles provided by multiple–aptitude batteries were not particularly useful (see p. 90). Although there were exceptions, the predictive validities of the several scores from multiple–aptitude batteries were repeatedly found to be not much better than the corresponding validity of one general factor estimated from the same battery. In their summary of 20 years of research on the interaction of ability factors with learning, Cronbach and Snow (1977) concluded:

> It has become fashionable to decry the use of measures of general ability, and sometimes their use has been prohibited in school systems. The attackers usually insist that the tests do not assess ability to learn, and it is often proposed to substitute measures of achievement or "learning styles." . . . Instead of finding general abilities irrelevant to school learning, we find nearly ubiquitous evidence that general measures predict amount learned or rate of learning or both. And, whereas

> we had expected specialized abilities rather than general abilities to account for interactions, the abilities that most frequently enter into interactions are general. (pp. 496–497)

Thus, special abilities failed either to predict educational outcomes better than general ability or to predict which students would profit from specialized educational interventions designed to match their particular patterns of abilities. Second, a hierarchical model of abilities was gradually adopted by American theorists which, while allowing for both broad and narrow abilities, clearly emphasized the role of general ability (see page 97).

Third, cognition returned to psychology. From Watson (1925) until Skinner (1953), American psychology was dominated by the belief that mind was not the proper subject–matter for psychology. Thinking and reasoning were considered complex behaviors that would be explained sometime in the future after elementary mechanisms of learning were adequately understood. By the mid–1960s, however, this promise was wearing thin. Psychology seemed not to be building toward the explanation of complex phenomena but, if anything, was digging increasingly deeper into reductionism. The emergence of the computer as a metaphor for mind, and as a vehicle for testing theories about thinking, finally dethroned behaviorism. Rather swiftly, the mainstream of psychology moved from conditioning, to perception, to thinking and problem solving. By 1985, Anderson proclaimed that "the goal of cognitive psychology is to understand human intelligence and how it works" (p. 1). Thus, in two decades, *intelligence* moved from the periphery of American psychology to its center.

The cognitive revolution had two rather different influences on theories of human intelligence. Some saw that the methods and theories of the cognitive psychologists provided a new way to understand what intelligence and other ability tests were really measuring. John Carroll, Earl Hunt, Robert Sternberg, and Richard Snow were leaders in this effort. Others, however, were not at all concerned with intelligence as an individual difference construct. These investigators sought to develop general theories of human cognition and, at times, to simulate their theories in computer programs that then displayed an artificial intelligence. Both of these efforts will be reviewed in this chapter.

The Challenge of Process

Although most research on intelligence has focused on the *products* of intelligent thinking, both theoreticians and clinicians have long called for greater attention to the *process* of intelligent thinking. In 1964, McNemar wondered whether intelligence could ever be more than an individual difference

construct. As he put it, would two supergeniuses marooned on an island ever discover the need for a construct like intelligence? He challenged psychologists to look beyond test scores to "come to grips with the *process*, or operation, by which an organism achieves an intellectual response" (p. 881). Others have made similar pleas. Seven years earlier, in his presidential address to the APA, Cronbach (1957) argued:

> Sophistication in data analysis has not been matched by sophistication in theory. The correlational psychologist was led into temptation by his own success, losing himself first in practical prediction, then in a narcissistic program of studying his tests as an end in themselves. A naive operationism enthroned theory of test performance in the place of theory of mental processes. (p. 675)

In this Cronbach echoed Thurstone (1947), who considered a factor–analytic study of abilities only the first step in a research program. Ability factors identified in such studies should be investigated in experiments designed to manipulate and thus identify "the processes which underlie" the factors (p. 55). But such experiments had little appeal in a psychology dominated by behaviorism, and so the research program Thurstone advocated had to await the rediscovery of mental process by the mainstream of experimental psychology.

Cognitive Science and the Computer

Recent research on intelligence has been driven by a renewed interest in cognition in many fields. *Cognitive science* is the term now commonly used to refer to this new blend of computer science, cognitive psychology, linguistics, neuropsychology, philosophy, and instructional psychology. Although roots of the cognitive revolution may be traced to many earlier sources, several observers see 1956 as the pivotal year in the development of cognitive science. In that year, Newell and Simon reported their success in devising a computer program that could actually prove theorems in logic; Bruner, Goodnow, and Austin published their *Study of Thinking*; and Miller published a seminal paper on short–term memory in which he argued that the capacity of this memory store seemed to be limited by "the magic number seven" (Newell & Simon, 1972, p. 4). The cognitive revolution gathered momentum in the 1960s, and achieved ascendancy during the 1970s (Gardner, 1985).

The computer has contributed importantly to this revolution in at least two ways. The first and more obvious contribution has been as a metaphor for human cognition. The second and more important contribution has been as a

tool for developing and testing theories of cognition. The computer as metaphor for cognition has taken several forms. At the simplest level, direct analogies have been made between the hardware of the computer and the human cognitive system. Computers have devices for encoding information from external sources (tape/disk readers, keyboards), temporarily storing it (memory buffers), transforming it (central processors), retaining it on long–term storage devices (tapes, disks), and producing output (printers, video displays). Although more sophisticated than previous metaphors for thought (such as a telephone exchange), the computer metaphor is incomplete. For example, some researchers have begun to question the extent to which theorizing about cognition has been constrained by the serial–processing, digital computer. New research programs based on parallel processing may circumvent some of these problems.

Some analogies between computers and human cognition go considerably beyond such comparisons of the superficial characteristics of hardware. In particular, it is argued that similar principles govern the functioning of any system that processes information. Fodor (1981) and others who espouse this computational metaphor for thought treat the mind as a device for manipulating symbols. At this level of abstraction, differences in hardware, whether electronic or neurophysiological, are thought to be irrelevant. Whether such an assumption is tenable is a hotly debated issue in cognitive science.

However, the computer provides much more than an analogy for thought. Its greater contribution has been as a tool for developing and testing theories of cognition, or, as Anderson and Bower (1973) put it, for experimenting on the nature of the connection between stimulus and response. In this way, the computer has changed the evidentiary base to include something other than human behavior. Theories of thinking and learning can be formalized as computer programs. Programs gain a measure of plausibility if they solve problems using sequences of steps that are similar to the steps used by successful human problem solvers, or when failing to solve problems, make errors that mimic human errors. A constant exchange between those who study human problem solving in the psychological laboratory and those who attempt to develop computer programs that display artificial intelligence (AI) serves to refine and extend both efforts.

Contributions of Cognitive Research

Cognitive science has contributed to the understanding of human intelligence in three ways. First, methods and theories of cognitive science have been applied to existing tests of intelligence, either through experimental analysis of tasks taken from intelligence and other ability tests, or through

careful study of the problem–solving or other information–processing characteristics of individuals identified as more or less able by existing tests. In this way, cognitive psychology offers a new source of evidence on the construct validity of tests. Second, tests of intelligence and narrow abilities are often used to predict performance in some non–test situation (e.g., conventional schooling). Careful study of the knowledge and processing demands of these criterial performances has led to the development of new measurement strategies and suggestions for the refinement of existing measures (Snow & Lohman, 1988). Third, cognitive science has sought to move beyond existing definitions of intelligence grounded in individual differences to develop general theories of thinking and learning. New measures are then developed to estimate particular processes or knowledge structures hypothesized by these theories. Patterns of individual differences on these new measures are then investigated, usually by determining relationships between new measures and scores on existing tests or experimental tasks.

THE THEORY OF FLUID AND CRYSTALLIZED ABILITIES

It is fitting that the most popular current resolution to the debate among Spearman, Thorndike, and Thurstone about the dimensions of intelligence was proposed by an Englishman who received his Ph.D. under Spearman (in 1929), completed a post–doctoral fellowship under E. L. Thorndike (in 1937), and conducted research with both Burt and Thurstone (Cattell, 1971, p. ix). In 1941, Raymond B. Cattell (no relation to James McKeen Cattell) proposed a pseudohierarchical model of human abilities with *two* general factors at the apex rather than the one advocated by Spearman. Each factor was defined by several of the primary factors Thurstone had identified. Cattell called these two factors fluid intelligence (*Gf*) and crystallized intelligence (*Gc*). (Note that these general abilities are symbolized by a *G* to indicate that they may not be the same as Spearman's *g*.)

In the earliest published account of the theory, Cattell (1943) argued that fluid ability was "a purely general ability to discriminate and perceive relations between any fundaments, new or old" (p. 178). Fluid ability was hypothesized to increase until adolescence and then slowly decline. It was thought to represent the "action of the whole cortex" (p. 178). Further, fluid intelligence was thought to be the cause of the general factor found among ability tests administered to children, and among the "speeded or adaptation–requiring" (p. 178) tests administered to adults. Crystallized intelligence, on the other hand, was thought to consist of "discriminatory habits long established in a particular field" that were originally acquired through the

operation of fluid ability, but that no longer required insightful perception (p. 178).

The empirical facts Cattell hoped to explain by this theory were the relative independence of individual differences in speed and power in adult intellectual performance, and their different patterns of growth and decline. The important psychological distinction in the theory was between process (fluid intelligence) and product (crystallized intelligence) (Cattell, 1963).

The theory of fluid and crystallized abilities attracted little attention, primarily because Cattell soon turned away from the study of human abilities and returned to his earlier research interest of applying the methods of factor analysis to the study of personality. He later wrote, "I had not learned . . . that more original and vital ideas than mine have collected dust on bookshelves for lack of exegesis by their parent or some scholarly leader" (Cattell, 1971, p. x).

Twenty years elapsed before Cattell returned to the theory of fluid and crystallized abilities with new data. In the 1963 formulation of the theory, *Gf* was hypothesized to reflect the physiological integrity of the organism useful for adapting to novel situations that, when invested in a particular learning experience, produced *Gc*. Thus, *Gf* was now hypothesized to be physiologically determined, whereas *Gc* was "a product of environmentally varying, experientially determined investments of *Gf*" (Cattell, 1963, p. 4).

Although intuitively appealing, the hypothesis that *Gf* reflects physiological influences and is thus a better measure of the true intelligence of an individual is perhaps the most controversial aspect of the theory. Several prominent theorists accept the fluid–crystallized distinction, and some also subscribe to the investment theory of aptitude, but without assuming that fluid ability represents something more innate than crystallized ability. For example, Cattell's student and collaborator Horn (1976) interpreted *Gf* simply as "facility in reasoning, particularly in figural or non–word symbolic materials" (p. 445). Cronbach (1977) went even further and argued that "fluid ability is itself an achievement" that reflects the "residue of indirect learning from varied experience" (p. 287). More recently, Horn (1985) echoed the same theme: "There are good reasons to believe that *Gf* is learned as much as *Gc*, and that *Gc* is inherited as much as *Gf*" (p. 289). *Gc*, said Horn, reflects individual differences in "accultural learning," whereas *Gf* reflects individual differences in "casual learning" and "independent thinking" (Horn, 1985, pp. 289–290). Horn and others point out that, if tests of fluid abilities were somehow better estimates of the physiological integrity of the organism and if achievement tests were more a product of experience, then scores on tests of fluid abilities should show relatively higher heritabilities,

which they do not (Horn, 1985; Humphreys, 1981; Scarr & Carter–Saltzman, 1982). These theorists also reject using tests of fluid ability as measures of "capacity" or "potential" against which achievement can be gauged (Cronbach, 1977; Humphreys, 1985; R. L. Thorndike, 1963). On the contrary, some argue that fluid abilities are among the most important *products* of education and experience (Snow & Yalow, 1982).

Tests of Fluid and Crystallized Abilities

Tests of fluid ability require novel problem solving. These tests require subjects to reason with moderately novel figural or symbolic stimuli. For this reason, complex spatial tests often load strongly on the *Gf* factor (Lohman, 1979). Memory span tests and other measures of what Jensen (1969) calls Level I ability often load significantly on the *Gf* factor as well (Horn, 1985).

Tests of crystallized ability, on the other hand, require the examinee to display understanding of concepts and skills taught in some domain, particularly in school. Verbal knowledge and skills are emphasized, although numerical computation and mechanical knowledge tests often load significantly on *Gc*.

Horn's Revision of *Gf–Gc* Theory

The most important change in *Gf–Gc* theory in recent years has been the addition of several other second–order factors to the model. These developments were summarized somewhat differently by Cattell (1971) and by Horn (1985). Horn (1985) identified ten second–order factors: two deep processing factors (fluid ability and crystallized ability), three perceptual organization factors (visualization, clerical speed, auditory thinking), three associational processing factors (short–term acquisition and retrieval, long–term storage and retrieval, and correct decision speed), and two sensory reception factors (visual sensory detection, and auditory sensory detection). Figure 6.1 shows how these factors can be arrayed along a continuum that progresses from surface to deep processing, or from infancy to adulthood.

The model is frankly speculative. Nevertheless, it summarizes much of what is known about the organization of human abilities, and is in the main consistent with the abilities Carroll (in press) has thus far identified in his massive review and reanalyses of 60 years of factor analytic studies of human abilities. Recent research on the four broad factors in this model that have been most widely studied is summarized next.

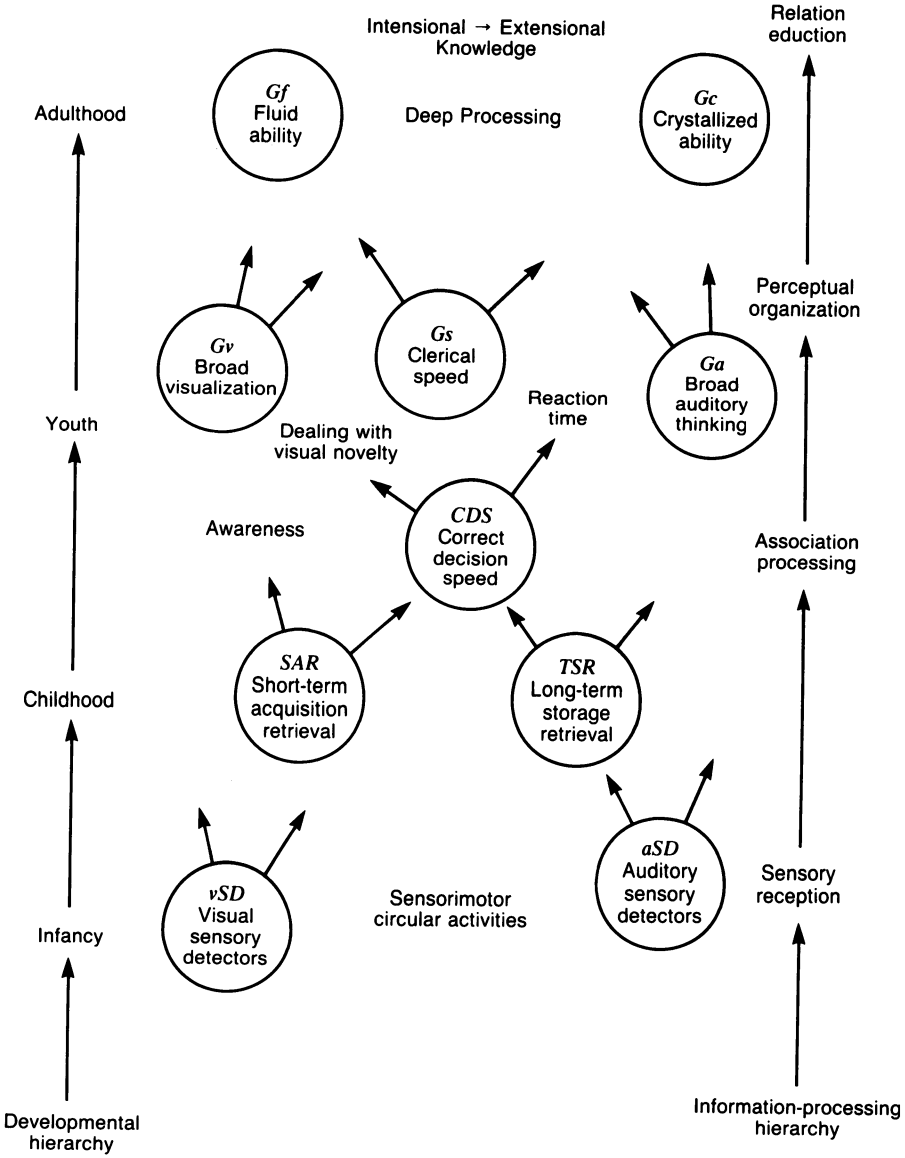


Figure 6.1. A model of ability organization within developmental and information–processing hierarchies.

(Adapted with permission of John Wiley & Sons, Inc., from "Remodeling Old Models of Intelligence" by J. L. Horn in *Handbook of Intelligence* by B. B. Wolman (Ed.), 1985, New York, John Wiley & Sons. Copyright © 1985 by John Wiley & Sons.)

PROCESS THEORIES OF ABILITY CONSTRUCTS

Verbal–Crystallized Ability

Verbal abilities hold a prominent place in all theories of intelligence. It is not surprising, then, that some of the first efforts to understand intelligence in terms of cognitive processes focused on verbal abilities. Hunt and his colleagues have reported several studies of the information–processing characteristics of subjects who differed in verbal–crystallized abilities. This work is of particular interest because it showed the strengths and weaknesses of both the newer cognitive–experimental approach and the traditional correlational approach to the study of intelligence. The aim of this line of research is aptly summarized in the question "What does it mean to be high verbal?" (Hunt, Lunneborg, & Lewis, 1975). The method used in this and several other studies was to select college students with extremely high or low scores on the verbal section of a college entrance examination, to administer to these subjects a battery of presumably well understood experimental tasks, to estimate information–processing scores for each subject on each experimental task, and then to relate these scores to scores on the reference verbal ability tests.

For example, in one experimental task, subjects were required to compare pairs of letters of the alphabet, and to respond yes if the two letters were physically identical (as in "aa" or "AA") or no otherwise (as in "aA" or "ab"). In a second task, similar pairs of letters were presented, but this time pairs were to be judged according to their names. Thus, in Task 1, the correct answer to the pair "Aa" would be no, whereas in Task 2, the correct answer would be yes. An information–processing model for Task 1 (Physical Comparison) would posit processes for encoding the appearance of the two letters, comparing these representations, and then responding. A model for Task 2 (Name Comparison) would include all of the processes required by Task 1 plus an additional process to retrieve the name codes. Thus, the difference between the time to respond to a given pair of letters in Task 2 and the same pair of letters in Task 1 provides an estimate of the time needed to perform this additional process. Correlations between these difference scores and measures of verbal comprehension are typically about $r = -.3$, suggesting that subjects high in verbal ability access name codes faster than subjects low in verbal ability.

These and other results are consistent with both a hierarchical model of human abilities and with current theories of the way knowledge is represented in memory. In particular, the information–processing tasks used by Hunt et al. (1975) appear to measure specific verbal abilities found in the lower branches of

hierarchical models of abilities. Performance on many of these tasks depends on the subject's ability to remember the order in which information was presented, sometimes represented in models of memory by a special type of memory code called a linear–order (Anderson, 1983). Such a code preserves the sequential structure of an event: what came first, then next, then next, and last. Spelling also requires this sort of memory code; one must not only remember the correct letters, but also their sequence.

Research relating scores on experimental tasks to scores on verbal ability tests also has revealed important limitations in efforts to generalize from laboratory tasks to test behavior. First, seemingly simple experimental tasks can measure different abilities in different subjects. For example, Hunt and others have used a sentence verification task in which subjects are shown a sentence such as "star above plus," then a picture such as $\begin{bmatrix} + \\ * \end{bmatrix}$, and then must determine whether the picture and sentence agree. However, minor variations in procedure can substantially alter the way subjects solve this task (Glushko & Cooper, 1978). More importantly, in any given procedure, subjects can differ in the way they solve the task: some creating a mental picture from the phrase and comparing it with the picture, and some converting the picture to a verbal description and comparing that description with the phrase (Macleod, Hunt, & Mathews, 1978).

A second limitation stems from the low correlations between scores representing particular information processes on experimental tasks and scores on reference tests of verbal abilities. Low correlations probably mean that much of the knowledge or some of the cognitive processes that account for general crystallized abilities (*Gc*) as measured by tests are not required by the experimental tasks. Experimental tasks in which subjects are required to infer the meaning of unfamiliar words from context sometimes show much higher correlations than do simple laboratory tasks with both *Gc* scores and general reasoning scores (Sternberg & Powell, 1983). This suggests that the low correlations obtained by Hunt et al. (1975) may properly estimate the contribution of specific verbal processes to *Gc*. Much of the remaining variability in *Gc* is better attributed to the ability to apply general reasoning skills and prior knowledge to the task of understanding verbal material and learning from it.

Spatial–Visualization Ability

Spatial tasks have long been used as psychological tests. Before 1915, Porteus had used such "performance" tasks to estimate the intelligence of linguistically different or handicapped examinees. Spatial tasks also figured prominently in the Army Beta examinations of World War I. Beginning with

Kelley (1928) and then El Koussy (1935), such tasks were studied in their own right, and several specific spatial abilities were identified (Smith, 1964). However, spatial or figural reasoning tasks have continued in their role as measures of general abilities, particularly *Gf* or fluid ability.

As with verbal abilities, cognitive research on spatial abilities may be divided into (a) attempts to develop general theories of spatial thinking that ignore individual differences, and (b) attempts to explain individual differences on existing tests of spatial abilities, either through correlations with laboratory tasks or through the construction of information–processing models for particular spatial tests. In contrast to recent research on verbal abilities, however, only a few studies have examined correlations between scores from laboratory tasks and scores from existing tests. Instead, most effort has been directed toward attempts to build information–processing models that describe how subjects solve particular spatial tests. This is because most spatial tests are process–intensive in the same way that most verbal tests are knowledge–intensive. In other words, although some interesting processing occurs when subjects take a vocabulary test (Sternberg & McNamara, 1985), most of the complex processing occurred when the words were learned. Conversely, although spatial knowledge has an important impact on spatial problem solving (Lohman, 1987), whether subjects solve such problems depends heavily on the processes they employ during the test.

Research on how subjects solve spatial tests has turned up several surprises. One persistent finding has been that rarely do all subjects solve figural tasks in the same way. For example, in a series of experiments on visual comparison processes, Cooper (1982) identified two markedly different strategies. Some subjects appeared to rely on a serial, analytic process to compare forms whereas others relied on a parallel, holistic process. Complex tasks—such as the paper–folding tasks or form–board tasks commonly seen in mental tests—elicit an even wider range of alternative solution methods. Some subjects solve items on such tests by generating mental images that they then transform holistically. Other subjects rely on general reasoning skills or external aids (such as line drawings) to solve problems. Others use even different processes. But most subjects use more than one type of processing, generally shifting from one strategy to another as problems increase in difficulty (Lohman, 1987). Strategy shifting may partially explain why complex spatial tests are often good measures of *g* or *Gf*. Appropriate flexibility in adapting solution methods to meet personal limitations and changing item demands appears to be a central aspect of any process theory of *Gf* (Snow & Lohman, 1988).

Fluid Reasoning Ability

There has been considerably more research on reasoning or general fluid ability than on either general crystallized or general visualization abilities. However, attempts to understand how subjects solve *Gf* tasks such as analogies, classification, and series completion that have ignored differences in processing strategy (by averaging over items) or reduced the need for alternative strategies (by drastically simplifying items) have generally produced experimental tasks that show little relationship with scores on reference *Gf* tests. Put another way, simple items that are all solved in the same way by all subjects probably require little of what we call intelligence.

The effects of simplifying a complex task so that it could be studied experimentally and ignoring within–person strategy shifts were perhaps most evident in Sternberg's (1977) early work on analogical reasoning. Sternberg hypothesized that subjects use several different or "component" processes when solving analogies such as "Up is to down as left is to (a) back (b) right" or A:B::C:D1, D2. According to Sternberg's theory, subjects (1) first read and understand each term in the analogy (*encoding*), (2) determine the relationship between the *A* and *B* terms (*inference*), (3) infer the relationship between the *A* and *C* terms (*mapping*), (4) generate an ideal answer by applying the *A–B* relationship to *C* (*application*), and (5) compare their ideal answer with the options provided (*comparison*). If none of the presented options meet the subjects' criterion for acceptability, they then recycle through some or all of the preceding steps (*justification*), and finally choose an option and respond (*response*). Component processes were assumed to be executed serially. Different models were then formulated by deleting particular processes (e.g., mapping, justification) and by specifying different modes of execution for a given process (e.g., self–terminating or exhaustive). Sternberg obtained three important results. First, models were quite successful in accounting for variabilities in response latencies and, to a lesser extent, in response errors. Second, the data from most subjects were well fitted by a single model, suggesting that most subjects used the same strategy. Third, estimates of speed of executing particular component operations showed small and inconsistent relationships with reference reasoning tests. Unexpectedly, the highest correlations were observed for the preparation–response component. Thus, the componential analysis appeared successful, but those components hypothesized to reflect the essence of reasoning seemed not to measure reasoning at all.

Later studies in which better–practiced subjects attempted more complex items did show significant correlations between component scores and scores on reasoning tests (Bethell–Fox, Lohman, & Snow, 1984; Sternberg & Gardner,

1983). It appears that problems must be more than trivially difficult before individual differences in reasoning are observed. Further, items must also vary somewhat in the processing demands they place on examinees. This means that problems must be moderately novel.

Novelty is an ancient theme in the psychology of individual differences. From Stern (1912/1914) to Sternberg (1985), theorists have argued that intelligence is best displayed when tasks are relatively novel. Cognitive psychologists are only beginning to understand how subjects transfer prior learnings to analogous situations (Gick & Holyoak, 1983). The problem, of course, is that what is novel for one person may not be novel for another person or even for the same person at a different time. It appears that inferences about how subjects solve items that require higher–level processing must be probabilistic, since the novelty of each item varies for each person.

Snow (1981) has integrated these and other research results in the following hypothesis on the nature of fluid and crystallized abilities. His perspective is similar to that adopted by the authors of the Fourth Edition of the Stanford–Binet:

> *Gc* may represent prior assemblies of performance processes retrieved as a system and applied anew in instructional or other performance situations not unlike those experienced in the past, while *Gf* may represent new assemblies of performance processes needed for more extreme adaptations to novel situations. The distinction is between *long–term* assembly for transfer to *familiar* situations vs. *short–term* assembly for transfer to *unfamiliar* situations. (p. 360)

Mental Speed

The fourth broad factor in Horn's (1985) model is sometimes called general speed, sometimes clerical speed, and sometimes simply, mental speed. There is a new interest in this construct, whatever it is called. However, like most other ability constructs, mental speed has a long history in educational and psychological measurement. E. L. Thorndike, Spearman, and Thurstone all addressed the question of whether mental speed should be distinguished from power or altitude.

Individual differences in mental speed have been studied in several paradigms, two of which are noted here. Research in the first paradigm initially sought to estimate the subjects "natural" rate of thinking (Hunsicker, 1925). This search led to the identification of several personality factors such as carefulness, persistence, and impulsivity that described subjects' typical trade–off

between speed and accuracy, and to the identification of several cognitive speed factors, such as perceptual speed, clerical speed, and eventually, to claims of a general speed factor.

Research in the second paradigm, which may be traced back to Galton, has sought to define intelligence as a physiological rather than as a psychological or sociocultural construct. Thus, the aim is to determine the integrity and efficiency of neurological mechanisms thought to underlie intelligent thought and action. Preferred indicators of intelligence in this paradigm are measures of sensory acuity, speed of detecting a stimulus or discriminating between two stimuli, and, in more recent work, patterns in recordings of electrical activity in the brain. Work in this paradigm had hardly begun before it was abandoned by most psychologists, partly because of studies like that of Wissler (1901), but also because of the success of Binet's test. However, interest in the area has been reawakened with some new findings.

One of the major new methods for investigating mental speed has been the study of reaction time (RT). The primary dependent measure in much cognitive research is response latency, usually on simple tasks. The main goal of cognitive researchers like Hunt, Snow, and Sternberg was to develop and test information–processing models of theoretically interesting cognitive tasks or of tests commonly used to estimate important ability constructs, not to propose new measures of mental speed. However, this was precisely the goal of another group of investigators. Led by Arthur Jensen in the U.S. and Hans Eysenck in the U.K., these researchers saw possibilities for new measures of intelligence in response latencies on simple tasks and other indices of cognitive efficiency presumably unaffected by intention or experience.

Jensen's work. One of the more extensive series of investigations on mental speed conducted in recent years is that of Jensen, who sparked new interest in the relationship between reaction time (RT) and *G* by showing significant correlations between choice (or discrimination) RT and measures of *G*. Jensen's work has generated much discussion, in part because his goal seems to be to isolate a culture–free measure of intelligence or *G*. Individual and group differences on such a measure could then not be interpreted "as reflecting only differences in cognitive contents and skills that persons have chanced to learn in school or acquire in a cultured home" (Jensen, 1980, p. 704).

The apparatus Jensen has used in his studies is shown in Figure 6.2. The box contains a center "home button" surrounded by eight light/button pairs. Different light/button pairs can be covered to vary the number of available stimulus–response pairs from 1 to 8. The task is to hold a finger on the home button until one of the exposed lights is activated, and then to turn it off as

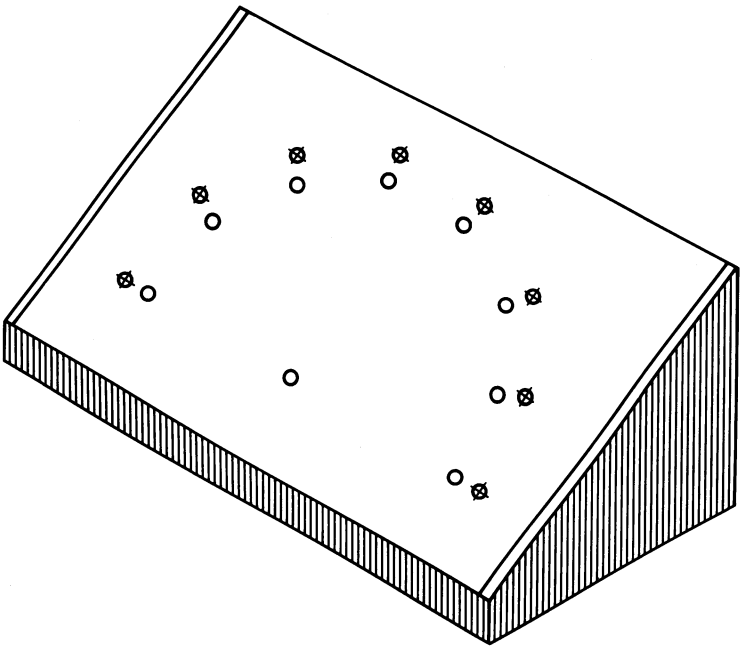


Figure 6.2. Apparatus used in Jensen's choice reaction–time studies. Red push buttons are indicated by circles; green lights, by crossed circles.

(Reprinted with permission of The Free Press, a Division of Macmillan, Inc., from *Bias in Mental Testing* by Arthur R. Jensen. © 1980 by Arthur R. Jensen.)

quickly as possible by moving the finger from the home button to the button directly below the activated light. Two time intervals are recorded: time between onset of the stimulus light and release of the home button (called reaction time), and the additional time required to move the finger to the button below the activated light and to press it (called movement time). In a typical experiment, subjects receive a few practice trials, followed by 15 trials at each of four levels of task complexity: 1, 2, 4, or 8 light/button pairs exposed. Typically, RT increases linearly with the log of the number of buttons exposed. Jensen finds that the slope of this function, which is taken as an estimate of the rate at which a person processes a single unit of information, is negatively correlated with G, $r = -.41$ being the most often cited correlation. In addition, the correlation between RT and G increases as task complexity is increased from 1 to 8 light/button pairs, suggesting that the greater the information processing burden, the greater the demand on G. Jensen's work has been praised by some (e.g., Eysenck, 1982) and criticized by others (e.g., Carroll, 1987; Longstreth, 1984).

Although there is still controversy about Jensen's work, there has been some consensus on the main findings. First, the correlation between *G* and RT is generally somewhat lower for the simple RT condition (1 light/button pair exposed) than for the discriminative RT conditions (2 or more light/button pairs exposed). Second, correlations between discrimination RT and *G* vary widely. However, replicable correlations are generally in the –.2 to –.4 range. Third, the variability in RT over trials often correlates as highly with *G* as does mean or median RT. Thus, attention control (or, conversely, distractibility) may be as important as speed of processing in this task. Fourth, Jensen's claim that RT increases linearly with the log of the number of exposed light/button pairs has been repeatedly confirmed. However, other investigators have been unable to confirm Jensen's claim that individual differences in the slope of this line correlate with *G*.

Eysenck's work. Eysenck (1982, 1988) has proposed a theory of intelligence with an even stronger physiological flavor. Following Hebb (1949), Eysenck (1988) distinguishes among biological intelligence, psychometric intelligence, and social intelligence. Biological intelligence "refers to the structure of the human brain, its physiology, biochemistry, and genetics which are responsible for the possibility of intelligent action" (p. 3). Biological intelligence is thought to be the purest, most fundamental intelligence, because it is "least adulterated by social factors." Eysenck claims that physiological intelligence can be measured by the electroencephalogram (EEG), evoked potentials, galvanic skin responses, and perhaps reaction times.

Psychometric intelligence is defined as that intelligence which is measured by psychometric tests. In addition to the core of biological intelligence, psychometric intelligence is also determined by cultural factors, education, family upbringing, and socioeconomic status. However, since only a fraction of the variance in psychometric intelligence (i.e., IQ) can be attributed to genetic factors (Eysenck estimates 70 percent), IQ should not be confused with the true, biological intelligence.

Finally, social intelligence reflects the ability to solve problems an individual encounters in life. But since so many noncognitive factors are reflected in such performances, Eysenck argues that "social intelligence is far too inclusive a concept to have any kind of scientific meaning" (p. 45). Thus, for Eysenck, intelligence is a concept that is best studied at the physiological (or even neurological) level, only indirectly represented in intelligence tests, and obscured almost entirely in performances in the real world. This is a rather extreme view, and is not widely shared, at least by American academics.

Summary. Critics of studies that report correlations between measures such as RT, inspection time, evoked potentials, and *G* somewhat cynically argue

that the best predictor of the correlation obtained is the date of the study: The first correlation reported is usually strikingly high, but then the magnitude of the reported correlation declines almost linearly with year of publication, eventually stabilizing on a value in the –.1 to –.4 range. Such correlations are theoretically interesting, but do not justify attempts to replace existing intelligence tests with RT measures, or interpretations of *G* as a purely physiological phenomenon.

One need not descend to the level of neurons to find a plausible account of the role of mental speed in models of intelligence. For example, the rate at which activation spreads through regions of memory, the rate at which an activated memory loses its activation, and the level of activation needed to allow further processing—these are all important constructs in modern theories of memory (Anderson, 1983). Direct study of these variables would seem more useful than the study of isolated tasks that have not been designed to reflect particular theories of cognition. Individual differences in mental speed undoubtedly have an important impact on all of cognition. But neither theory nor empirical evidence justifies attempts to define *G* in terms of speed, while ignoring the larger contributions of level or altitude in both process and knowledge to this construct we call intelligence.

ATTEMPTS TO MOVE BEYOND EXISTING TESTS

It has long been recognized that theories of human intelligence have been limited by the selection of tasks included in particular intelligence tests or in factor–analytic studies of abilities. Several theorists have proposed schemes for defining the universe of intelligent behaviors, cognitive functions, or tasks. The framework can then be used to select or construct tests of different facets of intelligence. Guilford's (1959, 1985) Structure of Intellect model (see pp. 96–97) is one example. Cattell's (1971) theory (see pp. 111–112) is another. This section summarizes the more recent work of Robert Sternberg and his triarchic theory of intelligence, and the efforts of Thorndike, Hagen, and Sattler (1986a) to construct a theory–based individual intelligence test.

Triarchic Theory

Sternberg (1985) has proposed a three–pronged or triarchic theory of intelligence. The theory is triarchic because it contains three subtheories: a contextual subtheory, an experiential subtheory, and a componential subtheory. The *contextual subtheory* attempts to specify those behaviors that would be considered intelligent in a particular culture. Sternberg argues that, in any culture,

contextually intelligent behavior involves purposive adaptation to the present environment, selection of a more nearly optimal environment, or shaping of the present environment to fit better one's skills, interests, and values. The nature of the adaptation, selection, or shaping can vary importantly across cultures. In the sociocultural milieu of the United States, Sternberg argues that the prevailing contextual theory of intelligence involves three main elements: problem–solving or fluid abilities, knowledge–based or crystallized abilities, and social and practical abilities.

However, even if a particular task is thought to require intelligence, contextually appropriate behavior is not equally "intelligent" at all points along the continuum of experience with that class of tasks. According to the *experiential subtheory*, intelligence is thought to be best demonstrated when the task or situation is relatively novel, or when learners are practicing their responses to the task so they can respond automatically and effortlessly. Although many have suggested that tasks must be moderately novel to measure intelligence, Sternberg's theory is unique in its claim that the ability to automatize processing is also a good indicator of intelligence. To date, no convincing evidence has been advanced to support this hypothesis.

Finally, in the *componential subtheory*, Sternberg attempts to specify the cognitive structures and processes that underlie all intelligent behavior. Contextually appropriate behavior at relevant points in the experiential continuum is said to be intelligent to the extent to which it involves certain types of processes. Three types of processes are hypothesized: *metacomponents*, which control processing, and enable one to monitor and evaluate it; *performance components*, which execute plans assembled by the metacomponents; and *knowledge acquisition components*, which selectively encode and combine new information, and selectively compare new information to old information.

Evaluation of the Triarchic Theory

Some have argued that intelligence as measured in the tradition of Binet and Wechsler is best construed as scholastic aptitude. This tendency to narrow the scope of intelligence tests has been countered repeatedly by those who would extend measurement to domains such as social intelligence (Thorndike, 1920a), creativity (Guilford, 1959), or musical ability (Gardner, 1983) that are sampled inadequately or not at all by existing tests. Those who would extend the purview of existing tests tend to view *intelligence* as an adjective rather than as a noun, and argue that tests of intelligence should sample from all domains of activity that are valued as intelligent in the culture. Sometimes

these unmeasured abilities are essential features of the theorist's own implicit theory of intelligence or that of a larger social group.

Those who view *intelligence* as a noun usually equate intelligence with individual differences in a particular type of cognition, such as "eduction of relations and correlates" (Spearman, 1923) or "judgment" (Binet & Simon, 1905b). However, others view the noun *intelligence* as a shorthand expression for all individual differences in cognition, and argue that a good test of intelligence presupposes a good theory of cognition (Hunt, 1986) or at least a good sample of "the repertoire of intellectual skills and knowledge available to the person at a particular point in time" (Humphreys, 1986, p. 98). Sternberg's triarchic theory attempts to satisfy both of these demands: His contextual theory recognizes the cultural relativity implied when intelligence is treated as an adjective, whereas his componential theory appears "intended to cover most if not all of the territory of cognitive psychology" (Carroll, 1986, p. 325).

Whether or not Sternberg succeeds in his efforts to develop new measures of practical intelligence or better measures of other aspects of intelligence, he has clearly succeeded in unifying diverse—even antagonistic—traditions in research on intelligence into a single framework. "With his prolific research, writing, and editing activities, Robert Sternberg has probably done more than any other contemporary psychologist to bring back into attention fundamental questions about intelligence—what it is, how it can best be observed and measured, and how it relates to other domains of behavior" (Carroll, 1986, p. 325).

The New Stanford–Binet

Although the studies by Sternberg and others have extended our knowledge about the nature of intelligence, they have not yet resulted in improved methods of measurement. Thus, the recent remodeling of the Stanford–Binet—the granddaddy of all intelligence tests—is an event of considerable import in the long history of intelligence testing. The Fourth Edition of the Stanford–Binet differs importantly from previous editions and from other intelligence tests in that it was explicitly designed to reflect the theory of fluid and crystallized abilities, and thus to wed theory with measurement practice.

Figure 6.3 shows the particular version of *Gf–Gc* theory on which the new Stanford–Binet is based. This theory is a combination of the hierarchical model of intelligence of Vernon, and the pseudohierarchical model of intelligence of Cattell. The three–level hierarchy includes a general reasoning factor, *g*, at the top. Three broad group factors—crystallized abilities, fluid–analytic abilities, and short–term memory—constitute the second level. Three more specific factors make up the third level.

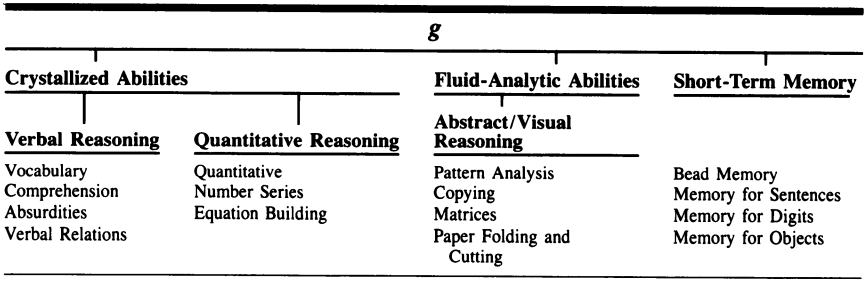


Figure 6.3. Model of ability organization for the Fourth Edition of the Stanford-Binet.

(Reprinted with permission of The Riverside Publishing Company from *The Stanford–Binet Intelligence Scale: Fourth Edition. Guide for Administering and Scoring* by R. L. Thorndike, E. P. Hagen, and J. M. Sattler. © 1986 by The Riverside Publishing Company.)

Following Snow (1981), *g* is interpreted "as consisting of the cognitive assembly and control processes that an individual uses to organize adaptive strategies for solving novel problems" (Thorndike, Hagen, & Sattler, 1986a, p. 3). Crystallized abilities are represented by both verbal and quantitative reasoning tasks. These abilities "are greatly influenced by schooling, but they are also developed by more general experiences outside of school" (p. 4). Fluid–analytic abilities are estimated by figural and spatial tasks. Fluid abilities are thought to involve "the flexible reassembly of existing strategies to deal with novel situations" (p. 4). Further, the authors acknowledge that these abilities are also developed, but from more general experiences than schooling. Finally, the short–term memory factor is represented by tests requiring memory for beads, sentences, digits, or objects. It is unusual to see a short–term memory factor represented at the same level as the fluid and crystallized factors. Nevertheless, research in three traditions—clinical, factor–analytic, and information–processing—attests to the importance of short–term memory.

The Fourth Edition of the Stanford–Binet blends old tasks and new theory. It is the first individual intelligence test based on a widely accepted model of human abilities. (The Kaufman Assessment Battery for Children was perhaps one other attempt. However, the theory on which it was based was not widely accepted by differential psychologists at the time the battery was assembled, and is even less well accepted today.) Yet the new Stanford–Binet preserves a link to the past by using a carefully selected array of familiar tests.

Speculations on the Future

Kant proposed a threefold categorization of mental faculties: cognitive, affective, and conative; or, knowing, feeling, and willing. By this account, a complete theory of mind must explain not only the cognitive dimension, but also the emotional and intentional dimensions as well. Indeed, theorists are once again beginning to argue that affect must be included in accounts of learning and cognition (Snow & Farr, 1987). Thus, one direction research on intelligence seems to be taking is to expand its horizons to include affective dimensions long recognized as central to intelligence (first by Binet and later by Wechsler), but rarely combined with the systematic study of the cognitive dimensions. A theory of intelligence thereby becomes more than an account of human cognition, but of affect and perhaps even volition as well. Even when *intelligence* is treated as a noun, its purview knows no bounds.

The second trend in research on intelligence is moving in the opposite direction. Binet's test was originally designed to predict performance in school. Whatever larger purposes he might have hoped the test might serve, or that others have actually used such tests for, it is clear that intelligence tests have always been most heavily used as measures of scholastic aptitude. Researchers have begun to uncover the reasons why such tests predict success in conventional forms of schooling as they have come to understand better the nature of the knowledge and thinking skills that are required by school learning tasks that are also estimated by intelligence tests. Items on intelligence tests often *appear* to differ markedly from the sort of school learning tasks they predict. For example, matrix completion problems and/or paper–folding problems do not appear to have much in common with understanding a story or solving an algebra word problem. Yet intensive analyses have revealed a commonality in the processes students use to solve both test problems and school learning tasks (Snow & Lohman, 1984).

Thus, somewhat paradoxically, new developments in the measurement of intelligence—particularly the sort of intelligence required by and developed through formal schooling—may well come about more through the careful study of achievement than through continued scrutiny of tasks modeled after existing intelligence tests. And there are reasons to be optimistic that such research may produce intelligence tests that are useful for instruction in more ways than are existing tests.

This possibility can be better understood if intelligence and achievement are viewed as points on a continuum of transfer (or novelty) rather than as

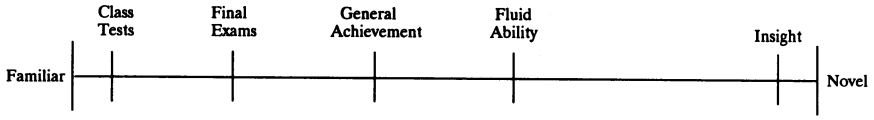


Figure 6.4. Hypothetical continuum of transfer for general achievement and ability tests.

qualitatively distinct constructs. The horizontal line in Figure 6.4 plots types of tests according to the amount of transfer they require for the typical examinee. At the far left, problems on the tests duplicate those taught. As one moves to the right, problems become increasingly novel, and require increasing transfer. For example, if students have learned to add numbers in columns, then one could present these same addition facts in column format to require minimum transfer; presenting the same facts horizontally would require some transfer; embedding the problems in a sentence would require more transfer; embedding them in a matrix problem in which the rule is "Add row 1 to row 2" requires even more transfer. Insight problems require the most transfer. As this example demonstrates, the continuum of novelty in Figure 6.4 is not limited to general ability but can apply to narrower ability constructs as well. It also illustrates the principle that the same task can elicit different processes from different people, depending on their prior experience.

Important educational objectives may be identified all along this line. Students must learn specific skills, but they must also learn to transfer their learnings to unfamiliar situations, and to be creative. Unfortunately, measurement problems increase as one moves from left to right on this scale. Tests that sample no more than those facts and skills explicitly taught are relatively easy to construct and defend, especially when only limited inferences are made from test scores. Tests that require transfer are more difficult to defend because problem novelty varies from individual to individual, and because such tests are usually constructed in ways that encourage broader inferences. Some argue that defensible tests of insight (on the far right) are nonexistent.

Much of the research on intelligence and intelligence tests conducted by Sternberg, Snow, Hunt, Pellegrino, and others during the 1970s could be seen as an effort to start at the line labeled *fluid ability* in Figure 6.4 and move to the left. Both Snow (1978) and Glaser (1976) claimed that the ultimate goal of their research on intelligence was to discover how the thinking skills required by such tests are also required for learning in schools. Although much has been learned from these efforts, specific recommendations for teaching have not been forthcoming. In part, this may be an inevitable conse-

quence of studying tests that were designed to work rather than to reflect a particular theory of cognition. A more fruitful avenue, for education at least, might instead be to begin somewhere near the left of Figure 6.4 and work toward the right. Perhaps then educators might finally learn what to teach the so–called "overachiever," who scores higher on tests of crystallized than on tests of fluid abilities. The recent work of Brown and Ferrara (1985) in estimating a student's "zone of proximal development" exemplifies one effort toward this goal.

SUMMARY AND EVALUATION

Much of the initial optimism about the potential impact of cognitive psychology on the study of human intelligence (e.g., Hunt, Frost, & Lunneborg, 1973; Sternberg, 1977) has been tempered by experience. Hunt now sees some fundamental incompatibilities between the correlational and experimental camps in psychology. He argues that

> Cronbach [1957] thought that general theories of psychological process ought not to ignore individual differences, and vice versa. He was right, and in a general sense the union of the camps is well underway. In my opinion . . . the way to achieve the *scientific* union is to concentrate on understanding how individual differences variables, such as age, sex, genetic constitution, and education, influence the processes of cognition. It does not seem particularly fruitful to try to derive the dimensions of . . . [a trait model] of abilities from an underlying process theory. (Hunt, 1987, p. 36)

There has been a similar tempering of enthusiasm about the prospects for an easy victory over the problem of human intelligence in other quarters of cognitive science, particularly artificial intelligence. Increasingly those who have attempted to develop artificially intelligent systems have come to question their efforts and the constraints that the digital computer has placed on their work. In a summary of this recent history of artificial intelligence, Dehn and Schank (1982) noted that "arrogance about the potential superiority of machine–specific intelligence slowly gave way to a growing respect for human intelligence and its operation. Characteristics of human intelligence that had at first seemed to be weaknesses began to be recognized as strengths" (p. 354). For example, humans tend not to consider all aspects of a problem or to generate and evaluate all possible answers to a problem before deciding upon a course of action. Computers are easily programmed with algorithms that painstakingly consider all factors in a problem before choosing the best answer. However, the computer begins to drown in computation as problems

increase in complexity, such as when the input is a visual scene, or when the number of alternatives that could be generated is unlimited, such as in a chess game. Therefore, AI has shifted from programs that solve problems by brute force to programs modeled after the "satisficing" rules of thumb humans use.

The recent shift to parallel–processing computers and to models of cognition that conform to current theories of brain function takes an even larger step away from the conventional digital computer and the constraints it imposes on efforts to model human cognition. However, some predict that even these efforts are doomed to fail, either because human cognition is not rule–bound (Dreyfus & Dreyfus, 1986) or because higher–level cognitive processes such as judgment and reasoning can be influenced by one's beliefs, values, and intentions (Fodor, 1981; Pylyshyn, 1984).

In short, there has been a growing respect for human intelligence, and a realization that it will not yield to ready explanation by the methods of cognitive science any more than it yielded to ready explanation by the method of factor analysis. Yet factor analysis contributed—and continues to contribute (Carroll, in press; Gustafsson, 1984)—to our understanding of human intelligence. Cognitive science will also continue to contribute to our understanding of intelligence, despite the dire warnings of pessimists. But it should do so with a little less arrogance and, hopefully, with a greater appreciation for the contributions of Binet, E. L. Thorndike, and others who have traveled this same path before.

References

Aikens, H. A., Thorndike, E. L., & Hubbell, E. (1902). Correlation among perceptive and associative processes. *Psychological Review, 9*, 374–382.

Alexander, W. P. (1935). Intelligence, concrete and abstract. *British Journal of Psychology Monograph Supplement No. 19.*

American Medical Association (1913). The problem of the feeble–minded among immigrants: Editorial. *Journal of the American Medical Association, 60*, 209–210.

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

Anastasi, A. (1985). Mental measurements: Some emerging trends. In J. V. Mitchell (Ed.), *Ninth Mental Measurements Yearbook* (pp. xxiii–xxix). Lincoln, NE: Buros Institute of Mental Measurements.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge: Harvard University Press.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: Freeman.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory.* Washington, DC: Winston.

Bagley, W. C. (1922a). Educational determinism: Or democracy and the I.Q. *School and Society, 15*, 373–384.

Bagley, W. C. (1922b). Professor Terman's determinism: A rejoinder. *Journal of Educational Research, 6*, 371–385.

Bethell–Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*, 205–238.

Binet, A. (1909). *Les idées modernes sur les enfants* [Modern ideas about children]. Paris: Flammarion.

Binet, A. (1911). New investigation upon the measure of the intellectual level among school children. *L'Année Psychologique, 17*, 145–201.

Binet, A., & Henri, V. (1895). La psychologie individuelle. *L'Année Psychologique, 2*, 411–465.

Binet, A., & Simon, T. (1905a). Upon the necessity of establishing a scientific diagnosis of inferior states of intelligence. *L'Année Psychologique, 11*, 163–191.

Binet, A., & Simon, T. (1905b). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique, 11*, 191–244.

Binet, A., & Simon, T. (1905c). Application of the new methods of the diagnosis of the intellectual level among normal and subnormal children in institutions and in the primary schools. *L'Année Psychologique, 11*, 245–336.

Binet, A., & Simon, T. (1908). The development of intelligence in the child. *L'Année Psychologique, 14*, 1–94.

Binet, A., & Simon, T. (1915). *A method of measuring the development of the intelligence of young children* (3rd ed.) (C. H. Town, Trans.). Chicago: Chicago Medical Book. (Original work published 1911)

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Trans.). Baltimore: Williams & Wilkens. (Original works published separately in 1905, 1908, 1911)

Bonser, F. G. (1910). The reasoning ability of children of the fourth, fifth and sixth school grades (Contributions to Education No. 37). New York: Teachers College Bureau of Publications.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 33*, 35–37.

Brigham, C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37*, 158–165.

Brown, A. L., & Ferrara, R. A. (1985). Diagnosing zones of proximal development. In J. Wertsch (Ed.), *Culture, communication and cognition: Vygotskian perspectives* (pp. 273–305). Cambridge: Cambridge University Press.

Buros, O. K. (1941). *The 1940 Mental Measurements Yearbook.* Highland Park, NJ: Mental Measurements Yearbook.

Buros, O.K. (1977). Fifty years in testing: Some reminiscences, criticisms, and suggestions. *Educational Researcher, 6* (7), 9–15.

Burt, C. (1909). Experimental tests of general intelligence. *British Journal of Psychology, 3*, 94–177.

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy and Training, 1*, 93–112.

Burt, C. (1940). *The factors of the mind.* London: University of London Press.

Burt, C. (1945). The assessment of personality. *British Journal of Educational Psychology, 15*, 107–121.

Burt, C. (1952). Autobiography. In E. G. Boring (Ed.), *A history of psychology in autobiography: Vol. 4* (pp. 53–73). New York: Russell & Russell.

Carroll, J. B. (1986). Beyond IQ is cognition [Review of *Beyond IQ: A triarchic theory of human intelligence*]. *Contemporary Psychology, 31*, 325–327.

Carroll, J. B. (1987). Jensen's mental chronometry: Some comments and questions. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 297–307). New York: Falmer Press.

Carroll, J. B. (in press). Factor analysis since Spearman: Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *The Minnesota symposium on learning and individual differences: Abilities, motivation, and methodology.* Hillsdale, NJ: Erlbaum.

Carter, L. F. (1965). Psychological tests and public responsibility: Introduction. *American Psychologist, 20*, 123–124.

Cattell, J. McK. (1890). Mental tests and measurements. *Mind, 15*, 373–381.

Cattell, J. McK., & Farrand, L. (1896). Physical and mental measurements of the students of Columbia University. *Psychological Review, 3*, 618–648.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22.

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

Cheng, C. Y. (1922). Psychological tests found in China during Han and Wei dynasties. *Chinese Journal of Psychology, 1*, 7. (From *Psychological Abstracts*, 1928, *2*, Abstract No. 461)

Cooper, L. A. (1982). Strategies for visual comparison and representation: Individual differences. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence: Vol. 1* (pp. 77–124). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1–14.

Cronbach, L. J. (1977). *Educational psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1978). Review of the BITCH Test. In O. K. Buros (Ed.), *Eighth Mental Measurements Yearbook* (p. 250). Highland Park, NJ: Gryphon Press.

Cronbach, L. J. (1986). Signs of optimism for intelligence testing. *Educational Measurement: Issues and Practice, 5*, 23–24.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Dehn, N., & Schank, R. (1982). Artificial and human intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 352–391). Cambridge: Cambridge University Press.

Dickson, V. E. (1922). The Oakland plan. In Terman, L. M., Dickson, V. E., Sutherland, A. H., Franzen, R. H., Tupper, C. R., & Fernald, G., *Intelligence tests and school reorganization* (pp. 30–50). Yonkers: World Book.

Dobzhansky, T. (1962). *Mankind evolving: The evolution of the human species.* New Haven, CT: Yale University Press.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine.* New York: Free Press.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Dvorak, B. J. (1947). The new U.S.E.S. general aptitude test battery. *Journal of Applied Psychology, 31,* 372–376.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. (1951). *Intelligence and cultural differences; A study of cultural learning and problem–solving.* Chicago: University of Chicago Press.

El Koussy, A. A. H. (1935). The visual perception of space. *British Journal of Psychology Monographs, 20.*

Eysenck, H. J. (1982). *A model for intelligence.* New York: Springer.

Eysenck, H. J. (1986). Inspection time and intelligence: A historical perspective. *Personality and Individual Differences, 7,* 603–607.

Eysenck, H. J. (1988). The concept of "intelligence": Useful or useless? *Intelligence, 12,* 1–16.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science.* Cambridge: MIT Press.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193–202.

Freeman, F. S. (1962). *Theory and practice of psychological testing* (3rd ed.). New York: Holt, Rinehart & Winston.

Garber, H. L. (1988). *The Milwaukee Project: Preventing mental retardation in children at risk.* Washington, DC: American Association on Mental Retardation.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (1985). *The mind's new science.* New York: Basic Books.

Garrett, H. E., Bryan, A. I., & Perl, R. E. (1935). The age factor in mental organization. *Archives of Psychology,* 176.

Gick, M., & Holyoak, K. (1983). Schema induction and analogical reasoning. *Cognitive Psychology, 15,* 1–38.

Glaser, R. (1976). The processes of intelligence and education. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 341-352). Hillsdale, NJ: Erlbaum.

Glushko, R. J., & Cooper, L. A. (1978). Spatial comprehension and comparison processes in verification tasks. *Cognitive Psychology, 10*, 391–421.

Goddard, H. H. (1911). Two thousand normal children measured by the Binet measuring scale of intelligence. *Pedagogical Seminary, 18*, 232–259.

Goddard, H. H. (1916). Introduction to *The development of intelligence in children*, A. Binet & T. Simon (E. S. Kite, Trans.). Baltimore: Williams & Wilkins.

Goddard, H. H. (1920). *Human efficiency and levels of intelligence.* Princeton, NJ: Princeton University Press.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). New York: Erlbaum.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14*, 459–479.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw–Hill.

Guilford, J. P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. *Psychological Review, 89*, 48–59.

Guilford, J. P. (1985). The structure–of–intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 225–266). New York: Wiley.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179–203.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology, 5*, 51–84.

Hebb, D. (1949). *The organization of behavior.* New York: Wiley.

Herrnstein, R. J. (1971, September). I.Q. *Atlantic Monthly*, pp. 43–64.

Herrnstein, R. J. (1973). *I.Q. in the meritocracy.* Boston: Little, Brown.

Holzinger, K. J. (1938). Relationships between three multiple orthogonal factors and four bifactors. *Journal of Educational Psychology, 29*, 513–519.

Horn, J. L. (1976). Human abilities: A review of research theory in the early 1970s. *Annual Review of Psychology, 27*, 437–485.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 267–300). New York: Wiley.

Horn, J. L., & Cattell, R. B: (1966). Refinement and test of the theory of fluid and crystallized ability intelligences. *Journal of Educational Psychology, 57*, 253–270.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417–411.

Huey, E. B. (1910). The Binet scale for measuring intelligence and retardation. *Journal of Educational Psychology, 1*, 435–444.

Huey, E. B. (1912). The present status of the Binet scale of tests for the measurement of intelligence. *Psychological Bulletin, 9*, 167.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475–483.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum Press.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201–224). New York: Wiley.

Humphreys, L. G. (1986). Describing the elephant. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 97–100). Norwood, NJ: Ablex.

Hunsicker, L. M. (1925). A study of the relationship between rate and ability (Contributions to Education No. 185). New York: Teachers College Bureau of Publications.

Hunt, E. (1986). The heffalump of intelligence. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 101–108). Norwood, NJ: Ablex.

Hunt, E. (1987). Science, technology, and intelligence. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology*

on testing and measurement: The Buros–Nebraska symposium on measurement and testing: Vol. 3 (pp. 11–40). Hillsdale, NJ: Erlbaum.

Hunt, E. B., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), *The psychology of learning and motivation: Vol. 7* (pp. 87–122). New York: Academic Press.

Hunt, E. B., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology, 7*, 194–227.

Intelligence and its measurement: A symposium. (1921). *Journal of Educational Psychology, 12*, 123–147, 195–216, 271–275.

Jackson, G. G. (1975). Comment on "Educational uses of tests with disadvantaged students." *American Psychologist, 30*, 88–92.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 1–123.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1982a). *Straight talk about mental tests.* New York: Free Press.

Jensen, A. R. (1982b). The chronometry of intelligence. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence: Vol. 1* (pp. 255–310). Hillside, NJ: Erlbaum.

Jones, L. V. (1949). A factor analysis of the Stanford–Binet at four age levels. *Psychometrika, 14*, 299–331.

Kamin, L. J. (1974). *The science and politics of I.Q.* New York: Erlbaum.

Kaplan, R. M., & Saccuzzo, D. P. (1989). *Psychological testing: Principles, applications, and issues* (2nd ed.). Monterey, CA: Brookes/Cole.

Kaufman, A. (1985). Review of the WAIS–R. In J. V. Mitchell (Ed.), *Ninth Mental Measurements Yearbook* (pp. 1700–1703). Lincoln, NE: Buros Institute of Mental Measurements.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children. Interpretive Manual.* Circle Pines, MN: American Guidance Service.

Kelley, T. L. (1914). Comparable measures. *Journal of Educational Psychology, 5*, 589–595.

Kelley, T. L. (1916). Further logical aspects of the Binet scale. *Psychological Review, 23*, 407–411.

Kelley, T. L. (1928). *Crossroads in the mind of man.* Stanford, CA: Stanford University Press.

Kelly, R. L. (1903). Studies from the psychological laboratory of the University of Chicago. Psychophysical tests of normal and abnormal children. *Psychological Review, 10*, 345–372.

Kent, G. H. (1940). Review of *The measurement of adult intelligence. Psychological Bulletin, 37*, 251–254.

Kirkpatrick, E. A. (1900). Individual tests of school children. *Psychological Review, 7*, 274–280.

Kite, E. S. (1916) *The development of intelligence in children* by A. Binet and T. Simon (E. S. Kite, Trans.). Baltimore: Williams & Wilkens. (Original works published separately in 1905, 1908, and 1911)

Kohs, S. C. (1914). The Binet–Simon measuring scale for intelligence: An annotated bibliography. *Journal of Educational Psychology, 5*, 215–224, 279–290, 335–346.

Kohs, S. C. (1917). An annotated bibliography of recent literature on the Binet–Simon scale (1913–1917). *Journal of Educational Psychology, 8*, 425–438, 488–502.

Krugman, M. (1939). Some impressions of the revised Stanford–Binet scale. *Journal of Educational Psychology, 30*, 594–603.

Kuder, F. (1986). Memoirs mostly. *Educational and Psychological Measurement, 46*, 1–10.

Kuhlmann, F. (1911). Binet and Simon's system for measuring the intelligence of children. *Journal of Psycho–Asthenics, 15*, 79–92.

Lippmann, W. (1922). Articles in *New Republic, 32*, 213–215, 246–248, 275–277, 297–298, 328–330.

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literature* (Tech. Rep. No. 9). Stanford, CA: Stanford University, School of Education. (NTIS No. AD–A075 972)

Lohman, D. F. (1987). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence: Vol. 4* (pp. 181–248). Hillsdale, NJ: Erlbaum.

Longstreth, L. E. (1984). Jensen's reaction–time investigations of intelligence: A critique. *Intelligence, 8*, 139–160.

Lorge, I. (1936). The influence of the test upon the nature of mental decline as a function of age. *Journal of Educational Psychology, 27*, 100–110.

Lubin, B., Larsen, R. M., Matarazzo, J. D., & Seever, M. (1985). Psychological test usage patterns in five professional settings. *American Psychologist, 40*, 857–861.

Macleod, C. M., Hunt, E. B., & Mathews, N. N. (1978). Individual differences in the verification of sentence–picture relationships. *Journal of Verbal Learning and Verbal Behavior, 17*, 493–508.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

McNemar, Q. (1942). *The revision of the Stanford–Binet scale: An analysis of the standardization data.* Boston: Houghton Mifflin.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871–882.

Mercer, J. R. (1979). *System of Multicultural Pluralistic Assessment (SOMPA). Technical manual.* New York: The Psychological Corporation.

Mullan, E. H. (1917). *Mentality of the arriving immigrant* (Public Health Bulletin No. 90). Washington, DC: U.S. Government Printing Office.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice–Hall.

Norsworthy, N. (1906). The psychology of mentally deficient children. *Archives of Psychology, No. 1.*

Otis, A. S. (1916). Some logical aspects of the Binet scale. *Psychological Review, 23*, 129–152, 165.

Otis, A. S. (1917). A criticism of the Yerkes–Bridges Point Scale, with alternative suggestions. *Journal of Educational Psychology, 8*, 129–150.

Page, E. B. (1984). Struggles and possibilities: The use of tests in decision making. In B. S. Plake (Ed.), *Social and technical issues in testing* (pp. 11–38). Hillsdale, NJ: Erlbaum.

Peak, H., & Boring, E. G. (1926). The factor of speed in intelligence. *Journal of Experimental Psychology, 9*, 71–94.

Peterson, J. (1925). *Early conceptions and tests of intelligence.* Yonkers: World Book.

Pichot, P. (1968). Alfred Binet. In *International encyclopedia of the social sciences: Vol. 2* (pp. 74–77). New York: Macmillan.

Pintner, R. (1931). *Intelligence testing* (2nd ed.). New York: Holt.

Pintner, R., & Paterson, D. G. (1915). The Binet scale and the deaf child. *Journal of Educational Psychology, 6*, 201–210.

Pintner, R., & Paterson, D. G. (1917). *A scale of performance tests.* New York: Appleton.

Plake, B. S. (1984). *Social and technical issues in testing: Implications for test construction and use.* Hillsdale, NJ: Erlbaum.

Porteus, S. D. (1915). Mental tests for the feebleminded: A new series. *Journal of Psycho–Asthenics, 19*, 200–213.

Pressey, S. L., & Pressey, L. W. (1918). A group point scale for measuring general intelligence with first results from 1,100 school children. *Journal of Applied Psychology, 2*, 250–269.

Pylyshyn, Z. W. (1984). *Computation and cognition.* Cambridge: MIT Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rist, R. C. (1970). Student social class and teacher expectations: The self–fulfilling prophecy in ghetto education. *Harvard Educational Review, 40*, 411–451.

Ruch, G. M. (1925). Minimum essentials in reporting data on standard tests. *Journal of Educational Research, 12*, 349–358.

Samelson, F. (1975). On the science and politics of IQ. *Social Research, 42*, 466–488.

Samelson, F. (1982). H. H. Goddard and the immigrants. *American Psychologist, 37*, 1291.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego: Jerome M. Sattler.

Scarr, S., & Carter–Saltzman, L. (1982). Genetics and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 792–896). Cambridge: Cambridge University Press.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research. *American Psychologist, 36*, 1128–1137.

Scott, C. A. (1913). General intelligence or "school brightness." *Journal of Educational Psychology, 4*, 509–524.

Sharp, S. E. (1898–1899). Individual psychology: A study of psychological method. *American Journal of Psychology, 10*, 329–391.

Simpson, B. R. (1912). Correlations of mental abilities (Contributions to Education No. 53). New York: Teachers College Bureau of Publications.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Smith, I. M. (1964). *Spatial ability*. San Diego: Knapp.

Snow, R. E. (1978). Theory and method for research on aptitude processes. *Intelligence, 2*, 225–278.

Snow, R. E. (1981). Toward a theory of aptitude for learning: Fluid and crystalized abilities and their correlates. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 345–362). New York: Plenum Press.

Snow, R. E., & Farr, M. J. (Eds.) (1987). *Aptitude, learning, and instruction: Vol. 3. Conative and affective process analyses*. Hillsdale, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*, 347–376.

Snow, R. E., & Lohman, D. F. (1988). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational Measurement* (3rd ed.)(pp. 263–331). New York: Macmillan.

Snow, R. E., & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 493–585). Cambridge: Cambridge University Press.

Sommer, R., & Sommer, B. A. (1983). Mystery in Milwaukee: Early intervention, IQ, and psychology textbooks. *American Psychologist, 38*, 982–985.

Sommer, R., & Sommer, B. A. (1984). Reply. *American Psychologist, 39*, 1318.

Spearman, C. (1904a). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Spearman, C. (1914). The theory of two factors. *Psychological Review, 21*, 101–115.

Spearman, C. (1920). Manifold sub–theories of "the two factors." *Psychological Review, 27*, 159–172.

Spearman, C. (1922). A friendly challenge to Professor Thorndike. *Psychological Review, 29*, 406–407.

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition.* London: Macmillan.

Spearman, C. (1925). Agreement on cooperation. *Journal of Educational Psychology, 16*, 423–425.

Spearman, C. (1930). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography: Vol. 1* (pp. 299–333). Worcester, MA: Clark University Press.

Spearman, C. (1931). Our need for some science in place of the word "intelligence." *Journal of Educational Psychology, 22*, 401–410.

Spearman, C. (1939). Thurstone's work re–worked. *Journal of Educational Psychology, 30*, 1–16.

Stern, W. (1914). *The psychological method of testing intelligence* (G. M. Whipple, Trans.). Baltimore: Warwick & York. (Original work published 1912)

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge: Cambridge University Press.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112*, 80–116.

Sternberg, R. J., & McNamara, T. P. (1985). The representation and processing of information in real–time verbal comprehension. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 21–43). Orlando, FL: Academic Press.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38*, 878–893.

Strasheim, J. J. (1926). *A new method of mental testing*. Baltimore: Warwick & York.

Terman, L. M. (1906). Genius and stupidity: A study of the intellectual process of seven "bright" and seven "stupid" boys. *Pedagogical Seminary, 13*, 307–373.

Terman, L. M. (1911). The Binet–Simon scale for measuring intelligence. *Psychological Clinic, 5*, 199–206.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Terman, L. M. (1919). *The intelligence of school children*. Boston: Houghton Mifflin.

Terman, L. M. (1920). The use of intelligence tests in the grading of school children. *Journal of Educational Research, 1*, 20–32.

Terman, L. M. (1922). The psychological determinist; Or democracy and the I.Q. *Journal of Educational Research, 6*, 57–62.

Terman, L. M. (1932). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography: Vol. 2* (pp. 297–331). Worcester, MA: Clark University Press.

Terman, L. M., & Childs, H. G. (1912). A tentative revision and extension of the Binet–Simon measuring scale of intelligence. *Journal of Educational Psychology, 3*, 61–74, 133–143, 198–208, 277–298.

Terman, L. M., Dickson, V. E., Sutherland, A. H., Franzen, R. H., Tupper, C. R., & Fernald, G. (1922). *Intelligence tests and school reorganization*. Yonkers: World Book.

Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L., Galbreath, N., & Talbert, W. (1915). The Stanford revision of the Binet–Simon scale and some results from

its application to 1000 non–selected children. *Journal of Educational Psychology, 6*, 551–562.

Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L. E., Galbreath, N., & Talbert, W. (1917). *The Stanford revision and extension of the Binet–Simon scale for measuring intelligence.* Baltimore: Warwick & York.

Terman, L. M., & Merrill, M. A. (1937). *Measuring intelligence: A guide to the administration of the new revised Stanford–Binet tests of intelligence.* Boston: Houghton Mifflin.

Terman, L. M., & Merrill, M. A. (1960). *Stanford–Binet Intelligence Scale: Manual for the third revision. Form L–M.* Boston: Houghton Mifflin.

Terman, L. M., & Merrill, M. A. (1973). *Stanford–Binet Intelligence Scale: 1973 norms edition.* Boston: Houghton Mifflin.

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology, 8*, 271–281.

Thomson, G. H. (1920). General versus group factors in mental activities. *Psychological Review, 27*, 173–190.

Thorndike, E. L. (1903). *Educational psychology.* New York: Science Press.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Science Press.

Thorndike, E. L. (1919). Tests of intelligence, reliability, significance, susceptibility to special training and adaptation to the general nature of the task. *School and Society, 9*, 189–195.

Thorndike, E. L. (1920a). Intelligence and its uses. *Harper's Magazine, 140*, pp. 227–235.

Thorndike, E. L. (1920b). Intelligence examinations for college entrance. *Journal of Educational Research, 1*, 329–337.

Thorndike, E. L. (1920c). The reliability and significance of tests of intelligence. *Journal of Educational Psychology, 11*, 284–287.

Thorndike, E. L. (1921). On the organization of intellect. *Psychological Review, 28*, 141–151.

Thorndike, E. L. (1923). On the improvement in intelligence scores from fourteen to eighteen. *Journal of Educational Psychology, 14*, 513–516.

Thorndike, E. L. (1924). Measurement of intelligence: I. The present status. *Psychological Review, 31*, 219.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence.* New York: Teachers College Bureau of Publications.

Thorndike, E. L., Lay, W., & Dean, P. R. (1909). The relation of accuracy in sensory discrimination to general intelligence. *American Journal of Psychology, 20*, 364–369.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Thorndike, R. L. (1963). *The concepts of over– and under–achievement.* New York: Teachers College Bureau of Publications.

Thorndike, R. L. (1985). The central role of general ability in prediction. *Multivariate Behavioral Research, 20*, 241–254.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *The Stanford–Binet Intelligence Scale: Fourth Edition. Guide for Administration and Scoring.* Chicago: Riverside.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *The Stanford–Binet Intelligence Scale: Fourth Edition. Technical Manual.* Chicago: Riverside.

Thorndike, R. M. (1978). *Correlational procedures for research.* New York: Gardner Press.

Thurstone, L. L. (1921). A cycle–omnibus intelligence test for college students. *Journal of Educational Psychology, 4*, 265–278.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451.

Thurstone, L. L. (1926). The mental age concept. *Psychological Review, 33*, 268–278.

Thurstone, L. L. (1928). The absolute zero in the measurement of intelligence. *Psychological Review, 35*, 175–197.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*, 406–427.

Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

Thurstone, L. L. (1936a). A new conception of intelligence. *Educational Record, 17,* 441–450.

Thurstone, L. L. (1936b). A new concept of intelligence and a new method of measuring primary abilities. *Educational Record, 17* (Suppl. 10), 124–138.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* No. 1.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Thurstone, L. L. (1952). Autobiography. In E. G. Boring (Ed.), *A history of psychology in autobiography: Vol. 4* (pp. 295–321). New York: Russell & Russell.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Town, C. H. (1911). The Binet–Simon scale and the psychologist. *Psychological Clinic, 5,* 239–244.

Town, C. H. (1915). *A method of measuring the development of the intelligence of young children* (3rd ed.) by A. Binet and T. Simon. (C. H. Town, Trans.). Chicago: Chicago Medical Book. (Original work published 1911)

Tuddenham, R. D. (1963). The nature and measurement of intelligence. In L. Portman (Ed.), *Psychology in the making* (pp. 469–525). New York: Knopf.

Undheim, J. O., & Horn, J. L. (1977). Critical evaluation of Guilford's structure of intellect theory. *Intelligence, 1,* 65–81.

Vernon, P. E. (1961). *The structure of human abilities* (2nd ed.). London: Methuen.

von Mayrhauser, R. T. (1987). The manager, the medic, and the mediator: The clash of professional psychological styles and the wartime origins of group mental testing. In M. M. Sokal (Ed.), *Psychological testing and American society* (pp. 128–157). New Brunswick, NJ: Rutgers University Press.

Wallin, J. E. W. (1911). A practical guide for the administration of the Binet–Simon scale for measuring intelligence. *Psychological Clinic, 5,* 217–238.

Wallin, J. E. W. (1923). The consistency shown by intelligence ratings based on standardized tests and teacher's estimates. *Journal of Educational Psychology, 14,* 231–246.

Warner, F. (1890). *A course of lectures on the growth and means of training the mental faculty.* Cambridge: Cambridge University Press.

Watson, J. B. (1925). *Behaviorism.* New York: Norton.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore: Williams & Wilkins.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams & Wilkins.

Weisenburg, T., Roe, A., & McBride, K. E. (1936). *Adult intelligence.* New York: The Commonwealth Fund.

Weiss, A. P. (1911). On methods of mental measurement, especially in school and college. *Journal of Educational Psychology, 2,* 555–563.

Wells, F. L. (1940). Review of the Wechsler–Bellevue Intelligence Scale. In O. K. Buros (Ed.), *The 1940 Mental Measurements Yearbook* (pp. 264–265). Highland Park, NJ: Mental Measurements Yearbook.

Whipple, G. M. (1904). Reaction–times as a test of mental ability. *American Journal of Psychology, 15,* 489–498.

Whipple, G. M. (1910). *Manual of mental and physical tests.* Baltimore: Warwick & York.

Whipple, G. M. (1914). *Manual of mental and physical tests: Part I: Simpler processes.* Baltimore: Warwick & York.

Whipple, G. M. (1915). *Manual of mental and physical tests: Part II: Complex processes.* Baltimore: Warwick & York.

Whipple, G. M. (1921). The National Intelligence Tests. *Journal of Educational Research, 4,* 16–31.

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Monographs, 3* (6, Whole No. 16).

Wolf, T. H. (1973). *Alfred Binet.* Chicago: University of Chicago Press.

Woodworth, R. S. (1912). Combining the results of several tests: A study in statistical method. *Psychological Review, 19,* 97–123.

Yerkes, R. M. (1919). Report of the psychology committee of the National Research Council. *Psychological Review*, *26*, 83–149.

Yerkes, R. M. (1921). *Memoirs of the National Academy of Sciences: Vol. 15. Psychological examining in the United States Army.* Washington, DC: National Academy of Sciences.

Yerkes, R. M., & Anderson, H. M. (1915). The importance of social status as indicated by the results of the point–scale method of measuring mental capacity. *Journal of Educational Psychology*, *6*, 137–150.

Yerkes, R. M., Bridges, J. W., & Hardwick, R. S. (1915). *A point scale for measuring ability.* Baltimore: Warwick & York.

Yerkes, R. M., & Wood, L. (1916). Methods of expressing results of measurements of intelligence: Coefficient of intelligence. *Journal of Educational Psychology*, *7*, 593–606.

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests.* New York: Holt.

Young, K. (1924). The history of mental tests. *Pedagogical Seminary*, *31*, 1–48.

Chronology

1890— Cattell calls for investigation of mental phenomena by "mental tests."

1891— Franz Boas collects anthropological data on 1,500 schoolchildren.

1894— Gilbert's study of sensory, reaction–time, and memory tasks.

— Kraepelin's work in Germany.

1895— APA appoints committee on mental tests.

— Binet and Henri's paper on individual psychology.

— Founding of *L'Année Psychologique*.

1896— Ebbinghaus's completion test.

1898— (through 1899) Sharp's comparison of current test tasks.

1899— Kirkpatrick's APA speech, calling for tests for schoolchildren.

1901— Wissler's study finding no correlation of reaction–time data to grades.

1902— E. L. Thorndike's paper, critical of current tests.

1903— Damaye's doctoral thesis, outlining standardized items for determining mentally retarded individuals.

— Kelly's call for norms of schoolchildren.

— E. L. Thorndike's *Educational Psychology* published.

— Norsworthy relates test scores to teacher's judgments of ability.

1904— E. L. Thorndike's *An Introduction to the Theory of Mental and Social Measurements* published.

— Spearman introduces two–factor theory (*g*).

— Commission for the Retarded proposes testing students for placement.

— Founding of Ministerial Commission for the Abnormal.

1905— Binet and Simon publish first scale.

1906— Norsworthy recommends profile of abilities.

— Terman's study of Binet's tests.

1908— Binet and Simon publish second scale.

1909— Binet publishes *Les idées modernes sur les enfants*.

1910— Ayres publishes *Laggards in Our Schools*.

1911— Binet and Simon publish final revision.

1911 — Wallin, Huey, Kuhlmann, and Goddard all administer Binet–Simon translations.

1912 — Stern introduces term *Intelligence Quotient.*

— Spearman introduces tetrad equation.

1914 — Knox publishes nonverbal test for immigrants.

1915 — Yerkes introduces Point Scale.

1916 — Yerkes and Wood introduce term *Coefficient of Intelligence.*

— Terman publishes the Stanford Revision of the Binet–Simon Scale.

— Thompson produces artificial correlations that yield *g*.

1917 — U.S. enters WWI. Yerkes heads Committee on the Psychological Examination of Recruits. Scott heads Committee on the Classification of Personnel.

1918 — Army Alpha/Army Beta.

— Oakland, California, adopts three–track system for pupil classification.

1919 — Terman proposes yearly group testing of pupils.

— (through 1920) E. L. Thorndike describes different kinds of intelligence.

1920 — Thomson introduces sampling theory, describing group factors.

1921 — Stoddard publishes *The Revolt Against Civilization.*

— Burt's revision of the Binet–Simon scales.

— National Intelligence Tests developed.

— Thurstone's Cycle–Omnibus Test.

1922 — Lippmann's series of articles criticizing conclusions from military testing and use of tests.

1923 — Brigham's analysis of army data, focusing on national origins.

1926 — E. L. Thorndike introduces four–dimension theory of intellect, and the CAVD.

1931 — Thurstone's method of factor analysis.

1933 — Hotelling's generalization of Pearson's method of principal component analysis.

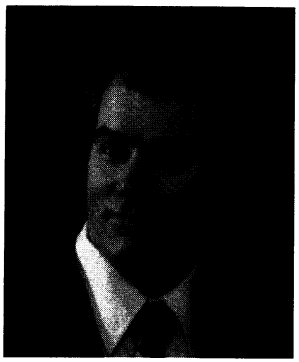
1935 — Alexander's study, finding *g*, several group factors, and two non–ability factors.

1937 — Forms L and M of the Stanford–Binet (2nd edition).

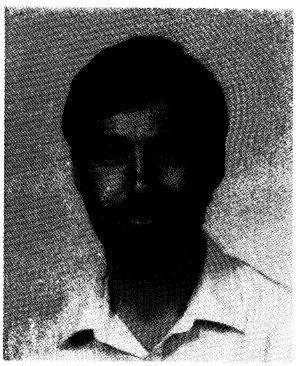
1938— Thurstone's study of Primary Mental Abilities.
— *Mental Measurements Yearbook* begins commissioned reviews of tests.
1939— Wechsler–Bellevue scale. Term *deviation IQ* introduced.
1941— U.S. enters WWII; Army General Classification Test (AGCT).
1943— R. B. Cattell proposes *Gf* and *Gc*.
1947— General Aptitude Test Battery (GATB).
— Differential Aptitude Tests (DAT).
1949— Wechsler Intelligence Scale for Children (WISC).
1953— Flanagan Aptitude Classification Tests (FACT).
1954— American Psychological Association recommends standards for test development and use.
1955— Wechsler–Bellevue replaced by Wechsler Adult Intelligence Scale (WAIS).
1959— Guilford's Structure of Intellect model.
1960— Form L–M of the Stanford–Binet (3rd edition).
1965— Congressional hearings on testing.
— Testing comes under renewed attack.
1967— Wechsler Preschool and Primary Scale of Intelligence (WPPSI).
1970— Cognitive Abilities Test (later known as the CogAT).
1972— Third edition of the Stanford–Binet renormed.
1973— Hunt's first study of the information–processing correlates of intelligence.
1974— WISC revision published (WISC–R).
1977— Sternberg's componential subtheory of intelligence.
1980— Jensen's mental speed studies.
1981— WAIS revision published (WAIS–R).
1982— Publication of Sternberg's *Handbook of Human Intelligence*.
1983— Kaufman Assessment Battery for Children (K–ABC).
1985— Horn's revision of the *Gf/Gc* theory.
— Sternberg's triarchic theory.
1986— Fourth Edition of the Stanford–Binet.

Name Index

Subject Index



Robert M. Thorndike is a professor in the Psychology Department at Western Washington University, Bellingham. He is the author of *Correlational Procedures for Research, Data Collection and Analysis: Basic Statistics,* and with Brislin and Lonner, *Cross–Cultural Research Methods.* He has written numerous book chapters and articles on statistics, measurement, and research methods. Dr. Thorndike, like his father and grandfather, received his BA from Wesleyan University. He received his doctorate degree from the University of Minnesota.



David F. Lohman is an associate professor in the Division of Psychological and Quantitative Foundations at The University of Iowa. He spent the 1988–89 academic year at the University of Leiden in The Netherlands as a Fulbright scholar. Dr. Lohman is the author of numerous journal articles and book chapters, most of which report his research on information–processing analyses of human abilities. Dr. Lohman received his BA degree in psychology from the University of Notre Dame and his PhD in educational psychology from Stanford University.