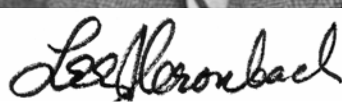
NATIONAL ACADEMY OF SCIENCES

LEE J. CRONBACH
1916—2001

A Biographical Memoir by
RICHARD J. SHAVELSON

Any opinions expressed in this memoir are those of the author and do not necessarily reflect the views of the National Academy of Sciences.

Biographical Memoir



LEE J. CRONBACH

April 22, 1916—October 1, 2001

BY RICHARD J. SHAVELSON

LEE J. CRONBACH WAS THE ONLY "educationist" ever elected to the National Academy of Sciences, as a psychologist. He earned his Ph.D. in educational psychology in the University of Chicago's Department of Education under G. T. Buswell in 1940. Coming from an education department, however, some questioned his status as a psychologist. His request, for example, for a joint appointment in education and psychology upon joining the Stanford University faculty in 1963 was turned down by the Psychology Department, even though he had served as president of the American Psychological Association, had published what became known famously throughout the social sciences as "Cronbach's Alpha" (the widely used reliability coefficient), and a decade later was to be recognized for his scholarship by the National Academy of Sciences (among many distinctions).

Cronbach made major contributions to the fields of educational psychology, psychological testing, and program evaluation throughout a career that spanned over five decades. The citation for his 1973 Distinguished Scientific Contribution Award from the American Psychological Association specifically noted his work on reliability theory, conceptualization of test validity theory, basic concepts in decision making based on psychological tests, integration

of various disciplines in psychology and thereby uncovering new problems for investigation, conceptualization and methods of program evaluation, and pioneering textbooks in psychological testing and educational psychology. The citation said that he had influenced the very foundation of psychological science.

Harbingers of his career in testing were evident at an early age in Fresno, California, where he was born on April 22, 1916, to a homemaker and salesman. According to his sister, Lee was overheard at age four years in a grocery market calculating the unit price of potatoes, drawing the conclusion that the market his mother shopped at charged far more than the market he was in with his babysitter. The eavesdropper reported this feat to Blanche Cummings, a school psychologist and disciple of Lewis Terman, who gave Cronbach an IQ test in 1921, publicized his test score (200), used him in Binet demonstrations, and signed him up in the Terman gifted program. His mother was eager for him to begin school and enrolled him in the upper second grade just as he turned five; she continued to push him and others to advance him throughout his schooling. He graduated from Fresno High School at age 14 and Fresno State College (majoring in chemistry and mathematics) at 18.

Cronbach's awareness of Terman and IQ testing attracted him to psychology. And as a college junior he "was entranced by the Thurstone and Chave monograph (1929) on measuring attitudes" (1989, p. 65); he constructed a Thurstone scale for a class project. He was particularly impressed with "Thurstone's inventive use of mathematics to sharpen the central construct and ferret out equivocal items; the virtue of rigorous engineering analysis of psychological measuring devices became fixed in my mind" (Cronbach, 1989, p. 65).

The connection between psychology and education came in a course on history of American education at Fresno

State (then a teachers' college). Asked to review and present a curriculum-improvement study to his fellow students, a class he "detested," he dissected the study with the scorn of a chemistry student, as he put it. After the presentation, which the students loved, the professor asked if he'd thought of a career in education research; Lee asked if there were such a thing. The suggestion came at a time when he had decided against pursuing chemistry and his path was set toward educational psychology.

To become employable upon graduating college in 1934 Cronbach completed a teaching credential at the University of California, Berkeley, in 1935. His teacher education proved to be telling: "courses required for mathematics teaching introduced me to the writings of Hilbert and Klein, hence to how model building preconditions the mathematical conclusions I had hitherto thought of as 'truths'" (1989, p. 65). He completed his education master's degree in 1937, producing a thesis in which he examined students' understanding of vocabulary in algebra, arguing that a teacher needs to know how words are *misunderstood* by students. No slouch, while working on the master's degree, in 1936 he applied for and was hired to teach at his old high school in Fresno. "Fresno High School hired me to teach, more because my mother was then president of the city PTA than because no teacher who had reached age twenty could be found" (1989, p. 66). He taught mathematics and chemistry briefly, for two and a half years (1936-1938).

In 1938 Cronbach entered full-time Ph.D. study in education at the University of Chicago under G. T. Buswell, earning a doctorate in 1940. His doctoral years then were few but full indeed. Within a month of his arrival he had completed his qualifying examinations. He noted that his "marginal pass entitled me to start dissertation research and cleared the way for marriage" (1989, p. 67). His dissertation

examined individual difference in learning to reproduce nonsense squiggles, extending an earlier study by getting precise data from filmed eye fixations (1941). He married Helen Claresta Bower and, again no slouch, rapidly became a proud father of a family of five children, born between 1941 and 1956. Upon reflecting on his doctoral education, Cronbach admitted that he had rushed through the doctorate "so fast that Chicago had little chance to educate me" (1989, p. 67); he missed, for example, the opportunity to take courses from the psychometrician L. L. Thurstone, who was in another department.

Nevertheless, Cronbach did have time to meet up with Ralph Tyler. In doing his dissertation Cronbach had become dissatisfied with the accuracy of his film measurements. (Dissatisfaction with measurements turned out to be a recurrent theme in Cronbach's career. I recall him telling me in the early 1970s that he seemed never to get to substantive questions because measurement problems needed to be solved first.[1]) To "fix" the measurements he needed a bit more time and financial support. The support came in a "mind-stretching" research assistantship with Tyler on the Eight-Year Study. This was a seminal study of 30 progressive high schools attempting to educate the whole student, focusing on social attitudes and personality, and on problem solving in each subject area. His brief year with Tyler exerted a lasting influence on Cronbach throughout his life, as we shall see, especially in his work on evaluation. Cronbach's admiration for Tyler became clear to me, again in the 1970s. I remember going into his office at Stanford to find him lying on a couch (he had back problems), talking to a gentleman sitting in a chair drawn up next to him. Interrupting himself in mid-sentence, he introduced me to Ralph Tyler. And then he continued speaking without missing a beat, only to stop himself in mid-sentence again, saying something like, "Why

am I lecturing you, Ralph? You know everything I know and much more" (quite an admission from Cronbach).

In 1940 opportunity struck, however serendipitously, and Cronbach left Tyler and took an assistant professorship at Washington State University (WSU, Western Washington) in the Psychology Department, which was housed within the School of Education. The university had such a tough time recruiting because of WSU's remoteness that, as Cronbach tells it, the president would travel around the country seeking faculty and hiring them on the spot. On his return to Pullman from the East Coast, stopping at the University of Chicago, the president realized that he had forgotten to fill the psychology position. He pounced on Cronbach, "despite my ignorance of mainstream psychology" (1989, pp. 68-69), who yearned to get back to the West Coast. In a challenging first year spent in class and in the library he taught introductory, social, child, applied, and industrial psychology and operated a reading clinic as well. "After two years in the stacks, I knew much of the psychological literature from 1900 to 1942" (1989, p. 68).

Cronbach's years at WSU (1940-1946) were full, and the seeds of future research and scholarship were sown. Indeed, he once told me that going to WSU instead of a Stanford or Harvard directly out of graduate school was the best thing that could have happened to him. At WSU he had many opportunities to pursue his scholarly interests without publish-or-perish neuroses: he taught psychology, ran a successful reading clinic, managed academic classification for an Air Force training detachment, surveyed a school system, taught a summer course on "evaluation," and "published dozens of scattered papers" (1989, p. 69). In 1945 he was assigned to teach a course on "mental measurements." He was so dissatisfied with current textbooks that he started at once writing *Essentials of Psychological Testing* (1949). The text not only

became the classic mental measurement text on the topic for the next 40 years but also helped out his rather large family financially. I recall being called into Cronbach's office one day in 1972 to account for why, as a young acting assistant professor, I was consulting on the side for a law school admissions test company. I explained that in my position with a wife and a child, my $14,000 annual academic salary from Stanford didn't quite make ends meet. He thought a moment and then said he understood. His advice: why not write a textbook like his *Essentials?* I told him I'd get on it right away…not in my lifetime but maybe another life.

Toward the end of the war Cronbach served as a military psychologist at the navy's sonar school in San Diego while on leave from WSU (1944-1945). He was hired initially to do selection-test validation but was shifted over to projects training enlisted men to detect submarines. He built situational exercises in which the difficulty of detection varied. In one project, for example, he verified the claim of experienced submarine listeners that they could detect whether a ship topside was opening or closing range. Most significantly, while at the navy laboratory, Cronbach was introduced to Shannon's information theory, which he initially believed might have application to reliability and validity theory. This proved not to be true but it did lead him to the work of Wald and the use of statistical models in reaching decisions. Wald's ideas suggested to Cronbach that test interpretation might be viewed as rules for decision making, and led to a number of significant publications.

Following the war Cronbach returned to Chicago as an assistant professor for two years. In this brief time he met up with Robert Havinghurst, Carl Rogers, and W. E. Henry, all of whom were doing personological studies. Their dynamic assessments of the person seemed to defy current-day validity

theory and practice. This activity led to several papers on reliability and validity while at Chicago, and to his chairing the Committee on Test Standards of the American Psychological Association (1950-1953) where he teamed up with Paul Meehl to reconceptualize test-validity theory in a the seminal paper "Construct Validity in Psychological Tests" (1955).

Cronbach left Chicago and went to the University of Illinois in 1948. After overstimulation at Chicago, he "welcomed the less vibrant collegiality of Urbana" (1989, p. 74) where he "worked up from substitute to regular in the faculty bowling league and sometimes joined a gang of would-be poker players; in Chicago, there would have been seminars on those nights" (1989, p. 74).

His appointment was in the Bureau of Educational Research, where he shared an office with Nathaniel L. Gage and found the freedom and encouragement to work on fundamental questions pertinent to education. He and Gage divvied up the field, with Cronbach focusing on psychometrics and Gage focusing on teacher characteristics and behavior. Cronbach's record of accomplishment while at Illinois is staggering. He served on the committee that produced the first standards for psychological testing, joined the American Psychological Association's Publication Board (1951-1953), served on the APA Board (1952-1958), and was elected president of the association in 1956. And he produced a remarkable record of scholarship in both quality and quantity, including *Essentials of Psychological Testing* (1949), "*Coefficient Alpha and the Internal Structure of Tests*" (1951), *Educational Psychology* (1954), "Conceptual and Methodological Problems in Interpersonal Perception" (1955), Psychological Tests and Personnel Decisions (1957), "Two Disciplines of Scientific Psychology" (1957, APA presidential address), and "Course Improvement through Evaluation" (1963).

After 16 years at Illinois, Cronbach found his "flow of enthusiastic ideas for research…to be drying up, and retreat from the privileges of a research professorship seemed proper" (1989, p. 74). He moved to Stanford University in 1964 as Vida Jacks Professor of Education. His work, once again, picked up steam, publishing "Alpha Coefficients for Stratified-Parallel Tests" (1965), "Generalizability of Scores Influenced by Multiple Sources of Variance" (1965), "Validation of Educational Measures" (1969), "How We Should Measure "Change"—or Should We?" (1970), "Test Validation" (1971), *The Dependability of Behavioral Measurements* (1972), "Beyond the Two Disciplines of Scientific Psychology" (1975), *Aptitudes and Instructional Methods: The Search for Interactions* (1977), and Designing Evaluations of Educational and Social Programs (1982). While at Stanford he received the Distinguished Scientific Contribution Award from the American Psychological Association (1973) and was elected to the National Academy of Sciences (1974) and the National Academy of Education (1965).

Cronbach retired in 1980 because "slow-moving commitments created a logjam, and my activities began to seem like 'work'" (1989, p. 75). Nevertheless, he remained intellectually active right up to the time of his death. Among other activities, he completed a book on a new theory of aptitude that had been started by his close colleague and friend, Richard E. Snow, who had just passed away from cancer (2001). He also published, posthumously, a paper, "My Current Thoughts on Coefficient Alpha and Successor Procedures" (2004), in which, upon the 50th anniversary of his coefficient alpha paper, he reflected on the uses, misuses, and misunderstandings of the reliability coefficient.

Cronbach's research questions, insights and productivity were prodigious, if not neatly developed over time. He

"pursued interests simultaneously and discontinuously" (1989, p. 79), intertwining substantive and methodological scholarship. However, there seems to me to be a center of gravity to all of his work: having to resolve methodological (measurement, statistical, design) issues before proceeding to answer or in reacting to answers to substantive questions. He really was influenced by the Thurstone and Chave work and seems to have believed that on the one hand he could to some degree engineer things to behave (1989):

> I was intensely interested in the logic of questioning people and capturing their characteristics in numerical form. As soon as I had data to inspect, I kept finding that methods of measuring and summarizing introduced artifacts—relationships that had nothing to do with the persons measured and everything to do with the choices the inquirer had made. Furthermore, these choices often buried important relationships. I became insistent that analytic methods should be matched to substantive ideas not chosen on criteria of convenience, familiarity, or statistical stylishness.

But, as we shall see, on the other hand he came to realize that in social, behavioral, and educational research, the standard errors were enormous, contextual factors led to complex interactions between people and their environments that were hard to generalize, and "facts decay" with a short half-life as society changes.

Cronbach's early research exemplifies his simultaneity and discontinuity—that is, his penchant to work on multiple projects at one time and to let one or another project lie fallow over a long period before picking it up again. Driven by interaction with colleagues with interesting research problems, he played his methodological cards. He worked on problems as diverse as:

- social person perception—"one reason for the present confusion [in social perception findings] is that we have tried to use methods which we did not understand sufficiently" (1989, p. 353),

- projective personality techniques—"my sympathetic formulation of the validity questions that had to be faced probably greased the skids for the decline of projective testing" (1989, pp. 80-81), and
- measurement challenges they posed for test-score reliability and validity, especially the validity conundrum (1989, p. 83):

> What was difficult was explaining how to validate measures of motivational and cognitive variables, and indicators of categories of mental illness. Aware of defense mechanisms and response biases, we could not accept content analysis as validating. . . Nor could one point to any kind of behavioral observation or clinical judgment sufficiently valid to be accepted as "criterion" for such variables.

As noted before, these challenges led to the famous Cronbach-Meehl (1955) paper reconceptualizing validity theory by recognizing that all validity could be encompassed under the umbrella of *construct validity*. Modestly, Cronbach notes (1989, p. 83), "My personal contribution was minor, but as committee chair I collaborated on the paper that fully set forth the idea; a coin toss that Meehl insisted on put my name first."

Three strands of research and scholarship emerge from the early days and Cronbach's penchant for simultaneity and discontinuity: measurement theory, individual differences and instructional methods, and program evaluation. His contributions to each are chronicled here, in that order.

While he published prodigiously on reliability throughout his career, his most widely cited measurement paper was published in 1951: "Coefficient Alpha and the Internal Structure of Tests." The coefficient, widely known as "Cronbach's alpha," proved useful for (at least) three reasons. First, it provided a measure of reliability from a single test administration so that repeated occasions or parallel forms of a test were not needed to estimate a test's consistency (following on the work

of, for example, Hoyt, Kuder, and Richardson). Second, the formula was general; it could be applied, for example, to dichotomously scored multiple-choice items or polytimous attitude scales. And third, at a time before computers, alpha was easily calculated from statistics well known by students with only a first course in statistics.

To put alpha in context, in the mid-twentieth century, dissatisfaction had arisen with split-half reliability coefficients. They depended on the particular splitting used to compute them (e.g., items in the first and second half of the test, odd and even items). This dissatisfaction stimulated efforts to develop alternative reliability coefficients that could still be computed using test-item data from a single test administration. Cronbach (1951) established alpha as preeminent among internal consistency reliability coefficients by demonstrating that alpha is the mean of all possible split-half reliability coefficients and showed that Kuder-Richardson formula 20, which preceded alpha, was a special case of alpha for dichotomous data. Alpha estimates the lower bound of the proportion of variance in test scores attributable to all common factors underlying item responses. Thus, alpha does not require the assumption that all items in a test be unidimensional (i.e., measuring only one aspect of individual differences). Consequently it is applicable to common educational tests that measure multiple abilities across items.

Alpha is sometimes misinterpreted as a *coefficient of precision*, which reflects the correlation between scores from one administration with a hypothetical second administration of the same test when no changes in the examinees have occurred. Instead, alpha is a coefficient of equivalence because it represents the correlation between two different tests. In this case the two tests consist of k items randomly drawn from a universe of items like those in the test and administered at the same time. However, alpha is a lower bound to the

coefficient of precision because the correlation between a test and itself would always be higher than the correlation between two different tests. Alpha does not provide any indication of test score inconsistency that might result if repeated testings were made separated in time.

For Cronbach his work on alpha linked mathematics with the real world. He came to the realization that "criticizing test theory [e.g., Kuder-Richardson's assumptions] thus becomes a matter of comparing what the mathematician assumes with what the psychologist can reasonably believe about people's responses" (1989, p. 82).

In an article published posthumously, "My Current Thoughts on Coefficient Alpha and Successor Procedures," Cronbach (2004; see also Shavelson, 2004) expressed doubt that alpha was the best way to study reliability. As was his wont, Cronbach had been dissatisfied with reliability theory from which coefficient alpha sprang for a long time. Following Goodenough (1936) and Thorndike (1947), he recognized that in practice different methods of calculating a reliability coefficient—internal consistency, parallel form, test-retest—defined "true score," the consistent part of a respondent's performance, and measurement error, the inconsistent part, somewhat differently. For example, remembering an answer to a particular question when the same test was administered twice meant that "memory" contributed to a respondent's consistency or true score, but not so upon taking parallel forms of the test. Moreover, he reasoned that measurement error was more complex than what a single undifferentiated term revealed; measurement error had multiple sources such as variations over occasions, raters, forms, tasks, and the like.

To improve researchers' limited understanding of reliability, Cronbach with Goldine Gleser's help set out to produce a handbook on measurement. Its goal was to tell social

scientists, education researchers, and psychologists "how to get help from mathematical systems for transforming the flow of behavior and events into quantitative conclusions... and suggest how to construct mathematical systems suited to that use" (1989, p. 84). The dynamic duo were not very far into the project when they realized that the "overfamiliar terrain of reliability was not well enough understood to be the base for explaining how to think" (1989, p. 87). What started out to be a handbook on measurement became a major reconceptualization of reliability theory in the form of generalizability theory (1972).

Generalizability theory extends reliability theory and affords analyses of complex measurement data to decompose observed test-score variance into an analogue of true-score variance and multiple sources of measurement error. The factorial, within-subjects, random-effects analysis of variance is used to statistically partition total variance into estimated variance components arising from true (person) score variance and facets of measurement error and their interactions such as error due to sampling of items, occasions, raters. Further, results from such studies may be used to compute alphalike reliability coefficients and investigate the effects of changing the number of items (or raters or occasions) on decreasing measurement error and increasing reliability. G theory, then, provided a melding of the psychological with the mathematical and produced a comprehensive conceptual framework and statistical model for identifying sources of measurement error. Coefficient alpha was then relegated to a minor corner of the much broader integrative theory.

Perhaps an example will make this comprehensive theory more concrete. In the early 1980s the newly legislated all-volunteer military force faced a recruiting challenge (Wigdor and Green, 1991). To insure the capacity of the military to fulfill its mission, more and more tax payers' money

was needed for recruiting the best and the brightest; the competition with higher education and the labor force was fierce. The military's request to Congress for recruiting dollars was based on strong validity findings: the Armed Forces Vocational and Aptitude Battery (ASVAB) did an excellent job of predicting recruits' training-school performance, and those scoring high were costly to recruit (e.g., due to incentives such as education benefits). However, Congress noted that the criterion was not job performance but school grades during training; no wonder ASVAB predicted well. Congress then launched an extensive study of military job performance measurement.

Performance assessments—hands-on measures of real-time job performance—were developed to be used as the criterion variable. These assessments were built by (randomly) sampling tasks and responses within a military occupational specialty. A sample of job incumbents performed a sample of job tasks under the watchful eye of sample judges who evaluated their performance. So, for example, machinist mates in the navy were observed by judges while working in a ship's engine room, mechanics in the air force were observed as they repaired a malfunctioning jet engine, and army or marine infantrymen were observed, for example, as they carried out sorties firing at enemy targets.

What was unknown at the time was whether judges (retired job incumbents) would be able to evaluate complex performance reliably. Generalizability theory provided the framework and statistical model for investigating the unknown. The measurement design—the performance assessment—focused on military personnel performance, the object of measurement. The design also contained multiple sources of measurement error, most notably task sampling and judge sampling. That is, the design was personnel × task × judge. Performance might vary over easy and difficult

job tasks, some judges might be lenient and others stingy, or some tasks might be difficult for some incumbents while other tasks difficult for different incumbents (personnel × task interaction).

Generalizability theory was used to estimate the magnitude of each source of variability in military performance scores: personnel (p), task (t), judge (j), and their interactions pt, pj, tj, and a residual (ptj,e). The G-study findings replicated across job performance measurements: judge sampling (e.g., judge disagreement) was negligible; task sampling was substantial. Moral of the story: save money and use one well-trained judge but increase the number of job tasks in the sample (see 1997 and Shavelson et al., 1999). Incidentally, the ASVAB general ability measure predicted job performance regardless of military occupational specialty, and about as well as it did training grades.

Cronbach's work on validity theory—the extent to which an interpretation of a test score is conceptually and empirically warranted—was no less significant than his work on reliability. As already noted, he reconceptualized the theory with Paul Meehl, placing construct validation at the center of psychological, educational, and social testing. Just as in other areas of science, for Cronbach validation was a process of theory building, testing, and improving. Validation, a never-ending process, examined a proposed test interpretation—a construct like self-concept or achievement—by testing the proposed interpretation logically and empirically against counter-interpretations. Moreover, what was validated, according to Cronbach, was not the test itself, for a test could be used for many purposes (e.g., prediction, diagnosis, placement). Rather what was validated was a proposed interpretation (1971). While reliability was an important characteristic of a test Cronbach believed that ultimately it served its master, validity, where sometimes tradeoffs were necessary between

broadly gauging a construct and narrowly constructing a homogeneous set of items to improve reliability: "Bandwidth, i.e., greater coverage, is purchased at the price of lowered fidelity or dependability. For any decision problem there is an optimum bandwidth. This conclusion departs from conventional theory, which assumes that it is always desirable to maximize dependability" (1957, p. 128).

Cronbach's work on individual differences and instructional methods with Snow focused on matching learning environments with students' aptitudes (1977). This research can be traced to early work with Goldine Gleser in the 1950s on personnel decision theory and on his presidential address at the American Psychological Association. In the personnel placement work he and Gleser concluded that optimal decisions about person-job matches must acknowledge the interaction of individual differences with job demands. Individuals with one profile of characteristics would be expected to perform well in one type of job while individuals with a second profile would be expected to perform well in another job with different task demands. (This was not found to be the case in the military job performance work described earlier and, as we will see, it did not pan out in the area of instruction.)

At about the time the Cronbach and Gleser book, *Psychological Tests and Personnel Decisions* (1957), went to press, Cronbach was preparing his presidential address for the American Psychological Association. The work on personnel theory gave him a fresh look at the schism in scientific psychology, one that formed the basis of the address, "Two Disciplines of Scientific Psychology" (1957). He called for a rapprochement between the "two disciplines," attempting to bridge the gap between the correlational studies characteristic of human abilities (individual differences) research that assumed these differences generalized across situations, and

the experimental studies that focused on differences between situations ("treatments") and viewed individual differences between people as noise (i.e., a source of error). In his own words: "Correlational psychology studies only variance among organisms; experimental psychology studies only variance among treatments. A united discipline will study both of these, but it will also be concerned with the otherwise neglected interactions between organismic and treatment variables. Our job is to invent constructs and to form a network of laws which permits prediction" (1975, pp. 681).

This integrated scientific approach would posit that individual differences might be highly predictive of performance in one type of instructional condition (situation) and much less so in another. It would seek matches between those treatment conditions under which individuals with a certain profile of characteristics flourished; different treatments would be needed for individuals with different profiles. Statistically speaking, an aptitude by treatment interaction was to be sought in the merging of two disciplines of scientific psychology—correlational and experimental. For example, Figure 1 shows an aptitude-treatment interaction in which the regression of outcome on aptitude differs under treatment conditions A and B. An assignment rule can be formed by setting up confidence intervals on the regression slopes. Assign low-aptitude students to treatment B, assign high-aptitude students to treatment A, and assign those falling in between (difference in slopes not statistically significant) to either treatment, perhaps the less costly one, all else being equal.

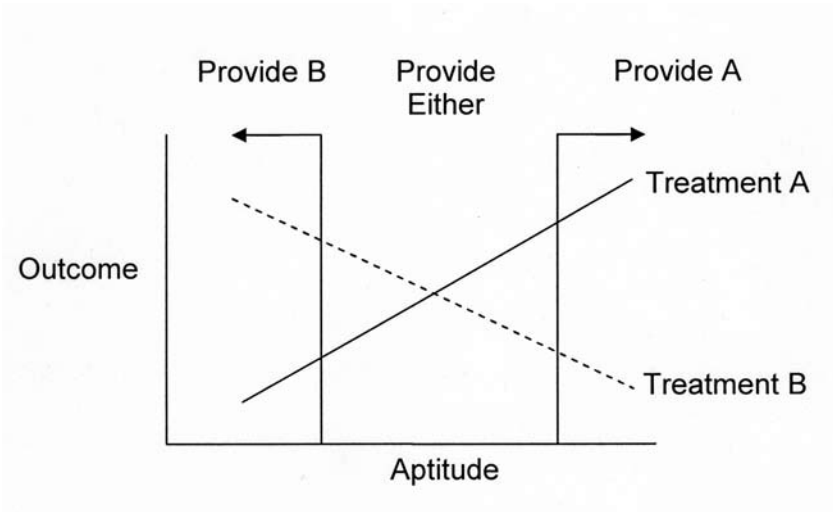


FIGURE 1 Regression of outcome on aptitude under treatment A and treatment B, showing an aptitude by treatment interaction such that those low on aptitude receive treatment B while those high on aptitude receive treatment A, and those in between receive either.

Cronbach's work on personnel decisions and his presidential address sent him and Snow on a 10-year trek in search of aptitude-treatment interactions (ATIs). They sought allocation rules like the one just described but sometimes more complicated: Assign people with one aptitude profile (high verbal ability, low spatial) to treatment A and people with the opposite profile to treatment B to optimize learning (1977). At the end of their search they concluded that "Almost no ATI effects were confirmed by multiple studies ....The evidence was negative. A spatial pretest, for example, may or may not predict outcomes from instruction filled with diagrams" (1989, p. 85).

The strongest ATIs involved general ability where students with above average intellectual development profited from instruction that provided them with considerable responsibility for organizing and interpreting while those below average

profited from a highly structured learning environment. Roughly 18 years later Cronbach revisited his presidential address in "Beyond the Two Disciplines of Scientific Psychology" (1975). He had come to recognize that ATIs were highly complex—as if human behavior involved myriad interactions acting on it simultaneously, like the reflections in a hall of mirrors, giving multiple views of an individual's behavior. Some ATIs were rapidly changing, and some context bound, far more than he had imagined earlier on. He concluded that "our troubles do not arise because human events are in principle unlawful; man and his creations are part of the natural world. The trouble, as I see it, is that we cannot store up generalizations and constructs for ultimate assembly into a network" (1975, p. 123, italics in original).

Cronbach's research went beyond measurement and instruction; he made significant contributions to the field of program evaluation. This line of work represented an enduring interest in the topic ever since he hooked up with Ralph Tyler in the Eight Year Study at the University of Chicago. He was skeptical of the sterile view of evaluation as a detached, objective, scientific activity with test content matched to curricula, appropriate experimental designs, and proper statistical tests. "[I] spoke up for the more constructive assignment of providing feedback on students' accomplishments and difficulties week by week, not delaying inquiry until the final examination; most of the conventional prescription was wrong for that kind of evaluation. I was, obviously, extending Tyler's view beyond the local teacher, using evaluation to reshape a course of study that would be widely distributed" (1989, p. 89).

In "Evaluation for Course Improvement" Cronbach (1963) defined evaluation broadly as the collection of information to make decisions about educational programs. These decisions might be made for course improvement, for planning

instruction for individual students, or for judging how good a program is. He argued that historically, systematic evaluation was introduced for the sake of course improvement. When evaluation is carried out for course improvement the focus is on its effects on students. For him, "*The greatest service evaluation can perform is to identify aspects of the course where revision is desirable*" (1963, p. 235; italics in original). Moreover, he argued that the comparative evaluation should not dominate plans for evaluation. He voiced his concern about a sole focus on this function. recognizing decision makers have to choose between courses:

> But formally designed experiments pitting one course against another are rarely definitive enough to justify their cost. Differences between average test scores resulting from different courses are usually small, relative to the wide differences among and within classes taking the same course. At best, an experiment never does more than compare the present version of one course with the present version of another. A major effort to bring the losing contender nearer to perfection would be very likely to reverse the verdict of the experiment (1963, p. 237).

Thirty-seven years later he voiced the same concern to me when I told him that I had agreed to chair the committee of the National Academy of Sciences that wrote *Scientific Research in Education* (Shavelson and Towne, 2002). He advised not to let the pressure for randomized trials overcome the reality of what they could produce under real-world conditions. For him a randomized trial was a case study highly dependent on local conditions and time.

His view of evaluation contrasted sharply with the views of others in the field. Donald Campbell and Thomas Cook extolled the virtues of the experimental-causal approach to program evaluation, a position prominent today with the federal No Child Left Behind legislation and the Institute of Education Sciences' What Works Clearinghouse. His view also contrasted with others in the field, such as Michael

Scriven's, who argued for two functions of evaluation, one formative such as improving a course as Cronbach advised, and the other summative. Summative evaluation focused on reaching a judgment on the merit and quality of a program; it often entailed comparisons, asking what alternative program produces the "biggest bang for the buck."

Cronbach turned to evaluation as his major focus as director of the Stanford Evaluation Consortium (1975-1979). The consortium, sponsored by the School of Education with Russell Sage Foundation support, was a research, service, and training group of faculty and graduate students, from the sociology, communications, and psychology departments, as well as education. The consortium served as a think tank, taking on evaluation projects as incubators for generating and testing ideas; the projects cut a broad swath for generating ideas and methods, ranging from education to health to juvenile delinquency programs. Cronbach and the consortium served as counterpoint to "scientific" approaches to evaluation with the "gold standard" being randomized, controlled field trials. They viewed evaluation as an art in which science and humanities both provided tools in crafting the design to the particular questions and situations driving the evaluation. Consequently, what is called "mixed methods" today were conceived and implemented by the consortium uniquely to fit each evaluation; a single method could not address all the pertinent questions raised in a program evaluation. The ideas and findings of the consortium were published in 1980 in *Toward Reform of Program Evaluation.*

Recognizing the consortium's publication reflected compromises inevitable when a group of scholars had to reach consensus, Cronbach (1982) set out his personal ideas on evaluation, noting that he did not believe they strayed too far from what his colleagues in the consortium believed. In *Designing Evaluations of Educational and Social Programs* we

see vintage Cronbach, harking back to the days of Ralph Tyler in 1938, who so influenced his writing on formative evaluation in 1963:

> Those who become investigators [evaluators] quickly learn that the formal, preplanned design is no more than a framework within which imaginative, catch-as-catch-can improvisation does the productive work. Even in basic research, nature does not stick to the script. Planned treatments go awry, and surprises lead the investigator down new paths. Questions posed to get the inquiry under way prove to be far less interesting than the questions that emerge as observations are made and puzzled over (1982).

His message was that "speaking of experiments and naturalistic case studies as polar opposites is a rhetorical device; evaluation planning is not a matter of choosing between irreconcilables" (1982, p. 44). He argued a kind of "let the punishment fit the crime" approach in that it makes sense to control some aspects of data collection in a naturalistic investigation but it also is important to use naturalistic observation even when the evaluation is rigorously controlled. "Experimental control is not incompatible with attention to qualitative information or subjective interpretation, nor is open-minded exploration incompatible with objectification of evidence" (1982, p. 44).

These two influential books that emerged from the Stanford Evaluation Consortium signaled the end of Cronbach's active university career; he retired from Stanford in 1980. It was fitting that his *Designing Educational Evaluations* was selected in 2000 as one of the top 100 education-related "Books of the [20th] Century" by the Museum of Education, University of South Carolina.

In the end then Cronbach found himself caught between science and practice whether in the classroom or in policy contexts. Science took him just so far, and he demanded science as far as it would take him. But he also recognized the contribution that other ways of knowing had to make

in understanding teaching and learning, and human action more generally. "The special task of the social scientist in each generation is to pin down the contemporary facts. Beyond that, he shares with the humanistic scholar and the artist in the effort to gain insight into cotemporary relationships, and to align the culture's view of man with present realities. To know man as he is is no mean aspiration" (1975, p. 126).

Cronbach's professional honors were numerous. He was president of the American Educational Research Association, the American Psychological Association, and the Psychometric Society, and a member of the National Academy of Sciences, the National Academy of Education, the American Philosophical Society and the American Academy of Arts and Sciences. He received many honorary degrees, including ones from Yeshiva University, the University of Gothenburg, and the University of Chicago. And he was honored by, for example, the Educational Testing Service for contributions to educational measurement, by the American Psychological Association for distinguished scientific contributions, by the American Psychological Society as a William James Fellow, by the American Educational Research Association for contributions to research in education, and by the Evaluation Research Society for contributions to evaluation methodology.

As I look back on my "travels with Cronbach" I realize just how fortunate I was that our paths crossed throughout our lives. He proved to be a caring and wise (if daunting at times) mentor. His ideas, intellect, and expectations set the bar about as high as it can get, and throughout my career (which is coming to the end as I retire in 2009) has guided my research, teaching, and mentoring. His research agenda also strongly influenced my work in psychometrics (especially G theory); in breaking new ground in assessing students' science achievement, undergraduates' learning, and enlistees' military job performance; and in the cognitive science

of working memory, intelligence, and school performance. I was fortunate to work closely with Cronbach up to about eight hours before he died (see Shavelson, 2004). All this said, I have found that either I disagree with him, or don't understand what he meant, or am afraid to believe in his conclusion that the best that the social, behavioral, and education sciences can achieve is "to pin down the contemporary facts." I continue to strive for and hope for a more enduring understanding of human cognition, affect, and conation, recognizing the complexities of human behavior in situation.

NOTE

1. Cronbach, along with Richard E. Snow, supervised my dissertation research. Upon graduating I joined the Stanford University School of Education faculty for three years, in part to teach Cronbach's test theory course while he was on sabbatical and when he returned we interacted regularly on measurement issues (1970-1973) before I moved to the University of California, Los Angeles. Cronbach and I continued our relationship over the years, and at the time of his death we had collaborated on his final paper, published posthumously, on 50 years after the publication of the "Cronbach Alpha" paper (2004; Shavelson, 2004).

REFERENCES

Goodenough, F. L. 1936. A critical note on the use of the term 'reliability' in mental measurement. *J. Educ. Psychol.* 27:173-178.

Shavelson, R. J. 2004. Editor's preface to Lee J. Cronbach's "My Current Thoughts on Coefficient Alpha and Successor Procedures." *Educ. Psychol. Meas.* 64(3):389-390.

Shavelson, R. J., and L. Towne, eds. 2002. *Scientific Research in Education*. Washington, D.C.: National Academy Press.

Shavelson, R. J., M. A. Ruiz-Primo, and E. W. Wiley. 1999. Note on sources of sampling variability in science performance assessments. *J. Educ. Meas.* 36(1):61-71.

Thorndike, R. L. (1947). *Research Problems and Techniques,* Report No. 3. AAF Aviation Psychology Program Research Reports. Washington, D.C.: U.S. Government Printing Office.

Thurstone, L. L., and E. J. Chave. 1929. *The Measurement of Attitude.* Chicago: University of Chicago Press.

Wigdor, A. K., and B. F. Green Jr., eds. 1991. *Job Performance Assessment for the Workplace*, vol. I. Washington, D.C.: National Academy Press

SELECTED BIOBLIOGRAPHY

1941

Individual differences in learning to reproduce forms: A study in attention. *Am. J. Psychol.* 54:197-222.

1949

Essentials of Psychological Testing. New York: Harper and Row.

1951

Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297-334.

1954

Educational Psychology. New York: Harcourt Brace and World.

1955

With N. L. Gage. Conceptual and methodological problems in interpersonal perception. *Psychol. Rev.* 62:411-422.

With P. E. Meehl. Construct validity in psychological tests. *Psychol. Bull.* 52:281-303.

1957

Two disciplines of scientific psychology. *Am. Psychol.* 12:671-648.

With G. C. Gleser. *Psychological Tests and Personnel Decisions*. Urbana: University of Illinois Press.

1958

Proposals leading to analytic treatment of social perception scores. In *Person Perception and Interpersonal Behavior*, eds. R. Tagiuri and L. Petrello, pp. 353-379. Stanford, Calif.: Stanford University Press.

1963

Course improvement through evaluation. *Teach. Coll. Rec.* 64:672-683.

1965

With P. Schonemann and D. McKie. Alpha coefficients for stratified-parallel tests. *Educ. Psychol. Meas.* 25(2):291-312,

With N. Rajaratnam and G. C. Gleser. Generalizabilty of stratified-parallel tests. *Psychometrika* 30:39-56.

With N. Rajaratnam and G. C. Gleser. Generalizability of scores influenced by multiple sources of variance. *Psychometrika* 30:395-418.

1969

Validation of educational measures. In Proceedings, 1969 Invitational *Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service.

1970

With L. Furby. How we should measure "change"—or should we? *Psychol. Bull.* 74:68-80.

1971

Test Validation. In *Educational Measurement*, 2nd ed., ed. R. L. Thorndike, pp. 443-507. Washington, D.C.: American Council on Education.

1972

With G. C. Gleser, H. Nanda, and N. Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

1975

Beyond the two disciplines of scientific psychology. *Am. Psychol.* 30:116-127.

1977

With R. E. Snow. *Aptitudes and Instructional Methods: The Search for Interactions.* New York: Irvington.

1980

With S. Ambron, S. Dornbusch, R. Hess, R. Hornik, D. Phillips, D. Walker, and S. Weiner. *Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements.* San Francisco: Jossey-Bass.

1982

With K. Shapiro. *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass.

1989

Lee J. Cronbach. In History of Psychology in Autobiography, vol. 8, ed. G. Lindzey, pp. 64-93. Stanford, Calif.: Stanford University Press.

1997

With R. L. Linn, R. L. Brennan, and E. H. Haertel. Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educ. Psychol. Meas.* 57(3):373-399.

2001

With L. Corno, H. Kupermintz, D. F. Lohman, E. B. Mandinach, A. W. Porteus, and J. E. Talbert. *Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow.* Mahwah, N.J.: Erlbaum.

2004

With R. J. Shavelson. My current thoughts on coefficient alpha and successor procedures. *Educ. Psychol. Meas.* 64(3):391-418.