Innovation and Intelligence Testing: The Curious Case of the Dog that Didn’t Bark

Robert J. Sternberg and James C. Kaufman
Yale University, USA

We argue that intelligence testing has been painfully slow to evolve, and that, arguably, it shows the slowest rate of evolution of any major technology. We consider a number of reasons for this glacial pace of evolution, and suggest what might be done to speed it up. The major obstacles to evolution seem to have been (a) the belief on the part of test developers that their tests are satisfactory as is, (b) the cost of developing new tests, (c) the validation of new tests against old ones, reinforcing the trend toward new tests that look like old ones, (d) failure fully to consider implicit theories of intelligence, and (e) test constructors who do test construction because they do well on and value existing instruments.

Keywords: Intelligence measurement, intelligence testing, intelligence theory

“Is there any point to which you would wish to draw my attention?”
“To the curious incident of the dog in the nighttime.”
“The dog did nothing in the nighttime.”
“That was the curious incident,” remarked Sherlock Holmes.


According to some theories of cognitive development, human beings are natural-born novelty-detectors. Infants prefer moderate novelty in stimulation (e.g., Kagan, Kearsley, & Zelazo, 1978; Kagan, Reznick, Clarke, Snidman, & Garcia-Collet, 1984), and there is even evidence that their preference for such novelty predicts later scores on standard psychometric tests of intelligence (Bornstein & Sigman, 1986; Fagan, 1984, 1985; Lewis & Brooks-Gunn, 1981). Adults, too, seem to be attuned to notice change, and when change stops occurring, they either habituate, as do infants, or show sensory adaptation, so that a stimulus that they first sensed stops being sensed (Helson, 1964). Evolutionarily, such sensitivity to change makes sense, because all organisms have had to evolve in a way that will permit them to adapt to often rapidly-changing environments. Those organisms that were insensitive to change presumably were quick bait for prey, and did not live to reproduce.

Because of our attunement to change, we often fail, like Dr. Watson, to notice when things do not change that presumably should. In the case of Watson, he failed to notice what should have been a reaction by a dog to an intruder. The ever-observant Holmes, in contrast, noticed that the oddity in the situation was that nothing happened. The thesis of this article is that a similar situation has arisen in an important area of psychology – intelligence testing.

Intelligence tests are clearly one of psychology’s most notable technological innovations. Since the turn of the century, they have been widely used for many different purposes, such as selection, diagnosis, and prediction (Anastasi, 1988; Cronbach, 1990). In the U.S., intelligence testing (and its offshoots) is a multi-billion-dollar business, and has a profound impact upon people’s lives from elementary school through postgraduate study.

Moreover, people involved in testing have been recognized as major contributors to psychology. According to a survey in Developmental Review, David Wechsler is among the top 10 most frequently cited psychologists in the field. Lee Cronbach, author of a leading U.S. textbook on assessment, is the most frequently cited author in the history of the major U.S. journal, Psychological Bulletin. Anne Anastasi, author of the other major textbook, won the National Medal of Science, the most prestigious award for a scientist in the U.S.

An interesting curiosity arises with regard to intelligence testing – how little it has changed over the years. The current version of the Stanford-Binet Intelligence Scale, the fourth edition (Thorndike, Hagen, & Sattler, 1986), differs in specific content, but little in concept from the original version of the
Binet-Simon scale first published in 1905 (see, e.g., Singer & Sattler, 1994). Similarly, the current versions of what are probably the most-widely used scales for measuring intelligence, the Wechsler series (e.g., the Wechsler Intelligence Scale for Children – Third Edition – Wechsler, 1991), also differ hardly at all from the original tests (see Wechsler, 1939). But should modern tests differ more from earlier ones?

The answer to this question of course depends on one’s point of view. Watson originally failed to see why the dog should have barked. But consider the development of other technologies, both inside and outside psychology.

Psychotherapy might be characterized as psychology’s other major applied contribution besides intelligence testing. Classical Freudian psychotherapy, which, like intelligence testing, dates back to the turn of the century, is still used. On the other hand, there have been many subsequent developments in psychotherapy – humanistic-existential, behavioral, cognitive, and eclectic, to name a few – and these innovations have been fairly widely adopted. For example, a survey of the orientations of 415 clinical psychologists in the U.S. (Smith, 1982) found that 12% reported themselves as primarily humanistic or existential, 6% as primarily behavioral, 12% as primarily cognitive, and 41% as primarily eclectic (combining orientations). Only 14% described themselves as primarily psychodynamic.

Consider what may be one of the most successful technological innovations of all times, computers. The first major electronic computer, the Automatic Sequence Controlled Calculator, also known as the Harvard Mark I, was first developed in 1944. It was 15 meters long and 2.4 meters high, and its operations were controlled by a sequence of instructions coded on punched paper tape. The second major machine, ENIAC, developed only two years later, weighed 30 tons and consumed 1500 square feet and was 1,000 times faster than its predecessor. By 1951 came UNIVAC, which was far more advanced than ENIAC in its use of vacuum tubes and memory components consisting of mercury delay lines. Today’s computers put the early ones to shame. For example, whereas ENIAC could execute up to 5,000 basic arithmetic operations per second, modern computers can easily execute 50,000,000 instructions per second and more. In terms of speed, then, they are at least 10,000 times more powerful than ENIAC, and the chips that are so much more powerful than the vacuum tubes of ENIAC are a tiny fraction the size.

Of course, one could consider other technologies: televisions, compact disk players, telephones, food processors, automobiles, and the like. The more technologies one considers, the more curious intelligence tests become, looking more and more like a dog that fails to bark, no matter what the provocation.

Our goal in this article is to reflect upon this strange state of affairs. We do not claim to have a final solution to the mystery, and we realize that there are those in the field, like Dr. Watson, who will see no mystery at all. But we believe there is one, and that psychologists need to understand why intelligence testing remains virtually alone among modern technologies in its slow pace of innovation, at least with regard to the content of the tests. We shall consider several possible reasons why, in our opinion, innovation has been so slow.

Suppose Spearman Got It Right

All technological innovations are based on underlying theory. Consider our two major examples beyond intelligence testing noted above. Classical psychodynamic therapy draws heavily on the theories of Sigmund Freud (e.g., Freud, 1949, 1963). These theories have, for many years, been under relentless attack by a number of theoreticians (e.g., Beck, 1976, 1986; Ellis, 1973; Rogers, 1959, 1961; Wolpe, 1958). The criticisms of Freud have been many, ranging from general nondisconfirmability of his theory to lack of scientific evidence for any aspects of the theory that might be disconfirmable to possible misinterpretation of his case studies. Whatever the criticisms, there are large numbers of clinical psychologists who do not view the available empirical data as providing much support for the Freudian point of view. At the same time, the data contrasting the effectiveness of various forms of psychotherapy remain equivocal (Stiles, Shapiro, & Elliott, 1986).

There have certainly been critiques of both the theory of general intelligence, or g (Spearman, 1927), on which the notion of an IQ is conceptually based, and on intelligence testing in particular (e.g., Feuerstein, 1979; Gardner, 1983; Piaget, 1972; Sternberg, 1985, 1990; Vygotsky, 1978). But the attacks differ in kind from those on Freudian theory. For one thing, the psychometric theory of g has spawned an enormous amount of data (see, e.g., reviews by Brody, 1993; Carroll, 1982; Jensen, 1980; Mackintosh, in press). For another, the data can be interpreted (although they are not always interpreted in this way) as supportive at a global level of Spearman’s concept of g. Thus, there may be a disanalogy with
Freud: What if Freud had gotten it right? Maybe Spearman did!

What, exactly, do we mean when we say the data have been supportive of Spearman’s major idea? We mean, basically, that when a large number of tests presumed to measure intelligence are given to various groups of individuals, factor-analytic methods tend to yield a general factor, or at least a general factor atop a hierarchy of less general abilities (Carroll, 1993; Cattell, 1971; Gustafsson, 1984, 1988; Horn, 1994; Jensen, 1980; Vernon, 1971). People can argue – and have – over the correct method for extracting factors, or over how to rotate factors. But even when the factors are rotated obliquely to yield primary mental abilities or similar such constructs (Thurstone, 1938), these primary mental abilities themselves tend to be correlated, and to yield a general factor when the factors themselves are factored.

One reason for the slow evolution of intelligence testing may be that Spearman had it right, except for the details. Arguably, psychometric theories after Spearman have really differed more in details than they have in fundamental postulates (Carroll, 1993; Sternberg, 1977, 1990). Most of them allow for a general factor, with group factors and specific factors nested under those factors. Those psychometric theories that do not allow for such a structure (e.g., Guilford, 1967) have not fared well empirically (see, e.g., Horn, 1967; Horn & Knapp, 1973).

Of course, there are non-psychometric theories that, in one way or another, discount \( g \). The forms of discounting have varied.

For example, Gardner (1983) has argued for a theory of multiple intelligences. According to Gardner, there are seven multiple intelligences – linguistic, logical-mathematical, spatial, bodily-kinesthetic, musical, interpersonal, and intrapersonal – each fundamentally independent of the others. Although the intelligences interact, functionally as well as anatomically, they are believed to be distinct. Gardner has suggested that the \( g \) factor results from the use of a narrow range of tests and testing formats, although he has not provided any concrete evidence in support of this assertion. The measures he has used (see Gardner, 1993) probably do not have the as-yet demonstrated psychometric properties (reliability, criterion validity) that would enable one fully adequately to test the claims.

Sternberg (1985a, b, 1988) has also questioned the \( g \) construct, arguing that psychometric \( g \) is a reality for the range of tests currently used on batteries to assess intelligence. But Sternberg, like Gardner, believes these batteries are too narrow in the abilities they tap. In his triarchic theory, Sternberg argues that intelligence comprises creative and practical as well as analytical abilities, but that intelligence tests measure almost exclusively the analytical abilities. Sternberg, Ferrari, Clinkenbeard, and Grigorenko (in press) tested these notions via a test based on the theory that was administered to 199 high school students. A multiple-choice test crossed sections measuring analytical, creative, and practical thinking, with items in three content domains – verbal, quantitative, and figural – for a total of \( 3 \times 3 = 9 \) multiple-choice subtests. Three essay tests were also used, one each to assess analytical, creative, and practical abilities. Sternberg and his colleagues found relatively low correlations among the sections (analytical, creative, practical), and a varimax-rotated factor analysis of the 12 subtests yielded only orthogonal specific factors with no underlying general factor.

Whether or not Spearman is viewed as “correct” with regard to \( g \) depends in large part on how one defines intelligence. Most theorists, including Gardner and Sternberg, accept a psychometric \( g \) for the current range of tests. The question then becomes whether these tests adequately represent intelligence. A number of theorists (see Sternberg, 1990, for a review) believe that they do not. But the majority of those using intelligence tests accept the psychometric paradigm, and thus something akin to the notion of \( g \). For them, therefore, there is no great incentive to use radically new tests, because most of the available tests provide roughly comparable measures of \( g \). This fact is sometimes referred to as the “indifference of the indicator” – no matter what test you use, you get more or less the same general factor (see Jensen, 1980).

**Tests and Markets**

A second reason for the relatively slow evolution of tests, we believe, is that a very substantial investment of time and money needs to go into the creation and development of an intelligence test, which leads publishers who are already risk-adverse to be even more so. The result is that publishers are likely to hesitate, for fear that they will pay a lot in development costs, and then will not recoup these costs.

Of course, it would be nice if national funding organizations were interested in funding research on the development of new tests. Our impression is that some national funding organizations have such an interest, for example, in Singapore. But such research is more likely to be applied than basic, and...
more likely to be for the government’s own tests than for ones that will become available commercially. In other countries, such as Japan, the concept of intelligence itself is not particularly well regarded, and is not terribly likely to generate government funding. Overall, then, government funding organizations, as well as private ones, may not place basic research that might ultimately lead to the development of intelligence tests at a high priority.

This situation is rather different from that in psychotherapy, where a new theory and therapy plan can be implemented with minimal or even no cost. The main difficulty is making the transition from a therapy plan to actual therapy, but in the past, this transition has been made with relative ease for a variety of different kinds of therapies – for example, cognitive therapy in the case of Aaron Beck, rational-emotive therapy in the case of Albert Ellis, client-centered therapy in the case of Carl Rogers, and behavior therapy in the case of Joseph Wolpe.

A further problem is that the test-buying market in many places is a relatively conservative one that has, in our opinion, not always been fully adequately trained. School and many clinical psychologists, for example, are only barely aware of new developments in thinking about intelligence and related constructs. Many of them have been trained to give conventional kinds of tests, such as the Wechsler, and actually feel threatened by the possibility that this training might become out of date. Of course, there are any number of other such psychologists actively looking for innovative products. But in order to recoup a major investment, a test publisher cannot afford only to sell to the “radicals,” and the vast majority of the market may be more or less content with what they have.

Even if they are not content, they are buying. And ultimately, what forces change is not contentment so much as competition. As long as test publishers can continue to show major profit from conventional products, no matter what they are, they have little incentive to promote innovation. On the contrary, the incentive is actually not to promote such products, because they could cannibalize the publishers’ own markets without bringing in any substantial new revenue.

In our own experience, developing items for more or less standard kinds of psychometric tests is challenging, but not beyond the pale. Developing novel kinds of item formats, however, is very difficult, and many of them do not actually work when the psychometric data come in. As a result, developing really novel tests is a time-consuming and often frustrating business. Because, in test development, time means money, the successive iterations needed to get a test to work add to costs, and make the prospect of developing new kinds of tests less appealing.

### The Gold Standard, but is it Really Gold?

For many years, the United States and other countries as well pegged the values of their monetary currencies to gold. Gold, then, was the standard by which the value of a currency was judged. Many countries, including the United States, eventually stopped pegging the value of their currencies to gold because they discovered that gold, like any other standard of value, itself is of only debatable value.

In the psychometric testing of intelligence, there is a gold standard, but in psychometrics, unlike in economics, that standard is still quite closely adhered to. The standard is prior tests of intelligence, most notably the Wechsler and the Stanford-Binet. A good (i.e., externally, convergently valid) test is one that correlates highly – preferably as highly as possible – with these and similar tests. A test that does not correlate with these past tests is in trouble. The test is often viewed as measuring something other than intelligence. Even if the test is reliable, it may simply be dismissed as measuring something reliably, but not intelligence, per se.

Of course, we would not want test publishers to stop publishing validity information, or test users to stop using it. At the same time, though, this particular validation procedure results in a built-in conservatism in the evolution of tests that is, from certain points of view, undesirable. There are many perspectives on intelligence (see, e.g., Gardner, Kornhaber, & Wake, 1996; Sternberg, 1982, 1990, 1994), of which the psychometric perspective is only one. Moreover, intelligence tests such as the Stanford-Binet and the Wechsler do not even fully represent the psychometric perspective. For example, they certainly do not cover all the abilities in the theories of Carroll (1993), Cattell (1971), or Guilford (1967).

The effect of using past tests as a primary means of validating new tests should not be underestimated. It results in new tests that are likely to be, at some level, clones of old tests. It is as though a first-generation computer were created, and then other computers were created that were clones, not only of the good features of that first-generation computer, but of the not so good features as well.

There is another standard that is used in the development of new tests, but really, it is a secondary
standard. This standard is school performance. The second most frequent kind of validation data is prediction of children's school performance. Such a criterion has the advantage that it is truly an external measure, and of something that most societies value. The main disadvantages are obvious. First, school grades are sometimes not very reliable. Second, they are contaminated by a number of variables, such as children's motivation and language background. Certainly they involve much more than intelligence. Third, it is difficult to compare grades across schools or even classrooms, because populations of students and grading practices differ so much.

There is a more insidious problem with grades, however. This problem is that they represent a strictly academic criterion for the validation of the tests. At one level, such a criterion makes sense. Getting good grades in school serves as one of the main potential accomplishments of children of school age. Moreover, the first serious intelligence test, that of Binet and Simon, was designed for strictly academic purposes – to separate out children who were truly mentally retarded from those who were not. Thus, it makes good sense that we would continue to validate tests in terms of academic performance. But here is a downside.

IQ versus People’s Conceptions of Intelligence

The downside of this procedure is that it puts a premium on the academic aspect of intelligence. But our research and that of others shows that there are nonacademic, or practical aspects of intelligence that are important for adaptation to everyday life and that apparently do not much correlate with IQ (see, e.g., Ceci & Liker, 1986; Dorner & Reither, 1978; Frensch & Funke, 1995; Sternberg & Wagner, 1986; Sternberg, Wagner, Williams, & Horvath, 1995), and probably don’t correlate much with school grades either. These aspects of intelligence will tend systematically to be excluded from tests if the tests are validated against academic criteria.

One might argue, of course, that these aspects of practical intelligence are something other than intelligence. For example, Herrnstein and Murray (1994) claim, as did Boring (1923) more than 70 years before them, that intelligence measured as IQ pretty well represents what we mean by intelligence. If this were true, then whatever it is that these practically oriented tests measure, it is not intelligence.

The evidence suggests, however, that IQ does not completely capture people’s conceptions of intelligence, among laypeople in the United States (Sternberg, Conway, Ketron, & Bernstein, 1981), professionals in the United States (Sternberg, 1985 b), nor among people in other countries either (Berry, 1974; Fry, 1984; Serpell, 1974; Sinha, 1983). For example, in our studies in the U.S. of people’s conceptions of intelligence, social- and practical-competence aspects of intelligence have proven to be an important part of people’s conceptions of intelligence. In some data that we have just analyzed (Yang & Sternberg, in preparation), we have surveyed Taiwanese conceptions of intelligence, and these conceptions are even more far afield from the conception of IQ than are the American conceptions. Elements akin to wisdom, for example, seem to play much more of a role. Moreover, mere translations of English-language tests probably are not entirely adequate for measuring the intelligence of people from diverse cultures, in any case.

Of course, there is no guarantee that people’s conceptions of intelligence are correct. Everyone could be wrong, as they were about phlogiston’s role in generating fire. But intelligence, unlike fire, has subjective elements built into it: Whereas fire is the same everywhere, what is adaptive to the environment is not. So people’s conceptions of intelligence need to be taken seriously. They are not and apparently have not been in test development, and as a result, the evolution of tests has been slowed.

Nonrandom Samples of Test Constructors

When we standardize tests, we are very concerned that we have random (or properly stratified) samples on which to base our data. But we do not think much, and usually not at all, about random or stratified samples of testers, and certainly not of test constructors. Yet, people who would go to work for test-construction companies are not a random sample of people, in general. Just as people who become police officers are likely to believe in the value of what the police do, or people who become medical doctors are likely to believe in the value of medicine, so people who become test-developers are likely to believe in the value of tests, and for now, that means conventional tests.

We know from research that similarity is one of the best predictors of attraction (Berscheid & Walster, 1978; Sternberg & Barnes, 1988). This principle applies in any human relationship, not just romantic
ones. When companies hire people, they look for those whose norms and values are likely to be compatible with their own. People who interview at a company that produces tests are likely to believe in the product; because if they don’t, their chances of being hired are presumably diminished. The same general principle would apply anywhere: We would be unlikely to hire at Yale, for example, a candidate who was anti-academic, or anti-Yale! The result of selective entrance of people into companies that produce tests is that those who are likely to apply for positions, get positions, and then be advanced to higher positions are those whose values are most compatible with those of the organizations. In all of the major organizations and probably many of the not so major ones, a likely outcome is the hiring and promotion of people who basically are satisfied or nearly satisfied with the kinds of products now being produced. This is not to say that such people would not consider innovations or even produce them, but it is to say that they are probably not the prime candidates for doing so. And many of them might actually resist evolutionary, not to mention revolutionary changes.

Conclusion

We have argued that the development of intelligence testing, as a technology, has moved at a snail’s pace. We have considered some reasons why this might be so. Doubtless there are other reasons, and moreover, many individuals might disagree with our reasons. But the facts would be difficult to controvert. What is more surprising than the slow pace of technological development, in our opinion, is the lack of recognition on the part of people in the field that there is even a problem – that the pace of technological innovation in this particular area is slower than that in any area at all of which we are aware. Perhaps it is because of the tests’ great success; but, of course, those who produced the earliest generations of automobiles, radios, televisions, and telephones also touted their products’ successes. The difference is that, in these other technologies, innovation proceeded at a rapid clip.

In our view, the most pressing question at this point is that of what is to be done. In our view, the future course of action is as simple as it is obvious. Testing companies, like many other companies, need to start thinking for the long-term rather than the short-term. And long-term thinking means support of basic research, something these companies have been loathe to do, with a few exceptions. This viewpoint would require a fundamental reorientation of their perspective. The truth is that even companies that have minimally supported basic research have had less than broad definitions of what is “basic,” and have not always been fully hospitable to projects or results that failed to support the utility of their products.

People’s lives are at stake. Tests are used to select people, to sort them once they are selected, to diagnose them, and to predict their future. Ghiselli (1966) pointed out that the validities of conventional occupational aptitude tests had not increased substantially in 30 years, and embarrassingly, they still haven’t improved another 30 years later! Although levels of these validities can be explained by a variety of factors, such as attenuation and restriction of range (Hunter & Schmidt, 1990), the failure of these validities to rise is nevertheless disappointing. Clearly, we have reached a point of diminishing returns with conventional tests. Sadly, the response of some testing organizations has been to emphasize only delivery vehicles (e.g., the use of computers) or scoring procedures (e.g., the use of latent-trait analysis), rather than the fundamental psychological presuppositions on which the tests are based. Developments in administration and scoring are important, but they are not a substitute for developments in the fundamental psychological theories underlying the tests.

There have been some notable developments. Kaufman and Kaufman (e.g., 1983) have published a series of tests that are theory-based, and a theory-based test by Das and Naglieri (in press) is about to be published. Feuerstein’s (1979) Learning Potential Assessment Device represents another lead, as does other work on dynamic testing (e.g., Brown & French, 1979). And there have been many exciting developments in the interface between cognitive theory and psychometrics (e.g., Embretson, 1985; Embretson & OeBoeck, 1994; Frederiksen, Mislevy & Béjar, 1993; Nichols, Chipman, & Brennan, 1995). But until we reorient ourselves to value substantive innovation in actual tests, we will probably remain in last place among the major technologies in rate of innovation.

Author’s Note

The work reported herein was supported under the Javits Act program (Grant #R206R50001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The
findings and opinions expressed in this report do not reflect the positions or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Author’s Address:
Robert J. Sternberg
Department of Psychology
Yale University
Box 208205
New Haven, CT 06520–8205
USA

References


