Sage Research Methods

The Uses and Misuses of Data and Models: The Mathematization of the Human Sciences

For the most optimal reading experience we recommend using our website. A free-to-view version of this content is available by clicking on this link, which includes an easy-to-navigate-and-search-entry, and may also include videos, embedded datasets, downloadable datasets, interactive questions, audio content, and downloadable tables and resources.

Author: W. James Bradley, Kurt C. Schaefer
Pub. Date: 2013
Product: Sage Research Methods
DOI: https://doi.org/10.4135/9781483348872
Methods: Exploratory data analysis, Measurement, Theory
Keywords: social science, intelligence, indicators, scale, social class, students, proxy, quality of life, organizations, decision making, persons, poverty
Disciplines: Anthropology, Business and Management, Criminology and Criminal Justice, Communication and Media Studies, Counseling and Psychotherapy, Economics, Education, Geography, Health, Marketing, Nursing, Political Science and International Relations, Psychology, Social Policy and Public Policy, Social Work, Sociology, Mathematics
Access Date: January 11, 2024
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Online ISBN: 9781483348872

Limitations of Measurement in the Social Sciences

In Chapter 5 we discussed the potential benefits that can be derived from a system of measurement. We also discussed the requirements a measurement must meet to provide these benefits and showed several examples of effective measurements. This chapter begins by examining the dark side of measurement. We will parallel our discussion of benefits from good measurement with a discussion of possible harm from mismeasurement. We follow this with a reexamination of the four requirements for measurement to see what limitations they place on us in collecting human information if we are to avoid such harm. This leads to a statement of several contextual norms. And we will conclude with several case studies, including a brief evaluation of the most controversial human measurement in our times—the measurement of intelligence.

Possible Social Harms from Mismeasurement

The first benefit discussed in Chapter 5 was predictive power. If Ed Thompson, our friend from the beginning of Chapter 5 with the weak back, sees a bag of salt labeled 40 pounds, he will feel safe in lifting it. But if the bag actually is a mislabeled 80-pound bag, he may injure himself. Incorrect or misleading measurements will result in wrong predictions. The consequences of such measurements can range from simple embarrassment to major harm.

For instance, standardized tests are often used to predict the potential academic success of young people. If such a test is inaccurate or misleading, a young person's ability to develop his or her gifts may be severely limited and society as a whole may be deprived of that person's contribution. Alternatively, suppose a charitable organization receives many proposals for projects and must select a relatively small number for funding. It develops a measurement scheme that enables it to compare projects. The purpose of such a measure is clearly predictive—to identify those projects that are going to bring the most benefit for the investment made in them. But if the measure is inadequate, poor selections may be made.

Our second benefit was that measurement enables us to transcend our subjectivity. But a misleading or inappropriate measure can fool us; it may result in misplaced and inappropriate decisions. Oftentimes these negative outcomes result from the illusion of precision and objectivity that a numerical measure provides.

For instance, an economic development organization may measure income production and use it as a basis

for decision making, while ignoring social disruption. Alternatively, suppose a prominent ingredient in some common food, A, is believed to be dangerous. Large numbers of people may choose to replace item A in their diet with item B, which has less of the ingredient. But item B may have another ingredient that subsequently is discovered to be equally dangerous. That is, “dangerous” has been inappropriately measured. As a third example, executives in a company may make decisions based on the “bottom line” while paying insufficient attention to other matters that are also critical to its success, such as employee and customer satisfaction and good will. The executives have mismeasured “success.” In each of these three cases, a precise, unambiguous, measurement—income produced, quantity of an ingredient, profit—was used as a proxy for a more general, more ambiguous concept—well-being, danger, or success. But it mismeasured the more general concept and thereby contributed to a potential for harm.

Poor measurement can lead to an impoverishment of understanding. If I learn that Billy's IQ score is higher than Bobby's, I may come away from that “learning experience” knowing less about Billy and Bobby than I did before. I may have forgotten my previous recognition that Bobby's interpersonal, intrapersonal, and artistic abilities are well above Billy's. But perhaps more seriously, I may have lost the recognition that both Billy and Bobby are complex persons whom I can never know completely, and I may have lost the ambiguity that should characterize my perception of them. And I may be unaware that Bobby was awake with a family crisis all night before he took the IQ test.

Our third benefit of measurement was that it provides a mechanism for identifying error. An adequate theory of error recognition has proven difficult for scientists and philosophers of science to develop. If two people take a meter stick and make the same measurement, they are very unlikely to obtain the same result. They immediately attribute the result to measurement error, rather than claim that the length is actually different for the two of them. But such an agreement is often lacking in areas that social scientists want to study. What makes agreement on methods for measuring length possible? The best answer available now seems to be that, intuitively, the two meter stick users share an underlying theory of length, guaranteeing that the mapping of objects into numbers called lengths is unique, given a particular unit (here, the meter). That is, they firmly believe that length is a property of the object such that differences in measurement have to result from the measurement process not from the object nor the measuring instrument. But measurements performed in the absence of such a theory and by techniques that have no substantive theoretical base are incapable of generating such a belief. Hence, differences in measurement may lead not to doubt about the accuracy of the particular act of measurement but to doubt about the measuring instrument or even to doubt about the uniqueness of the quantity being measured. If enough people develop enough doubt, this can lead to wide-

spread rejection of the measurement technique, to the loss of whatever benefits it might have obtained, and even to a reduction of credibility for the entire scientific enterprise.

Our fourth benefit of measurement was that it enables social justice, at least in the sense of equity. But mismeasures can cause serious inequities. Such inequities are particularly troubling because of the enormous social benefit that can be derived from accurate measures. Consider, for example, the “streaming” of children after elementary school done in Japan, France, China, and other countries. An accurate identification of a child's gifts can be enormously beneficial to the child and the culture by providing him or her an education suitable to those gifts. But misidentification can have lifelong negative consequences. Another critical justice issue associated with measurement is that measurements of people are internalized by those measured and by others who know the result. Inaccurate or misleading measures may result in serious misconceptions. We will return to this point later in this chapter.

Another benefit of effective measurement was that it raises people's confidence in the truthfulness of information being communicated. If measurement of a human quality is poorly done and leads to inconsistencies or doubt, the credibility of science in general, but more likely the social sciences in particular, can be undermined. This harms the general advance of scholarship and its dissemination by reducing people's confidence in its truthfulness.

Our last benefit was that measurement enables finer distinctions than those possible with natural language. Such distinctions are meaningful and often very helpful when the quantity can be measured on a ratio scale as with temperature, length, or time. But when a ratio scale is not appropriate, making such distinctions can be misleading. A familiar example of this is student grade point averages. That is, student grades are awarded on an ordinal scale and represent an approximate measure of a student's knowledge. These ordinal numbers are then treated as if they were cardinal numbers and averaged. But there is no significant difference between GPAs of 3.44 and 3.45, for instance.

In summary, then, there are significant potential social harms in mismeasurement.[1] We wonder how much of the current respect for the scientific study of the physical world, but suspicion of the scientific study of social phenomena, arises from the experience of some of these harmful effects. In the next two sections, we will examine the four requirements for measurement to see how these harmful effects can arise.

Limitations Associated with Definitions

Human phenomena are complex. Consider such important concepts as religiosity, prejudice, social class, poverty, and good will. Some (like good will) have proven so complex that decision makers who speak of them find it difficult to even attempt a definition. Yet precise definition is an essential prerequisite to measurement. One is tempted to reject measurement of such concepts altogether, but the potential benefits of effective measurement are so great that social scientists have not given up. Before we can assess the way that the social sciences have typically addressed this problem, we need to clarify the notion of definition itself.[2]

First, we can distinguish between univocal, equivocal, and analogous uses of words. *Univocal* uses of words are ones for which there is a one-to-one correspondence between the word and its meaning. Few words in ordinary language are used univocally, but scientific and mathematical words are typically used univocally and when they are not, their (unique) meaning ought to be clear from their context. Words are used *equivocally* when they have multiple meanings that may be quite unrelated. For instance, consider the word “can” as used in the sentences “I can do it” and “The beans were preserved in a tin can.” *Analogous* uses of a word occur when the same word refers to similar but distinct meanings. For instance, the uses of the word “religious” in “He is a religious man” and “That is a religious hymn” are analogous. The problem social scientists face in producing fruitful definitions is that many ideas they want to address are associated with words that are part of natural language and may have equivocal and analogous uses. But to be used scientifically, such words must be given univocal meanings. Social scientists have two choices: create new words or create special, univocal meanings for old ones. Both approaches are widely used although both are problematic. The introduction of new scientific words restricts understanding of the scientific work to those who have been initiated into its vocabulary. The redefining (or “precising”) of old words may lead to misunderstanding and confusion as people retain many of the older connotations of a word in the new, scientific, setting.

Besides these three different uses of words, we need to distinguish three different types of definitions. Once we have these concepts in mind, we will be able to see how the concept of definition is typically used in the social sciences and thus lay a foundation for a critique of measurement in these sciences.

- A *lexical definition* is a dictionary definition—a list of all popular usages of a word.
- An *essential definition* is a definition that has somehow fully captured the “essence” of an idea. Mathematicians, for example, spent many years in the 19th century debating the proper definition of

the word “function” before a definition was finally agreed on.

- An *operational definition* specifies a repeatable procedure by which one can determine whether a given object or event is an instance of a concept being defined.[3]

Mathematical and scientific definitions are typically operational. Mathematical definitions of abstractions aim to be essential and operational. Nevertheless, for social phenomena, definitions that are essential and operational are rare except in disciplines like mathematical economics and management science. The absence of definitions that are operational and essential helps to explain why there is a division in the social sciences between two different approaches: Some scholars choose operational definitions when definitions that are both operational and essential are not available. This enables them to measure and carry out statistical analyses. Other scholars regard such an approach as oversimplifying and misrepresenting. They seek essential definitions even when such definitions cannot be operationalized. Unless definitions that are simultaneously operational and essential can be found, this difference in approaches is inevitable. In fact, both approaches are necessary.

Scientific and mathematical definitions are closely related to models, as discussed in Chapter 2. That is, when used to describe some real-world phenomenon, they involve simplification and idealization and then formalization of a notion that (one hopes) captures some essential features of a characteristic of interest. Thus, they have the same limitations as models: the price one pays for simplicity, clarity, precision, and tractability is that the concept being defined is removed from context and, by being simplified, is at least partially misrepresented. In a sense then, it is impossible for scientific language to ever represent the truth about social phenomena. In fact, it seems to us that natural language shares the same limitation.

Our focus for the rest of this section will be on operational definitions inasmuch as their formulation typically precedes the production of human information. Operational definitions of real-world entities are positivistic, that is, they are expressed in terms of what can be concretely observed. “Theoretical terms”—terms for things that cannot be directly observed—are allowed, but they must be definable in terms of other things that are directly observable.

Operational definitions are common and familiar. In Chapter 5 we discussed the use of blood alcohol level as a proxy for drunkenness. Defining “drunkenness” as “having a blood alcohol level above .10%, as measured by a breathalyzer” is an operational definition with which most people would feel intuitively comfortable,

although they might argue about the choice of .10% as the critical value. Defining social class by various levels of annual income also would be an operational definition but one that would produce considerably less comfort. Defining a city's quality of life is a subtle enterprise, so for the sake of ranking cities, it is operationally defined. Typically such a definition involves a combination of such factors as health, crime, the economy, housing costs, education, transportation, weather, leisure, and the arts. Each of these is operationally defined in a way that yields a numerical measure. These measures are combined into a summary measure that forms the operational definition of "quality of life." Another example of operational definition is ranking of colleges. The magazine *US News and World Report* publishes an annual listing of "best" colleges and universities in various categories. It defines quality as a combination of academic reputation, student selectivity, faculty resources, financial resources, graduation rate, and alumni satisfaction. Each of these factors is measured numerically, some as a composite of several subfactors.

We can gain insight into some difficulties that have arisen in the use of operational definition by examining the history of the notion. The concept of operational definition was introduced in 1927 by P.W. Bridgman, a Harvard University experimental physicist, to address problems in the foundations of physics introduced by Einstein's theory of relativity and by the discovery of quantum mechanics.[4] Discussing length, Bridgman says,

> To find the length of an object, we have to perform certain physical operations. The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined. In general, we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations.* (p. 5, italics in original)

Thus, applying Bridgman's approach, length is that property of an object that is *defined* by the *act of measuring* it (whether by a meter stick, calipers, or other means).

Bridgman's concept of operational definition has a major weakness, however: It offers no explanation for why there should exist any meaningful correspondence between lengths (so measured) and the real number system. The credibility of the measurement of length as a real number arises from the fact that there is a natural correspondence between the properties of length and the real numbers, that is, lengths combine in the same way that real numbers do and people intuitively recognize this correspondence without needing a formal theory to demonstrate it. In other words, the theory exists whether we formalize it or not and it is the consistency

between the theory and the measurement that makes length a socially acceptable and hence useful measurement. It is not sufficient to define length (or any other characteristic) simply by the measurement technique without showing how the technique relates to properties of the characteristic.

We can now see why the use of operational definition has at times caused significant mischief in the social sciences. For instance, terms, such as “intelligence” and “customer satisfaction,” can be operationally defined as scores on a test. These terms already have considerable meaning in ordinary language, so people who are not familiar with the technical aspects of the operational definitions assume that their lexical meanings have been made more precise by the operational definition. In fact, this may not be the case. Bridgman himself recognized this difficulty and objected to the use of the term “operational definition” because it claimed too much. Because of this confusion we, along with some other scientists, prefer to replace the term “operational definition” with the more modest term “indicator,” as in “quality of life indicator,” whenever possible.

A fundamental and extremely important methodological problem underlies this difficulty. Social phenomena are complex and often poorly understood. They are embedded in complex contexts. Thus, understanding requires simplification and idealization and therefore all precise definitions of social phenomena require the use of proxies. Furthermore, to obtain replicable results, operational definitions must be used. Thus, a circularity arises—scientifically defining a concept requires a measurable proxy (so measurement must precede definition), but measurement requires that the concept first be unambiguously defined (so definition must precede measurement).

There is no easy resolution to this dilemma; scientists necessarily work in an artificial world that never quite matches the reality they are trying to study. Their work involves a continual tension.

- On the one hand, the entity one is trying to define is complex and embedded in a complex context. Also, any scientific theory demands unambiguous definition and intersubjective agreement. So operationally defined proxies are necessary.
- On the other hand, operational definitions can be constructed and proxies selected regardless of whether the resulting measurements correspond to the essential character of the entity being measured. Once proxies are selected and an entity is defined operationally, formal analysis can proceed and may continue to some length. But the results obtained may have little or nothing to do with the original concept one was trying to define.

Given this fundamental difficulty, about all social scientists who need to use measurement can do is to develop

indicators that capture the essential qualities of a concept as best they can, listen carefully to critics who call attention to the limitations of these indicators, and strive continually to improve the indicators and to understand the essential qualities of the concept. And we all need to continually remind ourselves that we are not actually measuring characteristics like intelligence, customer satisfaction, and quality of life but only proxies for them. That is, in social science our measurements, no matter how precise they may seem, are only as good as our selection of a proxy.

Limitations Associated with Theories, Rankings, and Methods of Measurement

We now look at the three remaining requirements for measurement and see what limitations they place on measurement of human information.

The “theory” requirement is that we have an understanding of the properties of the characteristic of interest that justifies correspondence with our measurement scale. As we saw in the last section, social entities are so complex that normally they are measured via proxies. Thus, the principal danger in the area of theory is that our proxy may satisfy the requirements for measurement whereas the characteristic we actually want to study does not. In this sense the proxy lulls us to sleep and we can begin to confuse the proxy with the actual characteristic.

For example, consider poverty, a complex phenomenon involving many social, cultural, and individual factors. These are difficult to measure so we use a proxy, perhaps gross income. Poverty then is defined as being below a specified income level called a “poverty threshold” or “poverty line.” This proxy is indeed a helpful indicator. But if the indicator of poverty is confused with poverty itself, we can be fooled into believing that an intervention to raise income is sufficient to lift someone out of poverty. Note that such an intervention may be helpful—the point is that it is not sufficient because poverty involves social and cultural factors as well as income. Similar considerations apply to two widely used economic indicators: gross national product (GNP) and gross domestic product (GDP). Both are used as proxies for economic well-being, a subtle and poorly defined concept. For example, GNP operationalizes economic well-being as the market value of final sales of products. Thus, GNP is precise and unambiguous. Popular discussions typically refer only to GNP or GDP, not to economic well-being. That is, the proxies have largely replaced the entity—economic well-being—for which they are proxies. But some aspects of economies are not included in GNP or GDP, such as black mar-

ket activity. Thus, we occasionally hear highly misleading reports, such as the news that the GNP of the former Soviet Union declined 50% per year for three successive years (1991 to 1993) and that it fell 25% in the first quarter of 1994 alone. In fact, much economic activity simply was not reported.

Next we consider limitations in the area of intersubjective ranking. In measuring weight, there is little difficulty with obtaining intersubjective agreement on ranking. If two people each lift two different weights, they will have little difficulty agreeing which is heavier unless the weights are very close. But the social sciences consider human qualities like intelligence, knowledge, customer satisfaction, quality of life, social class, and religiosity. These qualities are so complex that there is little likelihood of intersubjective agreement about which subjects have more of them. As with theory, this difficulty is usually addressed via the use of proxies. For instance, religiosity may be measured by frequency of attendance at a meeting place, such as a church, synagogue, or mosque. But note the same kind of problem we saw with theory: Agreement on a ranking of the proxy may give the illusion of agreement on a ranking of religiosity itself. But such agreement may be short-lived. For example, some classical work on the nature of prejudice found a strong positive correlation between religiosity as measured above and prejudice.[5] Yet subsequent study showed that when intrinsic and extrinsic motives for attending were distinguished, the high correlation only persisted with the extrinsically motivated attenders and not the intrinsically motivated attenders, though the average frequency of attendance was greater for the intrinsically motivated attenders. Thus, the proxy proved inadequate: whereas intersubjective agreement on a ranking of the proxy was easily obtainable (it was measured as a real number), a different proxy gave a different ranking. And a consensus was not immediately forthcoming regarding which best represented religiosity.

The need for a replicable means to measure also places limitations. Here are a few.

- Sometimes, even when definitions are clear and proxies available, measurement cannot be made. For instance, measurement may be too dangerous, subjects simply may refuse to be interviewed or may insist that the interviewer be an advocate for their concerns before confiding in the interviewer, or collection of data may be more expensive than the researcher can afford.
- Measurement techniques themselves may have limitations that simply must be acknowledged. Some measurements like blood alcohol level require the use of equipment that can be inaccurate. Some measurements require sampling and so only can be expressed by confidence intervals or by estimates. For example, consider cost of living indices that typically estimate cost of living by pricing items in an artificial “market basket,” or student knowledge that is measured by test questions that sample a student's knowledge.

▪ Data may be self-reported and therefore inaccurate—subjects may lie or be self-deceived. A classic study of this phenomenon was published by Richard T. LaPiere in 1934.[6] In this study LaPiere described his experiences traveling from 1930 to 1933 with a young Chinese couple. During their travels LaPiere and his Chinese associates approached 67 hotels, auto camps, and tourist homes for accommodation and ate in 184 restaurants or cafes. As much as possible, LaPiere allowed his Chinese guests to enter separately from him. LaPiere recorded detailed observations of the responses of clerks, waiters, and others to himself and his companions. They were only denied service once because of the ethnicity of his companions and they met 11 incidents of hesitancy or something they found temporarily embarrassing. On the other hand, on 97 occasions, LaPiere was treated better than he expected to be treated had he been alone. In the remaining 142 cases, they were treated as expected or treated well but with curiosity. In the second part of his study LaPiere mailed questionnaires to all of the establishments they had visited and asked “Will you accept members of the Chinese race as guests in your establishment?” He obtained a response rate of 51%. Over 90% of the respondents said they would not serve Chinese individuals! LaPiere's conclusion in the same paper was that “it is impossible to make direct comparisons between the reactions secured through questionnaires and from actual experience.” He concluded his article with this warning

> The questionnaire is cheap, easy, and mechanical. The study of human behavior is time-consuming, intellectually fatiguing, and depends for its success on the ability of the investigator. The former method gives quantitative results, the latter mainly qualitative. … Yet it would seem far more worthwhile to make a shrewd guess regarding that which is essential than to measure that which is likely to prove quite irrelevant. (p. 237)

The normative principle that arises from LaPiere's work is obvious: Questions that ask people what they “would do if” should be avoided.

▪ With questions involving categorical or ordinal data, the usefulness of the data depends on the clarity of the categories. For instance, a physical therapist once asked one of the authors of this book to rate the pain he was experiencing on a scale from 0 to 10. He mentally compared his pain to that likely when being crushed to death in an automobile accident and rated his pain a “3.” As he found out later, the physical therapist interpreted a 3 as meaning very minor pain and responded by pursuing an inappropriate mode of treatment. This problem (and others like it) easily could be avoided by giving a patient a written list of 10 carefully described scenarios of increasing pain and asking which comes closest to fitting his or her situation. In this way, a shared meaning for the measure would be

more likely.

Limitations of Ordinal Scales

One common misuse of measurement that does not arise directly from the four requirements, but should be mentioned, is the inappropriate use of arithmetic operations. As ordinal scales become more common (for good reason—there are many characteristics in the social sciences that do not lend themselves to interval, ratio, or absolute measure), it is frequently forgotten that arithmetic is not meaningful on these scales. For instance, many colleges and universities use a system of faculty evaluation in which students rank a faculty member on an ordinal scale. Typically the student is given a statement, such as “This faculty member presented material in a clear, well-organized fashion.” The student then is given response choices on a Likert scale: “1 = strongly agree, 2 = agree, 3 = neutral, 4 = disagree, 5 = strongly disagree.” The data are reported to the faculty member with the complete distribution of student responses and a mean score. Oftentimes the mean score is compared to other faculty members’ mean scores or an overall institutional mean when making promotion or tenure decisions. But a mean is not an appropriate statistic to calculate for the Likert scale or, in fact, for any ordinal data. Mathematically, the use of a mean requires a scale that is unique up to positive linear transformation, that is, at least an interval scale. More concretely, this means that there is no meaningful sense in which two scores of 2 (for instance) can be equated to one score of 1 plus one score of 3. The appropriate statistic to use for this data is a frequency count. If a one-number summary of the data is desired, the frequency of respondents rating the faculty member at 1 or 2 can be found easily. Because these are absolute data, arithmetic is meaningful and percentages can be calculated. These percentages then can be used meaningfully in decision making. Alternatively, a more comprehensive summary could be given by the cumulative percentages up to each value. Thus a report as (61.2, 90.0, 95.7, 99.0) indicates that 61.2% of the students strongly agreed with the statement, 90.0% at least agreed, 95.7% were not negative, and only 1% strongly disagreed. Only four numbers are needed because the last value will always be 100%.

This critique also applies to student grades, which are typically awarded on such a scale as “A, B, C, D, and F.” Such grades are ordinal, not numerical data. Yet typically they are assigned numerical values of 4, 3, 2, 1, and 0 and “grade point averages” (GPA) calculated. This custom probably has persisted because it provides a convenient way to compare students and because it does provide a somewhat meaningful summary of the

data—a grade point average near 4.0 means almost all As, whereas a GPA near 2.0 means few As and probably many Cs. But our previous critique still holds—a student with half As and half Cs gets the same GPA as a student with all Bs. But the students' performances are quite different. Still the use of the GPA as a measure of student achievement is probably not too harmful if the entire grade report is examined as well as the GPA and if no major decisions are made on the basis of fine distinctions.

Limitations Arising from Contextual Norms

So far we have focused on methodological norms—those arising from limitations of measurement itself. In this section we ask what limitations are placed on us by contextual norms.

The most obvious limitation of measurement in the social sciences is that controlled experiments with human situations are often nearly impossible or are unethical. When controls cannot be constructed, social science research is not as persuasive as natural science research and causality is difficult to establish, as we shall see in Chapter 8.

But the proscription of controlled experiments on human beings is not the only ethical constraint on measurement of human beings. We mentioned another briefly at the beginning of this chapter, that measurement changes people inasmuch as people internalize measurements performed on them. This internalization changes the persons measured and their relationships with others. Consider, for example, a literacy worker in an impoverished community. She measures the reading level of a man who has completed a sixth-grade education and is respected because of it. She finds he reads at a second-grade level. Her relationship with this man is immediately and irretrievably altered—perhaps for better or worse, but it is changed. When he learns the score, his self-concept will be altered, possibly in harmful ways. His relationship with his school and former teachers also will be changed. Similarly, a child in school who consistently receives word that she scores at the 99th percentile on standardized tests has much of her self-concept and her relationship with her teachers formed by this information. A child who scores below the 10th percentile also will have his or her self-concept and relationships formed but in a very different way. Even groups' self-concepts and relationships are affected by measurement. For instance, the measurement and labeling of groups as "disadvantaged" has resulted in such groups forming an identity and seeking equal treatment.[7] The ethical consequences of such measurements are great even if the measurements are meaningful and accurate. If they are invalid, the potential for

harm is enormous.

The medical profession espouses the contextual norm “Do no harm.” That norm is applicable to the measurement of human characteristics as well as to medical treatment. Human measurement only should be made after the potential impact on the persons measured has been carefully considered and adequate preparation made to deal with possible negative consequences. Furthermore, educational programs that teach people how to use measurements need to include training in the ethics of dealing with the changes those measurements produce in people.

Contextual norms can enter in another way. The purpose for which a particular measurement is performed can compromise contextual norms. As we discussed in Chapter 4, objectivity is a limiting ideal that is never fully attainable. But objectivity is sometimes ignored altogether. For instance, evaluation of a project often becomes an unpleasant necessity to keep grant money flowing. In such cases what is measured, how it is measured, and what is reported all are subordinate to a goal other than truthfulness, which is certainly a contextual norm.

Two further limitations arise from the fact that measurement is performed in a social context.

1. Human information, no matter how carefully and accurately gathered, often requires a discussion of values and principles before meaningful interpretation is possible. For instance, consider the unemployment rate. A cartoon portrayed a statistician reporting that his research had shown that a 7% unemployment rate was acceptable to 93% of the people. The 7% rate cannot be given a meaningful interpretation apart from values and principles that determine what unemployment rate is acceptable. Selection of such values and principles requires some societal discussion.
2. The impact of human information is never neutral, as it is reported in a context that involves human preconceptions, values, and beliefs. For instance, if a school official announces that 70% of the children enrolled in Franklin Delano Roosevelt High School are minority students, the 70% figure is reported on an absolute scale and involves no proxies, so it seems neutral. But it is received in a context in which there are multiple assumptions about the nature of Franklin Delano Roosevelt High School, many judgments about what constitutes an appropriate proportion of minority enrollment, and many different understandings of the word “minority.” In short, a simple, apparently objective piece of human data has enormously different meanings to different people. Reporters and users of such data need to consider the context in which they will be reported.

Case Studies

We conclude this chapter with three case studies that will suggest how normative thinking in the conduct of social science can be done and be helpful.

Intelligence

Although the measurement of intelligence has a somewhat sordid history, it began with noble objectives. Alfred Binet (1857–1911), whose name has been attached to one of the most famous IQ tests, the Stanford-Binet test, was commissioned in Paris in 1904 to conduct a very practical study: to find means of identifying children who would not be successful in regular classrooms but could benefit from some form of special education. Binet developed a test with a large number of short tasks related to everyday problems of life and graded these in order of difficulty. He included tasks suitable for many different ages. A child's “mental age” then was identified by the number of tasks he or she could complete. After Binet's death, this mental age was converted to an “intelligence quotient” by dividing it by the child's chronological age and multiplying by 100. The test score is thus an operational definition of intelligence.

After Binet's work, IQ testing began in the United States and intelligence was hypothesized to be genetically related. Differences in the scores of ethnic and racial groups surfaced and these were used as arguments for the inherent superiority or inferiority of different groups. In perhaps the worst abuse of measurement known, during the 1930s Jewish refugees who anticipated trouble with the Hitler regime sought to emigrate to the United States but were denied. The grounds were that, on the basis of IQ test scores, they were genetically inferior. Controversies about racial and ethnic links to intelligence have periodically resurfaced, as in the publication of *The Bell Curve* by Richard Herrnstein and Charles Murray in 1994. In this book the authors point out that African Americans score an average of about 15 points below white Americans on IQ tests. The book's most controversial feature is a claim that between 40% and 80% of this difference can be attributed to genetic factors.

In this section we want to examine only the measurement of intelligence; we will return to causative explanations of IQ differences in Chapter 8. We hope to demonstrate that the analytic techniques we have presented can help us understand why the history of intelligence measurement has been so controversial and also help make some normative judgments about the measurement and use of IQ.

Note the situation: Although intelligence has been widely studied since the 19th century (and reflected on much earlier), there does not exist a widely agreed-on theory of intelligence. But there are strong incentives to measure intelligence. Societies would like to enable their most able young people to be well trained, employers would like to be able to select the most able employees from applicants, militaries would like to identify potential officers, and colleges would like to select the most promising applicants. Thus, a test that is easily administered and yields a numerical score is very attractive. So again, we encounter the temptation to cut corners.

There have been three main approaches to the study of intelligence. The approach that has been the basis of IQ testing after Binet sees intelligence primarily as a single mental factor, usually denoted *g,* for “general intelligence.” The evidence for *g* began with the work of Charles Spearman in 1904. Spearman developed a statistical technique, called factor analysis, to aid in distinguishing the number of distinct “factors” in a collection of multiattribute data. That is, suppose an intelligence test consists of a collection of a dozen subtests (for instance, vocabulary, geometric visualization, etc.), each subtest consisting of closely related questions. A student is awarded a score on each of the 12 subtests and each subtest is then regarded as one attribute. Students who score well on one subtest tend to do well on other subtests also, so scores on the subtests tend to be correlated. Factor analysis enables the identification of one or more principal components that make up these scores. Note that the components identified are mathematical abstractions; they may or may not correspond to anything real. Nevertheless, because IQ test scores consistently yield a single dominant component, that component has been reified and called *g.*

Recent work by Howard Gardner represents a second approach to understanding the nature of intelligence. Gardner argued for seven distinct intelligences: linguistic, musical, logical-mathematical, spatial, bodilykinesthetic, interpersonal, and intrapersonal. Other scholars have argued for an even larger number of distinct intelligences.

A third approach is that of Jean Piaget. Piaget started his career in Binet's laboratory in Paris around 1919, after Binet had died. He quickly turned his studies away from the quantitative measure of intelligence to the study of the underlying structure of intelligence. In Piaget's view, which he called genetic epistemology, human intellect develops through a sequence of four main stages, with many substages. These stages have the same basic form, independent of culture. The different stages represent different types of reasoning.

We are now at the point at which we can conduct a normative analysis of IQ: *The principal methodological*

problem is that IQ tests measure intelligence on a numerical scale, but at present there exists no theoretical justification for a correspondence between intelligence and numbers. That is, IQ scores for an individual are typically relatively consistent over time and are often correlated (or inversely correlated) with other important factors, such as income, social class, and rates of out-of-wedlock births. Thus, they are indeed measuring something important. But because IQ scores are not based on a meaningful concept of intelligence and because no one has demonstrated that intelligence corresponds to numerical values, IQ tests cannot properly be called "intelligence tests." Herrnstein and Murray[8] partially recognized this critique and dismissed it.

> Before something can be measured, it must be defined, this argument goes. And the problems of definition for beauty, justice, or intelligence are insuperable. To people who hold these views, the claims of the intelligence testers seem naive at best or vicious at worst. These views … are generally advanced primarily by non-specialists. (pp. 17–18)

After this comment Herrnstein and Murray did not return to the critique and wrote another 645 pages. But the critique cannot be diminished this lightly, especially when one considers the enormous social harm that IQ measurements have been used to justify. Even Binet was acutely conscious of the limitation we have addressed here. Writing about his intelligence scale in 1905, Binet[9] asserted, "The scale, properly speaking, does not permit the measure of the intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured."

The application of factor analysis to the study of IQ test scores is also problematic. Factor analysis of intelligence test scores simply identifies the existence of a dominant factor; it does not tell us what that factor is. Spearman and others have interpreted this result as unitary intelligence, *g.* But, as we saw in Chapter 4, such an interpretation is heavily dependent on the presuppositions one brings to the act of interpreting. For instance, the dominant factor just as well could be interpreted as symbol manipulation skill, because all of the tests depend on the test taker's capacity to manipulate verbal, visual, or oral symbols. That is, the case for g is far from conclusive.

Contextual norms also help to analyze the measurement of intelligence. Some commonsensical norms that clearly apply are respect for persons, service to others, compassion, and stewardship of human resources. From the point of view of these principles, intelligence testing was originally well-intentioned, but its subsequent use at times has been deplorable. It has been used to select an elite group and to justify discriminatory

practices toward those who are judged inferior (as in the 1930s). Such uses obviously violate all of these norms.

In conclusion, it is clear that IQ tests are measuring something. It is not so clear what that something is. Calling it “intelligence” is dishonest, inasmuch as no theoretical foundation has been advanced to demonstrate a correspondence between any definition of intelligence (other than the operational definition: “Intelligence is what IQ tests are measuring.”) and an IQ score. Nevertheless, we do not want to exclude all mental measurements. For instance, SAT and ACT (American College Testing service) tests have proven to be moderately effective indicators of potential success in college. Thus, they provide a means for colleges to compare students from different high schools. When used for this purpose and in conjunction with other indicators, such as high school grades, recommendations, and personal interviews, they can be genuinely helpful. Such measures as IQ scores also can be helpful in decisions, such as finding appropriate educational programs for slow learners, if we recognize the limitations of IQ scores and use them in conjunction with other data and personal knowledge of the test taker.

Stress

One influential study in psychology is the work of Thomas Holmes and Richard Rahe on stress.[10] Holmes and Rahe were studying the link between stress and illness. From their clinical experience, they compiled a list of 43 life events that they judged to be stressful, that is, that required individuals to make significant psychological adjustments to adapt to the event. Then they asked 394 subjects to rate the stressfulness of these events on a numerical scale. The following instructions were given to each subject.[11]

> In scoring, use all of your experience in arriving at an answer. This means personal experience where it applies as well as what you have learned to be the case for others. Some persons accommodate more readily to change than others; some persons adjust with particular ease or difficulty to only certain events. Therefore, strive to give your opinion of the average degree of adjustment necessary for each event rather than the extreme. … “Marriage” has been given an arbitrary value of 500. As you complete each of the remaining events, think to yourself, “Is this event indicative of more or less readjustment than marriage? Would the readjustment take longer or shorter to accomplish?” (p. 173)

The ratings of the 43 events then were averaged, the mean was divided by 10, and rounded to the nearest whole number. The results are given in Table 6.1.

TABLE 6.1 The Social Readjustment Rating Scale

Rank	*Life Event*	*Mean Value*
1	Death of spouse	100
2	Divorce	73
3	Marital separation	65
4	Jail term	63
5	Death of close family member	63
6	Personal injury or illness	53
7	Marriage	50
8	Fired at work	47
9	Marital reconciliation	45
10	Retirement	45
11	Change in health of a family member	44
12	Pregnancy	40
13	Sex difficulties	39
14	Gain of new family member	39
15	Business readjustment	39

16	Change in financial state	38
17	Death of a close friend	37
18	Change to different line of work	36
19	Change in number of arguments with spouse	35
20	Mortgage over $10,000	31
21	Foreclosure on mortgage or loan	30
22	Change in responsibilities at work	29
23	Son or daughter leaving home	29
24	Trouble with in-laws	29
25	Outstanding personal achievement	28
26	Wife begin or stop work	26
27	Begin or end school	26
28	Change in living conditions	25
29	Revision of personal habits	24
30	Trouble with boss	23
31	Change in work hours or conditions	20
32	Change in residence	20
33	Change in schools	20
34	Change in recreation	19

35	Change in church activities	19
36	Change in social activities	18
37	Mortgage or loan less than $10,000	17
38	Change in sleeping habits	16
39	Change in number of family get-togethers	15
40	Change in eating habits	15
41	Vacation	13
42	Christmas	12
43	Minor violations of the law	11

Note that the dollar amounts in items 20 and 37 reflect 1967 prices. To use the Social Readjustment Rating Scale (SRRS), an individual circles the items on the list that have occurred in his or her life in the previous 12 months and totals the points assigned to those items, giving a score in “life change units” (LCUs); this score then is regarded as a measure of the amount of stress a person has experienced in the previous 12 months.

Subsequent to the development of the SRRS, statistical analysis showed that correlating a person's LCU score with health indicators only accounts for about 10% of the variation in health. Furthermore, the scale has been critiqued on several grounds: It includes positive and negative events in the same scale. It includes events over which a person has control along with events over which a person has no control, although research has shown that sudden, negative, uncontrollable events are far more effective predictors of illness than positive, controllable events. It does not take into account the meaning of the event to the person. For example, a planned pregnancy has a very different stress effect than an unplanned one. Many of the items also are vague. But despite these criticisms, the SRRS has stimulated much research and has been given a great deal of attention in popular magazines and newspapers.

The techniques for normative analysis of measurement we have presented here can, we believe, enable an

additional useful critique of the SRRS. First, note how the scale was established. The SRRS is a ratio scale in that two fixed points—marriage as a 500 and absence of stress as a 0—were established, and values were assigned to other events in relation to the two fixed points. (Although the 0 was not explicitly mentioned in the directions to subjects, it is implicit in the comparison to 500. That is, subjects are judging what fraction of 500 corresponds to a particular stressor, thus rating on a scale with fixed points of 0 and 500.) Once a ratio scale has been established many capabilities follow, though the only one used by Holmes and Rahe is additivity. For instance, if the ratio scale is valid, one can meaningfully say that foreclosure of a mortgage or loan is twice as stressful as change in number of family get-togethers. Also, the incremental stress of death of a spouse as compared to divorce is the same as the incremental stress of an outstanding personal achievement over a minor violation of the law. We suspect that most persons would be uncomfortable with these multiplicative and incremental comparisons and, in fact, Holmes and Rahe don't make them. But these operations are intrinsic in use of a ratio scale; if we are skeptical about these operations, we ought also be skeptical about the additivity that Holmes and Rahe did use.

Let's examine the SRRS in the light of the four requirements for measurement. First, as critics have pointed out, there are serious definitional problems. Holmes and Rahe defined stress as the need for psychological readjustment to an event. This is not an operational definition but rather an attempt at an essential definition. But common lexical definitions include stressful states (not just change) and distinguish stress caused by negative events from stress caused by positive events. Furthermore many of the events listed are described ambiguously. So the likelihood of a socially useful measure of stress is diminished by the somewhat nonstandard definition of stress that Holmes and Rahe used and by the inconsistent interpretations of the events likely to result from their ambiguous descriptions. Even so, this in itself does not make the measure socially useless or harmful—recall how social class still maintains some usefulness as a concept despite its ambiguous definition.

The second requirement for measurement is intersubjective agreement on ranking. Unfortunately, this important requirement was finessed by simply averaging the subjects' scores. Considerably more insight into stress could have been gained by examining subjects' rankings of the stress-producing events, doing correlational studies of the lists of rankings, identifying when there was strong agreement and disagreement, and discussing areas of disagreement with subjects to find out why they perceived the stressors differently.

The third requirement is a theory of the properties of the entity being studied that justifies its correspondence

with the desired scale. From our perspective, this is the major weakness of the SRRS. The authors began by assuming the appropriateness of a ratio scale and then asked subjects to rate stressors relative to fixed points. But the essential prior question of the appropriateness of this scale was not addressed. Why should one assume that stress is one-dimensional? Stress may have multiple components that affect health in different ways. It is not difficult to conceive of the notion that certain types of stressors cause ulcers, others cause headaches, and others affect the immune system. Such a notion may or may not be correct, but an analysis of the structure of stress must precede the establishment of a numerical measure. Furthermore, Holmes and Rahe assume additivity when they use a ratio scale. But we see no *a priori* reasons to believe that stresses are linearly additive. Stresses may combine in a multiplicative fashion; perhaps some people are able to discount the effect of additional stressors in some fashion. Whatever the case may be, the assumption of a simple linear additive model is unwarranted.

And last, the measuring instrument is vulnerable to the critique that “What would you do if” type questions are being asked. We have seen in the discussion of Richard LaPiere's work that such questions are notoriously unreliable.

The SRRS is a good example of “corner cutting”—trying to gain the benefits of a numerical measure without paying the price of the substantive theoretical work required to determine whether such a measure is appropriate. But what about contextual norms? Can they add anything further to our understanding of the SRRS? Our main concern is the potential harm to the credibility of measurement in the social sciences resulting from the use of a poorly grounded measuring instrument like the SRRS; many people intuitively recognize it as an oversimplification, even without understanding the foundations of measurement.

Project Evaluation

In this case study we examine an evaluation technique commonly used by granting agencies. Our example here is based on an actual agency, but the details have been modified to preserve anonymity. We call our (fictitious) agency Third World Economic Development (TWED).

TWED's long range strategic plan states: “Our vision is to enable and empower people to undertake independent responses such that the poor and their community flourish.” It also identifies its “core values” as stewardship of natural resources, independence, and justice. These terms are not explicitly defined and this causes

some problems, as we shall see.

TWED is quite explicit about its goals, objectives, and the strategies it uses to achieve them. Its international objective is “To serve 100,000 families in poverty bringing them toward self-sufficiency, to help organize local community groups, and to work with established organizations enabling them to independently identify and resolve their own community problems and needs.” TWED employs five strategies to achieve these goals.

1. Organizational development and collaborative planning (providing consultation for other similar agencies)
2. Community development (training and leadership development, increasing income through food production or small industry, improving health care largely through preventive measures, and increasing functional literacy rates)
3. Leadership development (training local groups to provide assistance)
4. Disaster response (direct aid, refugee resettlement)
5. Constituency participation (volunteer programs, on-site visits)

TWED uses a somewhat informal approach to select new sites at which to apply the previous five strategies. Nevertheless, once a site has been selected TWED uses quantitative methods in its ongoing planning for that site. A single “international planning form” is used with minor variations at several levels of planning—the individual local site, the country, and the project (a collection of local sites within a country). It consists of four parts and is completed annually by staff directly involved at each level. The first part is a simple summary of dates (when the project was begun and when it is expected to become independent), other organizations involved, and anticipated cost. The second part asks the evaluator to assess either the community in which the project is being conducted or the organization conducting the project. Table 6.2 is a copy of this second part for a community. It is explained below.

TABLE 6.2 TWED Community Capacity Indicators

Organizational Capacity Indicators	92	93	94	95	96	97	*Levels of Independence*
Technical							5 = independent
Management (incl. financial)							4 = adequate quality
Networking and Resource Development							3 = needs improvement/ cooperation
Board Control							2 = unsatisfactory results
Holistic Outreach							1 = not functioning
Average							

The third part of the international planning form asks about the number of families involved and requests some financial data, including cost per family. Part four asks for specific objectives for the project. These differ, depending on the nature of the project. For example, on health care it asks for the number of children from ages 0 to 6 and the number of families to which they belong, the current death and malnutrition rates, and the target rates at the site.

Once these data have been collected, they are used to make a variety of comparisons, such as site to site within the country, country to country, and project director to project director. Comparisons are not made mechanically by, for instance, comparing average scores. Effort is made to look at the entire assessment and to take different contexts into account.

We now have enough information to evaluate this measurement technique. We will focus primarily on the community capacity table. Note that the measurement scale being used for each indicator is an ordinal scale. As pointed out earlier in this chapter, there are two requirements for using an ordinal scale: unambiguous definitions (both of the indicator being measured and of the scale categories) and intersubjective agreement on rankings.

First, we consider definitional issues. Table 6.2 serves as an operational definition of quality for TWED—the five indicators constitute five equally weighted aspects of quality. Nevertheless, the five indicators also oper-

ationally define the core values on which TWED bases its work, despite the long-range plan's core values of stewardship, independence, and justice. That is, inasmuch as decisions on funding and staff performance evaluation are made primarily on the basis of scores earned on the five indicators, these indicators, not the list of core values, serve as the primary guides for the work of project directors. These two sets of values are not necessarily inconsistent, but the existence of two statements of values, one of which is operationalized and one of which is not, may be confusing and could result in a misapplication of the work. In fact, looking over the five indicators and the measurement scale, one would conclude that TWED's core values actually are operational effectiveness (which could be interpreted as stewardship) and independence. It is not clear how justice is being operationalized.

We don't want to be overly critical of TWED on this point. The observation we have just made—of a potential difference between an organization's explicitly stated values, goals, or both and the operational definitions that actually guide its decision making—is extremely common. Nevertheless, it would be wise for any organization to think carefully about whether a gap exists between the two.

Given the use of these five indicators of quality, we next must ask if they are unambiguously defined. Explicit definitions are not stated for them, but implicit definitions are given in a separate questionnaire in which evaluators are asked a series of questions grouped under the headings of the five indicators. For instance, under “technical capacity” questions are asked about staff expertise and creativity, experience working with target groups, and assessment of community problems, needs, and priorities. Each question asks for a response on a five-point scale where “1 = not functioning, 2 = unsatisfactory, 3 = needs improvement, 4 = adequate, and 5 = excellent.” For the indicator “management,” 22 such questions are asked; under “networking,” two are asked. Under each of “board control” and “holistic outreach,” three questions are asked. The vast difference in the number of questions asked in each category suggests that in fact management capacity may be the most important of the five factors to the TWED staff. But the summative instrument does not reflect this and the averaging process treats them equally.

The key definitional issue is whether all TWED evaluators and their supervisors understand the terms in the same way. The questions help a great deal to communicate this common understanding. But it could be enhanced even more by succinct, explicit definitions. Staff training also could help; we will comment more on this point later.

Another definitional problem is that the constituency may not agree with TWED's definition of quality, partic-

ularly if the constituency is defined broadly enough to include the poor with whom TWED is working. Without an agreement on the definition of quality, measurement of it is meaningless. That is, North American staff can form an operational definition and can train field personnel to use it. But at least some people in the many diverse cultures the field staff work with are likely to have quite a different concept of quality. This is especially critical because TWED's goal statements place a high value on independence. Hence, there is a potential inconsistency between the relatively well-defined operational definition TWED has formed and its value of independence. One possible resolution of this inconsistency is to allow local cultures to modify the definition in dialogue with TWED staff. A common definition throughout the organization does provide a great benefit by providing for common understanding. But because it originates with the granting agency, it may not serve the goal of independence.

Beyond these problems of definition, use of an ordinal scale such as the one TWED has used requires intersubjective agreement on the meaning of the five levels on which each indicator is ranked. The TWED scale has some problems here. Most serious is the ambiguous definition of the five levels.

Unfortunately, people's intuitions in complex situations as those being evaluated by TWED vary a great deal. Thus, explicit instruction is needed to ensure meaningful intersubjective agreement. This could take several forms.

- Explicit definition of the categories “Excellent,” “Adequate Quality,” and so forth
- Metaphors, such as “Ripe,” “Sprouting,” “Newly Seeded,” and so forth
- Preparation of a list of scenarios that illustrate each of the five levels for at least some of the five indicators. An evaluator could look at the scenarios and say “Now which of these come closest to fitting my situation?”
- Explicit training for staff in the use of the planning form. Training might best be done with sample case studies that could be evaluated by staff and then discussed.

Even if there were intersubjective agreement on the definitions of the categories, there remains a problem with the use of the average at the bottom of Table 6.2. Granting agencies frequently use ordinal scales similar to the scale used by TWED, then average the scores on different indicators. Such an average has the decided benefit of being an easily computed summary measure. But it is technically meaningless because the data are not quantitative. Consider two sites. Suppose one is rated an “excellent” (5) on one indicator and a “not functioning” (1) on another. The second site is rated “needs improvement” (3) on both. The average for each

site is a 3 and hence the method equates these two sites. But there is no meaningful sense in which they are equal. A better summary measure is the frequency count (for example, “This site rated two excellents and three adequates”). Alternatively, a vector listing the frequencies at each level, such as (2,3,0,0,0) could be used. The choice of a summary measure depends on the values of an organization. In some settings, high quality performance in a few areas is very important, whereas poor performance in other areas can be overlooked. In such a situation, a frequency count of the “excellents” would be a good measure. In other situations good performance across all categories is of great importance. In such a case, the frequency of “excellents” plus the frequency of “adequates” might be a good measure; a penalty could even be assessed for low scores by subtracting their frequency.

It also would be wise for TWED to drop the numbers associated with the levels because numbers are normally interpreted as cardinal values rather than ordinal values—stating the levels as numbers immediately introduces the temptation to do arithmetic on them. Even using the letters A, B, C, D, and E should be avoided inasmuch as these letters are so commonly converted to numbers and averaged in schools. It would be better to use verbal descriptors like “exc,” “adeq,” “nimp,” “unsat,” and “nf” that have no natural numerical associations yet still allow for frequency counts.

The use of quantitative objectives (as in the fourth part of the form) is helpful, but, as TWED itself has realized, the indicators selected are short-term and do not necessarily measure sustainable changes after the project is completed. They represent “output” rather than “outcome” changes. TWED is currently working to replace these with more long term measures.

In summary, then, a measurement-based approach to evaluation is a helpful tool for assessing quality. Nevertheless, because such a tool operationalizes the values of the organization, considerable care must be exercised in selecting the right proxies for quality. That is, the indicators selected must accurately represent the values of the organization. Such a tool also requires clear definitions shared by all evaluators and requires significant effort to ensure intersubjective agreement on rankings. Finally, in a situation like TWED's in which sites are located in many different cultures and involve many different types of projects, local variations in definitions of quality are needed, even though this causes some difficulties with communication.

Notes

1. For some astounding examples of mismeasures in the area of intelligence that had serious consequences, see Gould (1981).

2. For our discussion of definition in this section, we are very much indebted to Professor John Edelman of the Philosophy Department of Nazareth College of Rochester, NY His unpublished manuscript explicating these ideas more fully is available on request. The college's address is 4245 East Ave., Rochester, NY, 14624.

3. For those interested in a concrete example of a definition that is simultaneously essential and operational, the word *function* is defined by mathematicians as follows: Let A be a set, B be a set, and consider a set of ordered pairs of the form (*a, b*) where *a* is a member of A and *b* is a member of B. The three sets are a function if every a in A corresponds to one and only one *b* in B.

4. See Bridgman (1927).

5. See Allport (1958).

6. LaPiere (1934, pp. 230–237).

7. See the article by Guillemin and Horowitz, Section 10.3.C in Callahan and Jennings (1983).

8. Herrnstein and Murray (1994).

9. As quoted in Gould (1981, p. 151).

10. See Holmes and Rahe (1967). For a clear summary and evaluation of Holmes and Rahe's work see Hock (1992).

11. Holmes and Rahe (as quoted in Hock, 1992, p. 213).

https://doi.org/10.4135/9781483348872