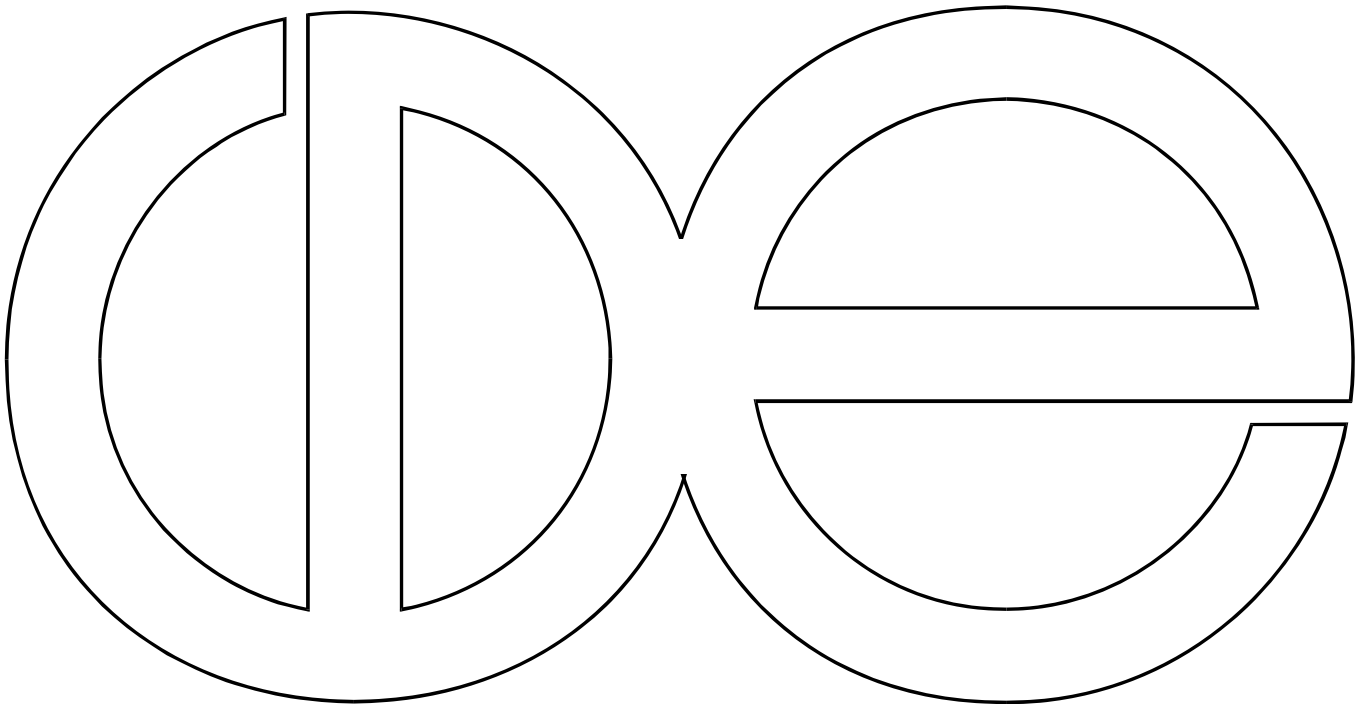# Center for Demography and Ecology

University of Wisconsin-Madison

Comparable Metrics: Some Examples

Robert M. Hauser

CDE Working Paper No. 2010-08

Comparable Metrics: Some Examples[1]

Robert M. Hauser

Center for Demography of Health and Aging

University of Wisconsin-Madison

February 8, 2010

There is a delicate balance between standardization and validity of measures of social scientific constructs. There are three good reasons to standardize measurement. First, standardization introduces the possibility of common understandings. Second, it permits and may encourage the accumulation of evidence. Third, it permits and may encourage valid comparisons across time, place, or units of observation, e.g., persons, families, settings, localities, or organizations. Because standardization typically entails loss of information, standardized measures have drawbacks that parallel their benefits. Shared understandings, when based on a common, but incomplete measure, may miss important features of a phenomenon. Even extensive evidence may be uninformative or misleading. Comparisons may miss important differences among the units or cases being compared. Thus, one of the tasks of social scientific research is to negotiate and balance competing demands for construct validity and standardization of measurement.

Trade-offs between standardization and validity are complicated because measurement cannot be separated from representation – who or what is being measured – and from analysis – how data will be described and used. In this essay, I will illustrate these ideas with several

---

examples, a few historical and others based on my experience as a researcher and observer of the social, behavioral, and economic sciences. For that reason, my text is more autobiographical than systematic, and it is undoubtedly lacking on that account.

Public Metrics: Some Examples

Standardized measurement enables communication, debate, and – sometimes – action, whether in the worlds of policy, politics, science, or among the lay public. The following examples are presented in declining order of success, based on my judgment of the validity and usage of the measures.

The Unemployment Rate

People in the United States know, or at least think they know what an unemployment rate means. That rate was a social-scientific invention, a behavioral measure based on reports of job search during a reference week by members of the labor force. The unemployment rate is simply the ratio of the number of unemployed individuals to the number of members of the labor force (employed plus unemployed). It was invented during the Great Depression of the 1930s after it became clear that the previous "gainful worker" concept offered little insight into the state of the economy (Hauser 1964). It was given legal status in the Employment Act of 1946 (Brinkley 2001).

To be sure, almost from the beginning – and continuing to this day – there has been discussion and debate about defects in the unemployment measure. The officially unemployed do not include "discouraged workers," persons who have abandoned the job search because they believe that no jobs are available, and they do not include employed individuals who are working fewer hours than they would like (United States National Commission on Employment and

Unemployment Statistics 1979). These defects are especially visible during recessions because the standard definition of unemployment under-estimates the extent of economic distress.

The Poverty Line

The official poverty rate in the United States is another social-scientific invention that has had a long and visible life in social policy, social science, and common parlance, despite major weaknesses that greatly limited its validity and usefulness from the outset and across the past four decades (National Research Council 1995). The poverty standard compares pre-tax income to economic need, based on a ratio of total income to the cost of a minimal diet, adjusted for family size. Because the poverty line is absolute, updated only to reflect changes in the Consumer Price Index, because living standards have changed, and because the share of food in family budgets has changed, the standard has become increasingly obsolete. The Earned Income Tax Credit (EITC), the most effective income support program in the U.S., has never moved a family across the official poverty line because it is a post-tax income supplement, while the poverty standard is based on pre-tax income. Putting the matter simply, the official poverty concept does not order families accurately with respect to their access to economic resources. All the same, the poverty standard and measure have become embedded in social and economic structure, though with occasional embellishments and modifications. For example, eligibility for the National School Lunch Program (NSLP) and some other social programs is based on a multiple of the poverty standard. Official poverty has also been over-used in thousands of research papers and books; it is daunting to see how our perceptions about poverty and the poor would differ if a standard measure of greater validity were widely accepted (National Research Council 1995:247-92).

The use of the NSLP as a measure of socioeconomic standing in the National

Assessments of Educational Progress (NAEP) provides an especially compelling negative

example of the loss of information in standardized measurement. For many decades, under

federal law, elementary and secondary pupils have been entitled to a free school lunch if their

income falls below 125% of the official poverty line and to a reduced-price lunch if their income

falls below 185% of the line. Eligibility for the NSLP is ascertained by individual schools and

kept in each student's record. The National Assessment Governing Board (NAGB) requires that

academic achievement differentials be reported by "socioeconomic status" (SES) among other

key variables. Because NSLP participation is nominally an indicator of poverty status and is

readily available in student records – not requiring a potentially unreliable or invalid report by

the student – it has been adopted as a standard measure of SES in NAEP. This choice is doubtful

on its face because a free or reduced-price lunch is a treatment, not merely an indicator. While

the NSLP standard is nominally the same across grade levels, school systems, and years, those

presumptions cannot be supported. As noted above, the official poverty line has become

increasingly obsolete. Students in the upper grades are less likely than those in lower grades to

participate in the NSLP. In recent years, the federal government has become increasingly relaxed

about treating NSLP eligibility as an individual student characteristic, and now schools are often

granted blanket eligibility.

Academic Achievement Levels

A more recent example of a nominally social-scientific, standardized measure that has

become visible and influential in public discourse and policy is the labeling of academic

achievement levels in K-12 education as below basic, basic, proficient, or advanced.[2] The use of

these labels was initiated by the National Assessment Governing Board (NAGB) of the National

Assessments of Educational Progress (NAEP) in the late 1980s as a way of improving the

communication of test score distributions to the public, and it has become ubiquitous in reports

of academic achievement in diverse subjects and in state as well as national tests. NAEP had

begun in 1969 with an assessment of a sample of 17 year-olds in citizenship, science, and

writing. It evolved to cover diverse subjects among students in the 4th, 8th, and 12th grades with a

sampling scheme in which each participant completed only part of a test and score distributions

for aggregates of students were estimated from multiple imputations (Jones and Olkin 2004).

From the outset, controversy enveloped the use of the NAEP achievement level standards and

the judgmental methods by which they were determined. For example, early reports by the

National Academy of Education and by the National Academies each referred to the standard-

setting process as "fatally flawed" (Glaser et al. 1993a; b; National Research Council 1999c).

Another persistent criticism is that achievement standards have been set too high, thus

encouraging and reinforcing continuing and sometimes inaccurate judgments about the quality of

American students and schools. Methodological criticisms of achievement levels have led to

subsequent rounds of rejoinder, research, and revision (Bourque 2004:211-18).

Given the strength and the sources of negative evaluations of student achievement levels,

one might have expected them to disappear. In this case public and political demands for

understandable metrics of educational accountability trumped scientific review. The No Child

Left Behind Act (NCLB) specifically proclaimed the unattainable goal of 100% "proficiency" in

---

[2] Standards are described for basic, proficient, and advanced performance, but the below basic category is left as a residual.

all major subjects – a goal that, as Robert Linn (2003) has noted, surpasses current achievements in any nation or in any school in America. For example, "the proficient standard is set at nearly the 75th percentile at grades 4 and 8 and a little higher than the 80th percentile at grade 12" (p. 5). While pursuing the proficiency goal, NCLB permitted states to set their own achievement standards, based on their own tests, and many states adopted low standards or lowered existing standards in order to set more realistic goals. To counter, or at least to expose such efforts, NCLB made NAEP mandatory at the state level; thus, there may be massive discrepancies between score distributions in state NAEP reports and in the states' own reports of educational progress under NCLB. In this case, the creation of a supposedly scientific set of standards has led to their reification in law, to the creation of competing standards, and to comparisons of populations in differing, but nominally identical metrics.

Adult Literacy

The development of a standard metric for adult literacy presented issues that resemble those of K-12 achievement and were briefly in public view, but appear to have been resolved for the moment. The 1992 National Adult Literacy Survey (NALS) combined a standard social background questionnaire, an inventory of reading practices, and brief assessments of reading ability in prose, documentary, and quantitative texts (PDQ). These were administered in a national household survey, supported by the National Center for Education Statistics (NCES), which was supplemented by selected state-representative samples and by a sample of the incarcerated population (Kirsch et al. 1993; Kirsch and Kolstad 2001). Literacy score distributions were cut into five (numerically labeled) levels at points that supposedly indicated discrete breaks in competence. They appear to have been chosen somewhat arbitrarily, for equal

6

test-score intervals are bounded by round numbers in the three central, closed-ended categories (National Research Council 2005a).

No one would have noticed these details, but for the events surrounding a press conference where the findings of the study were presented. Members of the press repeatedly asked the investigators, "How many Americans are illiterate?" – a question for which they were completely unprepared. Eventually, the Secretary of Education, Richard Riley, intervened and pointed to the line between levels two and three, thus branding 49 percent of American adults as illiterate. The error of this extemporaneous public standard-setting was compounded by later confusion about the 80 percent difficulty level that was used to identify and describe items at the cut-points (National Research Council 2005a: 69-71). Consequently, when the National Center for Education Statistics was about to undertake the National Assessment of Adult Literacy (NAAL) in 2003, it asked the National Research Council to determine standards for adult literacy that could be used in NAAL and applied retroactively to NALS in order to compare literacy levels across the decade among all adults and specific population groups.[3] The task was completed using a combination of a heavily monitored but explicitly judgmental method (bookmarking) in combination with descriptive population characteristics. This process yielded categories with explicit descriptions corresponding roughly to readiness for successive levels of formal education (National Research Council 2005a), along with experimental findings suggesting that expert judges are not fully responsive to variation in literacy standards. The new standards for adult literacy had greater credibility, but the new NCES study also had much less visibility than its predecessor.

---

[3] This was feasible because much of the assessment content was carried over from 1992 to 2003.

President Clinton's 1997 proposal for "Voluntary National Tests" (VNT) of reading (at grade 4) and mathematics (at grade 8) was a dramatic and failed effort to create a common metric for the assessment of academic achievement and changes in it. The idea was to administer exactly the same assessment to students all over the nation and to report individual and aggregate test scores to all relevant parties – students, parents, teachers, school administrators, public officials, and the public at large. By measuring all students' performance and progress in exactly the same way, there would be no way to avoid responsibility for poor performance, and the tests would presumably have diagnostic value as well.[4] Leaving aside issues of feasibility, many constituencies were strongly opposed to the VNT. For example, minority groups were concerned that low test scores would be stigmatizing, while Republicans declared that a national test would inevitably lead to a national curriculum, thus imperiling the tradition of state and local school control.[5] A compromise between the Clinton administration and the Republican-controlled congress invoked severe limits on the process of test development, and the project was eventually terminated.

In the course of the VNT, the congress made two specific proposals about academic measurement in an attempt to achieve strict comparability of student performance without developing a common test. The first proposal was to equate the scales of existing tests. An NRC report reached the following conclusions (National Research Council 1999b:4):

---

[4] An NRC report, mandated by congress, recommended against the use of the VNT for "high stakes" decisions about students.
[5] Compare this with the provisions of NCLB, enacted with high priority in 2001 by the Bush administration.

"1. Comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible.

2. Reporting individual student scores from the full array of state and commercial achievement tests on the NAEP scale and transforming individual scores on these various tests and assessments into the NAEP achievement levels are not feasible."

The second proposal was to alter existing tests by inserting modest numbers of common items that might be used to equate the overall scales. This, too, was rejected, though not quite with the same force as its predecessor (National Research Council 1999a):

"Embedding part of a national assessment in state assessments will not provide valid, reliable, and comparable national scores for individual students as long as there are: (1) substantial differences in content, format, or administration between the embedded material and the national test that it represents; or (2) substantial differences in context or administration between the state and national testing programs that change the ways in which students respond to the embedded items."[6]

The striking thing about the VNT case is that the congress directly addressed technical issues of comparability in measurement in its proposals to the National Research Council. Although both proposals were scientifically naïve, they had at least the virtue of attempting to establish national comparability in the measurement of individual academic performance.

---

[6] There is, of course, ample evidence that similar observations apply in social survey measurement.

Accumulating Evidence, Comparing Effects

The fundamental scientific value of standardized measurement is that it permits evidence to be accumulated. Social scientific examples of standardization range from qualitative classifications, like race-ethnicity and social class; to numerical scales describing psychological traits, social standing, or economic amounts; to normalized measures of the fit of statistical models and the effects of variables in such models. To this list one might add such social inventions as voting and the metric system, but that would exceed my ambitions (Duncan 1984).

Although I have mentioned both the accumulation of evidence and the comparison of findings in the introduction to this essay, those two processes are so closely linked in the practice of research that I have made no effort to treat them separately in the following discussion.

Social Class

In sociology, no construct is as ubiquitous as social class, and disagreements about its proper measurement fill hundreds of volumes.[7] Issues in class measurement range from the reality behind the measures, i.e., whether social classes are organized entities or merely statistical abstractions, to the dimensions of social structure and process that define class membership, e.g., relationships to the means of production, market position, consumption practices, or political power. In recent sociological research – predominantly studies of social class mobility across generations – there have been three main contenders to be the gold standard for class measurement: A neo-Marxist classification developed by Erik Wright (1997); a neo-Weberian classification developed by Robert Erikson and John Goldthorpe (1992); and variants of the Edwards Scale, a socioeconomic classification of occupations by the U.S. Bureau of the Census

---

[7] Social class is used here in specific reference to a classificatory scheme, not as a synonym for social and economic standing.

that was developed in the 1930s.

It would be difficult to over-estimate the use of the Edwards Scale in social research in the U.S. For at least seven decades, that scale – sometimes described as the Census major occupation groups, and sometimes condensed into upper and lower white collar and blue collar occupations – was extensively used in classifications of social, economic, and political outcomes and descriptions of local and national labor markets. With the major changes in U.S. Census occupational classification systems in 1980 and in 2000, it has become more difficult to maintain comparability in the use of that classification across time. One well-known example of the use of the Edwards Scale – elaborated with selected distinctions of class-of-worker (self-employment vs. wage and salary employment) and industry, was in national studies of occupational mobility between cohorts and generations by Blau and Duncan (1967) and by Featherman and Hauser (1978; Hauser and Featherman 1977). In studies of intergenerational mobility, that classification was useful for two reasons. It captured tendencies toward occupational inheritance – persistence along the main diagonal of a classification of men's occupations by their father's occupations (Blau and Duncan 1967:90-97; Featherman and Hauser 1978:139-61), and it also represented a central socioeconomic dimension of occupational standing (Hauser and Logan 1992; Klatzky and Hodge 1971; Rytina 1989; 1992a; b).

Wright's classification scheme combines broad occupation categories with distinctions of ownership, size of establishment, and supervisory or managerial responsibility. It has chiefly been used in the work of Wright and his international collaborators, but it has received a good deal of scholarly attention, primarily in juxtaposition to the Erikson-Goldthorpe class schema. Among other criticisms, Halaby (1993; Halaby and Weakliem 1993; Wright 1993) showed that Wright's typology had less explanatory power relative to earnings than an additive combination

11

of the component variables. Miech and Hauser (2001) analyzed data on education, social class, and health among 54-year old male and female high school graduates from Wisconsin. They found that the Wright scheme was not significantly associated with an array of health outcomes once educational attainment was controlled.

Like Wright's classification, the Erikson-Goldthorpe class scheme uses data on occupation, self-employment, number of employees, and supervisory status. It was initially developed for use in the CASMIN project, a comparative study of intergenerational class mobility among developed nations (Erikson and Goldthorpe 1985; 1987a; b; 1992; Erikson et al. 1982). Provided the auxiliary information is also available, the classification can be constructed from detailed occupational data in ISCO (International Standard Classification of Occupations) and those of other nations; thus, it is relatively easy to adopt and use. It has had high visibility among members of Research Committee 28 (RC28) of the International Sociological Association. RC28 is a loosely bounded international group of scholars whose work focuses on social mobility; it meets once a year in North America and a second time each year in another nation. A hundred or more scholars attend each meeting, so it provides ample opportunity for diffusion of research measures and methods. Consequently, the Erikson-Goldthorpe scheme has been used widely both in national studies and in international comparative research. Unfortunately, as shown by Hout and Hauser (1992), the classification used by Erikson and Goldthorpe in the CASMIN project combines categories that are heterogeneous in prestige, education, and income and thus suppresses the main, socioeconomic dimension of the occupational hierarchy. Correspondingly, the widespread adoption of that scheme has yielded cumulative, comparable findings, but those findings systematically underestimate the import of the central dimension of occupational stratification.

The history of these three classification schemes exemplifies the both strengths and weaknesses of common metrics. On the positive side, the schemes have been used extensively in cumulative and comparative research and social reporting. Also, it is feasible to code job descriptions into each scheme – even within the same set of survey data – provided that the descriptions have been coded in detail and that the relevant auxiliary information has also been collected. On the negative side, each scheme competes with the other two, thus reducing the set of comparable studies and observations. Moreover, each scheme has theoretical or empirical weaknesses. The Census major groups have a stronger empirical than theoretical grounding, and it has become more difficult to maintain comparability across time, even within the U.S. federal statistical system, as the parent, detailed occupational and industrial classification systems have evolved. The Wright and Erikson-Goldthorpe class schemes each have a strong basis in sociological theory, but each also has notable empirical weaknesses.

There is another, broader problem with the use of any of the standard measures of "social class," namely, that scientifically naïve researchers and consumers of social data tend to believe that these, or closely related measures of social standing, taken alone, fully represent the social and economic standing of a person, household, family, or family of orientation. This simplistic view fails to recognize the complexity of contemporary systems of social stratification, where inequalities are created and maintained within a substantially, but by no means highly correlated mix of psychological, educational, occupational, and economic dimensions. This, more than the details of class measurement, is the greatest disadvantage of standardization in the measurement of social class.

Occupational Prestige

Scalar measures of occupational standing parallel and to some degree have competed with measures of social class like those discussed above. Their recent history begins with the measurement of occupational prestige, based upon lay or expert reports of the "general social standing" of occupations. Inkeles and Rossi (1956) found that popular ratings of occupational prestige were highly correlated in six industrialized countries: United States, Great Britain, New Zealand, Japan, the Union of Soviet Socialist Republics, and Germany. The correlations ranged from 0.83 to 0.97 – and were mainly in the high end of that range – with the exception of a correlation of 0.74 among ratings of only 7 occupations that were in common between the USSR and Japan (pp. 331-2).[8] Hodge, Siegel, and Rossi (1964) used data from their 1963 NORC survey along with earlier studies of occupational prestige in 1925, 1943, and 1947 to show that there had been little if any change in perceptions of occupational standing in the U.S. across four decades (Counts 1925; Reiss 1961; Smith 1943).[9] Among 23 to 90 occupational titles that could be compared between pairs of studies, the correlations of ratings all exceeded 0.93 (p. 297).

These findings suggested the possibility of creating a common metric of occupational prestige that could be used in international and inter-period comparisons. The idea was pursued by Treiman (1976; 1975; 1977), who exhaustively mined the expanding store of historic and national ratings of occupational standing and developed a Standard International Occupational Prestige Scale (SIOPS). Contrary to claims that there was a unique occupational culture among African-Americans in the U.S., Siegel (1970) showed that ratings of occupational prestige were virtually the same among Blacks and whites and that ratings in both populations largely reflected

---

[8] In that study, there were at most 30 occupational titles in common between pairs of countries.

[9] The temporal stability of occupational prestige was reconfirmed in a 1989 NORC survey (Nakao and Treas 1994).

14

the socioeconomic characteristics of occupations in the dominant (white) population. A clever study in Israel showed that one could extract a prestige scale from a task in which participants were asked to sort occupations by their similarity, without reference to any criterion (Kraus et al. 1978). Occupational prestige appeared to be largely invariant with respect to time, place, the population of raters, and even the design of the rating scheme.

These ideas and findings were not universally accepted. For example, Haller and colleagues showed that in less developed societies there were significant differences in ratings of occupational prestige, relative to the common pattern (Haller and Bills 1979; Haller and Holsinger 1972; Haller and Lewis 1966). In fact, Inkeles and Rossi (1956) had noted cross-national differences among developed nations in the ratings of agricultural and service occupations. Coxon, Davies, and Jones (1978) undertook a comprehensive critique and reconsideration of the entire body of occupational prestige studies, based on multi-dimensional scaling of extensive data obtained from a quota sample of 381 individuals in Edinburgh, Scotland. It was little noticed.[10] Five American social historians mounted a challenge to the historic validity of a common scale by creating a new scale based on their studies of 19th century American cities (Hershberg et al. 1974), to which Hauser (1982) responded at length, showing that the correlations among occupational prestige ratings by the five social historians were about the same as those between their scale and prestige ratings from 20th century America.

SIOPS did not become the international standard for studies of occupational correlates, but it was not primarily a consequence of the critiques of the universality or invariance of occupational prestige. There were two other reasons. First, Treiman's work came at a time when

---

[10] A search in Google Scholar finds 37 citations to Coxon et al. (1978), as compared to 947 citations of Treiman (1977).

sociologists around the world were largely committed to studies of social stratification in their own nations and even more committed to the idea that unique characteristics of their own nations were more important than comparisons among nations. That is, almost no one was willing to suffer the loss of information entailed in the use of a common scale. Second, despite the appealing properties of occupational prestige, empirical findings from studies of correlation of the occupational standing of fathers and sons showed that prestige was not the main dimension of occupational persistence. Rather, intergenerational occupational correlations of the socioeconomic status of occupations – measures of typical educational or income levels of occupational incumbents – were substantially higher than those of occupational prestige. Indeed, occupational prestige behaved as if it were an error-ridden indicator of occupational socioeconomic status (Featherman et al. 1975).

Occupational Socioeconomic Status

In the U.S., the early studies of occupational prestige covered only modest numbers of occupational titles, and they were by no means representative of the entire occupational structure. The 1947 North-Hatt study covered only 90 occupation titles, and it was only after the 1963 NORC study that prestige scores became available for all occupations in the U.S. (Siegel 1971). In the absence of a complete set of prestige scores, Duncan created a proxy measure, "A Socioeconomic Index for All Occupations," by regressing a prestige measure for 45 occupational titles in the North-Hatt study on age-standardized educational attainment and income of occupations held by men in the Census of 1950. That is, he used the regression weights from the matched set of occupation titles to construct an index for all occupations (Duncan 1961).

While the Duncan SEI (Socioeconomic Index) was developed for use in research with vital statistics, e.g., in analyses of differential mortality, the first highly visible use of the SEI

was in Duncan and Hodge's (1963) regression analysis of intergenerational mobility in a 1951 sample of 1,105 Chicago men. Among other observations, Duncan and Hodge wrote (p. 631):

> "It must be conceded that the occupational SES score lacks the properties of a true interval scale, and any reasonable monotonic transformation of it might be justified as easily as the particular set of values used here. … There is little question that the SES index is more suitable for measuring vertical mobility than is the classification of occupations by census major occupation groups. The well-known heterogeneity of occupations within each category of the latter is easily illustrated."

The SEI became established as the preferred measure of the socioeconomic standing of occupations with the publication of Blau and Duncan's (1967) monograph, *The American Occupational Structure*. That work analyzed the first large nationally representative survey of intergenerational social mobility in the U.S., which was based on a supplement to the March 1962 Current Population Survey. By updating the original SEI in light of minor changes in the Census occupational classification system between 1950 and 1960, Hauser and Featherman (Featherman and Hauser 1978; 1977) were able to reproduce and expand the 1962 survey in 1973 and to compare mobility chances between the 1960s and the 1970s.

The SEI was not unique among efforts to create a standard measure of the socioeconomic status of occupations in the U.S. The Hollingshead Index of social position gained visibility after the publication of *Social Class and Mental Illness* (1958). Despite its extraordinarily weak empirical basis – and the fact that its content was never formally published – the Hollingshead Index (1957) has been used widely, especially in epidemiological research (Krieger et al. 1997;

Liberatos et al. 1988). The idiosyncratic and particularistic basis of that index is suggested by a

passage in Hollingshead's (1971:566) commentary on a critique by Haug and Sussman (1971):

"The problem of allocation of a given individual's occupation to a

particular place on the economic scale is occasionally difficult. Haug and

Sussman puzzle over why a correction officer was assigned a rating of 2 while a

policeman rated 5. This particular correction officer was a professional social

worker attached to the juvenile court. He held a Master of Science degree from a

recognized school of social work. Policemen were rated 5 because they are

trained men and were generally regarded in the community as skilled municipal

employees. … The occupational lists are incomplete. No claim has ever been

made by me that they are complete, even for the New Haven community in 1951.

… If we had drawn a different sample in New Haven at that time and interviewed

household members, we would probably have found a number of occupations that

were not drawn in the first sample. I agree that an optimal listing of occupations

should include a wider range than we found in the New Haven community … ."

The Nam-Powers index of socioeconomic status (1963; Nam and Powers 1983; Nam and

Terrie 1982; 1988; Nam et al. 1994) was a far more credible competitor to the Duncan SEI. Nam

and Powers' procedure was to array occupation-industry combinations in Census data in order by

their median educational levels and by their median income levels. They then obtained the

percentile point of each entry in the cumulative distributions of workers and averaged the two

percentile points to obtain index values. The Nam-Powers index has been updated in successive

decennial censuses, and it has been used, for example, in studies of health and mortality

differentials (Rogers et al. 2000). One stated advantage of the Nam-Powers index is that its

metric – relative position in the distribution – is strictly comparable across time, even when the position of specific occupations on the scale has changed, yet that same feature means that the index cannot be used to measure changes in occupation levels across time.

Perhaps because Duncan was such a dominant contributor to research on social stratification, or because of the use of prestige measures in its construction, the Duncan SEI and its variants have been more used in research on social stratification. In many cases, the SEI has been described as a prestige measure by writers who are unfamiliar with the details of its construction. While the SEI became a standard metric for occupational standing in American research on social stratification, it became less useful over time in its original form, and new problems in its use were identified. As Census occupation classifications evolved, it became more difficult to classify occupations into units to which Duncan SEI scores had been assigned. Thus, Stevens and Featherman (1981), Nakao and Treas (1994), and Hauser and Warren (1997) each recreated something like the SEI using contemporary prestige scores for much larger sets of occupation titles as criterion variables. Moreover, it became increasingly clear that indexes based on male workers alone were not valid in analyses of women's occupational attainments, and vice versa (Warren et al. 1998). As in the case of "social class" – and especially because SEI scores were based on the education and income of occupational incumbents – naïve researchers often misused the SEI by assuming that it purported to represent individual or family socioeconomic status *in toto*, rather than occupational standing alone. Finally, Hauser and Warren (1997) demonstrated that, in studies of intergenerational mobility, one did not need a criterion variable like occupational prestige to obtain optimal weights for the aggregate educational and income levels of occupations and, moreover, that the optimal weight for occupational income in such analyses was approximately zero. That is, the main dimension of intergenerational occupational

persistence in the U.S. has been the level of education typical of occupations. In the words of Hauser and Warren (p. 251), "While composite measures of occupational status may have heuristic uses, the global concept of occupational status is scientifically obsolete."

In summary, the story of the Duncan SEI is a case history of the rise and fall of a standard sociological metric. At the time of its development, almost half a century ago, it filled the important unmet need in sociological research for a scalar measure of occupational standing that was defined for each and every occupation title. Moreover, as researchers compared its behavior with that of occupational prestige measures, after the latter had been obtained for all occupations, it became clear that the SEI had substantially greater predictive validity. Thus, the SEI was an important tool in the cumulative body of research on social stratification for about half a century. However, as time went on, the index became obsolete because of changes in occupational classification and in the role of women in the labor force and because new research showed that a single indicator of occupational standing based on the education of occupational incumbents was more valid in many applications than a composite index.

The Duncan SEI was developed for use with an American system of occupational classification and based on the characteristics of male American workers, so it was never given serious consideration as a metric for international comparative studies. However, informed by the greater validity of socioeconomic than of prestige-based occupational scales, Ganzeboom, De Graaf, and Treiman (1992) developed an SEI-like index of occupational standing that is well-suited for comparative work. Rather than using a prestige criterion, they used an optimal scaling procedure in which 271 occupational categories in ISCO were assigned scores to maximize the ability of occupation to mediate the relationship between educational attainment and income in data from 31 surveys in 16 nations from 1968 to 1982. The resulting ISEI (International

Socioeconomic Index of occupational status) was validated in comparison with Treiman's

SIOPS and the Erikson-Goldthorpe class scheme in several new bodies of data. Unlike

Treiman's SIOPS, the ISEI has been well accepted among international scholars.

Normalized Metrics in Comparative Analysis

Transformations and normalizations of metrics are analytic schemes to achieve

comparability in levels or effects. They range from truly useful to utterly misleading.

Log Transformation

One of the simplest and most powerful transformations, under appropriate circumstances,

is the log transformation.[11] In the social sciences, this transformation is most often applied in

analyses of economic amounts (Heckman and Polachek 1974). It also proves useful in

psychophysical measurement (Stevens 1971), and Stevens' work has led to research using

magnitude estimation of attitudes (Hamblin 1971; 1974), utilities (Kemp 1991), linguistic

acceptability (Bard et al. 1996), and the goodness of jobs (Jencks et al. 1988).

The natural log transformation is especially useful in the analysis of economic amounts.

First, the distributions of such amounts tend to be skewed to the right, and the log transform

often yields a reasonable approximation to the normal distribution. Second, a linear relationship

between two log-quantities can be interpreted as elasticity, the percentage change in one quantity

induced by a percentage change in another. Third – and most important in the present context –

the log transform eliminates the original monetary unit. For example, the effect of a year of

education on earnings in Mexico may be expressed in pesos and that in the U.S., in dollars, but in

log pesos or dollars, the two effects are directly comparable. Moreover, one only need

---

[11] There are, of course, many other useful power transformations, and Mosteller and Tukey (1977) provide an excellent guide to choices among them.

exponentiate such estimates to obtain effects in proportional terms. To be sure, comparisons among mean or other summary values of distributional location require use of the original metric of the variables or some linear transformation like purchasing power parity or "the Big Mac Index."

One problem with the log transform is that the quantities to be transformed must be positive. A common, but by no means foolproof solution to this problem, e.g., in the case of zero earnings, is to use a started log, $ln(x + 1)$. A problem here is that, just as the log transform draws in large numbers (at the high end of a distribution), it extends the scale at the low end. Thus, the log transform may replace a distribution that is excessively skewed to the right with one that is excessively skewed to the left. One appropriate solution to this problem is to choose a larger starting value, $ln(x + c)$, where c is chosen to provide a roughly symmetric distribution of the transformed variable.

Scale distortion and Interaction Effects

Measuring a variable on the same scale in two populations does not guarantee that effects of them or on them are comparable. That depends on the functional form of effects and on the way in which the scale was constructed. Consider the effect of years of education on the occupational status of white and black men. Study after study has reported that the returns to schooling are greater in the white than in the black population. Yet this finding may be an artifact of the location of whites and blacks in the educational distribution, that blacks typically have less schooling than whites, in combination with a nonlinear effect of schooling, namely, that a year of schooling beyond high school is more valuable than a year of schooling prior to high school graduation. Among white and black men in the 1973 Occupational Changes in a Generation survey, differential status returns to schooling disappeared when a piecewise linear effect of

schooling was fitted (Hauser et al. 2000:198).

In *On the Success of Failure*, Alexander, Entwisle, and Dauber (1994; 2003) found that

Baltimore first-graders who were held back a year gained more in math achievement in the

following year than students who had been promoted. They offered this as evidence that grade

retention had salutary effects on academic achievement. However, the test on which these groups

of students were compared had been constructed on the assumption that there was a tendency for

achievement to grow more slowly at higher than lower levels of initial achievement. Thus, the

low initial performance of retained students (along with regression effects) all but guaranteed

that they would show large achievement gains in the year following retention. Recalibration of

achievement changes in the metric of grade-equivalents eliminated this effect (Alexander 1998;

Hauser 2005; Shepard et al. 1996; 1998).

Vignette Measurement

One of the mantras of social survey measurement is that you can measure the same thing

by asking the same question in the same way. It isn't necessarily so. One problem here is that

ordered response categories may be located differently by respondents: What is "truly offensive"

to me may be merely "annoying" to you. A classic example of this situation, reported by King et

al. (2004) is of a comparative study of political efficacy in China and Mexico. Taken at face

value, the Chinese considered themselves to have greater efficacy than the Mexicans. King, et al.

developed methods to adjust metrics analytically using vignettes, that is, asking participants in

each nation to rate political efficacy in several sets of hypothetical circumstances, on the

assumption that each participant used the same scale to rate their own efficacy as in the

hypothetical vignettes. After adjustment for responses to the hypothetical vignettes, political

efficacy was rated as higher in Mexico than in China. Such methods have been applied

extensively in international comparisons of health (Salomon et al. 2004).

These effects are not limited to international comparisons. Researchers almost always find that women report more health problems than men, yet the distributions of responses to the standard "general health" question – "In general, would you say your health is excellent, very good, good, fair, or poor?" – are virtually the same among women and men. The 2004-06 round of the Wisconsin Longitudinal Study included an experimental module in which participants were assigned randomly to respond to various vignettes describing general health before they were asked to report their own general health. After adjustment for differential response to the vignette items, women reported poorer general health than men, consistent with findings about specific health conditions (Grol-Prokopczyk et al. 2009).

Normalization of Latent Variables

Problems of establishing comparability in observables, e.g., in the case of the Voluntary National Tests, may be mitigated when the variables to be measured can be treated as indicators of latent variables. Consider the following model:

$$\eta_{il} = \beta_l \xi_{il} + \zeta_{il} \tag{1}$$

$$y_{ikl} = \lambda_{kl}^y \eta_{il} + \varepsilon_{il} \tag{2}$$

$$x_{ijl} = \lambda_{jl}^x \xi_{il} + \delta_{il} \tag{3}$$

where η_{il} and ξ_{il} are latent (unobservable) variables, y_{ikl} and x_{ijl} are observable indicators of η_{il} and ξ_{il}, respectively; β_l, λ_{kl}^y, and λ_{jl}^x are coefficients to be estimated; and ζ_{il}, ε_{il}, and δ_{il} are independently distributed random disturbances. The indexes are $i = 1, …, I$ for individual observations; $j = 1, …, J$ for x-variables; $k = 1, …, K$ for y-variables; and $l = 1, …, L$ for

24

populations. To normalize the metrics of the latent variables, let

$$\lambda^y_{k'l} = \lambda^x_{j'l} = 1 \tag{4}$$

where k' refers to the same y-variable in each population and j' refers to the same x-variable in each population; that is, the same indicator of each unobservable has a slope fixed at 1.0 in each population. Even if no other indicators of x are the same in any two populations and no other indicators of y are the same in any two populations, equation 4 is a sufficient condition to establish comparability of β_l across all populations.

To be sure, this is a less than ideal design. Preferably, one might have the same indicators of each latent variable in each population and be able to show that the conditions

$$\lambda^y_{kl} = \lambda^y_{kl'} \tag{5}$$

and
$$\lambda^x_{jl} = \lambda^x_{jl'} \tag{6}$$

hold across all pairs of populations. With as few as two indicators per variable in each population, conditions 5 and 6 establish over-identifying restrictions on the measurement models of equations 2 and 3.

There are obvious parallels between this scheme and the proposal to use a common set of embedded questions in diverse tests of academic achievement. That is, one might take a summary measure of performance on a set of common, embedded questions to determine the metric of latent academic achievement and use scores on other tests with similar content as additional indicators of the same latent construct. To the extent that those other tests are also used in different populations, one could impose testable restrictions on the model.

Productive Measurement

The preceding example pertains to reflective measurement, that is, where observables are

specified as effects of a latent construct. There are also instances where a metric is constructed as an effect of multiple causes.

One biomedical example, which has also been pursued in social and epidemiological research, is the construct of allostatic load. It is described as an adverse physiological consequence of exposure to the neural or neuroendocrine stress response, that is, as cumulative wear and tear on the body (McEwen 1998; 2000; McEwen and Seeman 1999; McEwen and Stellar 1993; McEwen and Lasley 2002). The measurement problem here is that there is no unique physiological indicator of the construct, so it has been measured in research as a sum of the number of leading indicators of disease processes that exceed specified levels (Singer and Ryff 1999:103). The measure is easy to construct, given the several physiological indicators, but by the same token it is not clear whether each of those indicators actually reflects a unitary stress process, whether unit weights are appropriate, or whether the indicators simply represent specific disease processes.

The concept of comparable worth has played a major part in research and policy about pay equity (England and Dunn 1988; Hartmann 1985). Jobs entail complex mixtures of skills, tasks, responsibility, and environments, so it is daunting to show – without simply accepting prevailing levels of compensation – that one job should be more or less compensated than another. The idea in analyses of comparable worth is to estimate a policy-capturing earnings regression, that is, to write an equation that predicts earnings from a weighted combination of job characteristics that are widely accepted as relevant to compensation. In principle, the regression estimates yield a comparable metric for compensation that can be used to assess the actual compensation levels of population groups or types of employment.

Meta-Analysis and Effect Size

The introduction to Wachter and Straf's (1990:xiii) *Future of Meta-Analysis* provides a succinct definition and description:[12]

> "Meta-analysis refers to the application of quantitative methods to the problem of combining results from different analytic studies. Typically, a statistical analysis is made of numerical summaries of each study. Meta-analysis is not a statistical method *per se*, but rather an orientation toward research synthesis that uses many techniques of measurement and data analysis."

There is a very large and sophisticated body of research about methods of meta-analysis (Cook 1992; Cooper 1998; Cooper and Hedges 1994; Hunter and Schmidt 2004), and a full review of its relationships with the development of common metrics would far exceed the scope of this review.

My take on meta-analysis is that, to paraphrase Schumpeter (1942:13), meta-analysis is the crippled sister of pooled analyses of primary data. My hope is that the enterprise will become less important in proportion to the rise of data-sharing as the norm in social scientific research and the growth in the capability of researchers to obtain and re-analyze data from diverse sources (Committee on Science Engineering and Public Policy et al. 2009; Hauser 1987; National Research Council et al. 2000; National Research Council 1985; National Research Council 2005b).

My focus here is on the use of "effect size" in standard deviation units as the metric of comparisons among studies in meta-analysis, e.g., by Rosenthal (1994). Olkin (1990:6) stated the

---

[12] Cooper (1998:107-8) provides a succinct history of meta-analysis.

issue clearly:

> "The introduction of an effect size had the positive effect of moving us away from p-values and vote counts to parameters and models. This was a critical step forward. However, at times combinations of estimates were made when the underlying populations were different, thereby leading to a variety of biased results."

Cooper (1998:126-53) offers an extensive exegesis of effect size measures before observing, "The most desirable technique for combining results of independent studies is to integrate the raw data from each relevant comparison or estimate of a relationship." Standard deviations (and other standardized measures, like correlations) give the misleading impression that they are in a comparable metric when they are not. They vary from population to population such that effects described in that metric may appear different when they are actually similar, and *vice versa*. Laird (1990:49-50) offers a striking example, that the use of an effect size based on a meta-analysis of aphasia treatment by Greenhouse, et al. (1990) vastly overstates the effectiveness of that treatment in a concrete metric shared by most of the studies. She observes, "In my view, effect sizes are necessary evils which are sometimes unavoidable, but which should be used only when all else fails" (p. 49).

Not only may effect sizes misinform, but they also substitute for and thus discourage the use of concrete metrics that could become standards if they, rather than standard deviation units, were more widely used to report findings. Recall that unemployment rates, official poverty, and achievement levels all have entered public discourse precisely because they are so often used. Where multiple outcome measures have been obtained within the same study, variants of the

measurement model described earlier could be useful in calibrating outcomes in a common, observable metric.

The Bottom of the Barrel: R^2

The coefficient of determination (R^2) shares all of the potential defects of parametric effect sizes in meta-analysis, yet it is often treated as if it were in a comparable metric, perhaps because it varies between 0 and 1, even where variables of interest have actually been measured in the same metric. One pertinent example comes from research on the determinants of educational attainment among women and men in the Wisconsin Longitudinal Study (Sewell et al. 2004). Based upon the coefficient of determination, investigators reported that educational attainment depended more upon social and economic background among women than among men. In fact, socioeconomic origins had larger effects on the educational attainment of men than of women, but the coefficient of determination was larger among women than men because there was less variation in women's attainments.

A more contemporary example, on a larger scale, is an analysis of cross-national differences in the effects of socioeconomic status on school achievement in PISA, the Program for International Student Assessment (Organisation for Economic Co-operation and Development 2007b; a). Figure 4.10 (from the OECD report) plots mean national scores on the science composite by the percentage of variance explained by the PISA SES Index. The text accompanying that figure states:

> "Figure 4.10 highlights that countries differ not just in their overall
> performance, but also in the extent to which they are able to moderate the
> association between socio-economic background and performance. PISA suggests
> that maximising overall performance and securing similar levels of performance

29

among students from different socio-economic backgrounds can be achieved simultaneously. The results suggest therefore that quality and equity need not be considered as competing policy objectives" (Organisation for Economic Co-operation and Development 2007c: 190).

This discussion is problematic because neither axis of the diagram is well-chosen. In the context of the analysis, adjusted rather than observed mean levels of achievement should be used to indicate the quality of science education. In an ideal situation, one would base such an adjustment on a full model of achievement in science – including many more background, parental, and student characteristics beyond economic, social, and cultural status. At the least, the adjustment should take account of national differences in the PISA SES Index. [13] Then, as just explained, the second axis of the graph should be the variance about the regression line, indicating (inversely) how academic performance follows socioeconomic status. This relationship is shown in Figure 2. In that figure, unlike Figure 4.10, the horizontal and vertical lines mark the average values of performance in science and of error variance for all 55 countries, not just the OECD countries.

There is essentially no relationship between observed means and percentages of variance explained in Figure 4.10 (r = -0.04). Thus, the discussion of this figure in the text points to examples of four types of nations, which appear in roughly equal numbers representing the four possible combinations of achievement in science and fit of the regression model. In contrast, there is a moderate relationship between the adjusted means and error variances in Figure 2 (r = 0.34). That is, high performing countries tend to have greater equality of opportunity, in the

---

[13] To be sure, the text recognizes the import of socioeconomic background for science achievement, and a consistent analysis would have taken that into account in the construction of Figure 4.10.

sense that the scatter of individual observations about the regression line is greater, while low

performing countries tend to have less equality of opportunity, less dispersion of individual

observations about the regression line. Again, the position of nations on the vertical axis of

Figure 2 (science achievement) is similar to that in Figure 4.10, with the exceptions noted above,

but as shown in Figure 1, there is very little relationship between the percentages of explained

variance and the variances of observations about the regression of science achievement on the

PISA SES Index.[14]

Figure 2 thus offers a very different picture from Figure 4.10 of the relationship between

educational opportunity – lack of fit to the regression line – and national levels of academic

performance in science. For example, in Figure 4.10, the United States appears near the center,

slightly below the OECD average in science achievement and somewhat above average in

percentage of variance explained. In Figure 2, the U.S. is slightly above average in science

achievement (for all nations) and far above average in equality of educational opportunity, for

there is a relatively high level of scatter of science achievement about the values predicted from

the PISA SES Index. Why is a high percentage of variance explained in the U.S.? The variation

in the PISA SES Index in the U.S. is the same as the OECD average, but the regression of

science achievement on the PISA SES Index (49) is almost 25 percent above than the OECD

average (40) (Organisation for Economic Co-operation and Development 2007a: Table 4.4a, pp.

123-24). Thus, the U.S. performs badly on one indicator of educational opportunity (the

regression coefficient), but far better on another indicator, the scatter of individual student

achievement about values predicted from socioeconomic background. Israel appears as slightly

---

[14]Since the X-axis of Figure 4.10 goes from high to low percentages of explained variance, while the X-axis of Figure 2 goes from low to high estimates of error variance, the spatial representation of effects in the two diagrams is the same. The strength of the relationship between the PISA SES Index and science achievement declines from left to right.

below average both in science achievement and in the impact of socioeconomic background in Figure 4.10, but Figure 2 shows Israel as far below average in the impact of socioeconomic background. Bulgaria appears as below average in science achievement in both figures, but it is depicted as having very high dependence of science achievement on social background in Figure 4.10 and moderately low dependence of achievement on background in Figure 2. Plainly, it is possible to add to these examples of divergent findings.

One might imagine adding the regression coefficient of science achievement on the PISA SES Index as a third dimension of the display. In this way both aspects of the dependence of science achievement on SES would be represented, but the findings would not be confounded by statistically (though not substantively) irrelevant differences in the variability of socioeconomic background. Unfortunately, it is not possible to distinguish between the effects of these two variables (the regression slope and the error variance) on mean country achievement levels. The correlation between the two is moderately high (r = 0.70), while their correlations with mean science achievement are similar (0.35 and 0.34, respectively). That is, the error variances are *larger* in countries with steeper slopes of science achievement on the PISA SES Index. In a regression analysis of the adjusted means, the slope coefficient dominates, but there is actually no significant difference between the effects of the two explanatory variables. In other words, data are not available for a large enough number of countries to identify significant differences between the associations of the achievement-SES slope and the error variance with adjusted country means.

Discussion

The preceding examples suggest a few rules of thumb and *caveats*, but no hard and fast rules for the creation of sound, standard and comparable social, economic, and behavioral measures:

- Repeated use gives meaning to a metric; unfortunately, overuse may reify it.

- Meet a real scientific and/or policy need. If no one else will use a measure, it's not worth the effort. Widespread use truly is rewarding; check the citation indexes.

- Seek simplicity in content and construction. To the extent that an indicator is hard to ascertain, complicated to construct, and admits multiple interpretations, it will be less useful.

- Avoid relative measurement: Above all, avoid percentile ranks, standard deviations, and shares of variance.

- Avoid descriptive terms for arbitrarily or subjectively determined ranges of a quantitative indicator. Such terms invite misinterpretation.

- Study the operational and analytic behavior of a measure to assess its validity, and not merely the details of its construction.

- Weigh the balance between internal and external validity. Information loss may vary positively with comparability, and, sometimes, loss is gain.

There is no more important and scientifically rewarding task than the development of standard metrics that will be useful in theory and in practice. This selective review of past efforts provides a far more cautionary account of the prospects for useful and valid common metrics in social science than I had intended and expected to offer. The lesson, I think and hope, is not that such metrics are out of reach, but that creating them requires a great deal of thought and sustained

effort. The balancing act among validity, generality, accessibility, and sustainability is delicate indeed.

References

Alexander, Karl L. 1998. "Letter to the Editor." *Psychology in the Schools* 35:402-04.

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber. 1994. *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary Grades*. Cambridge: Cambridge University Press.

—. 2003. *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary Grades*. Cambridge: Cambridge University Press.

Bard, EG, D Robertson, and A Sorace. 1996. "Magnitude Estimation of Linguistic Acceptability." *Language* 72:32-68.

Blau, Peter M. and Otis Dudley Duncan. 1967. *The American Occupational Structure*. New York: John Wiley and Sons.

Bourque, Mary Lyn. 2004. "A History of the National Assessment Governing Board." Pp. 201-31 in *The Nation's Report Card: Evolution and Perspectives*, edited by L. V. Jones and I. Olkin. Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Brinkley, Alan. 2001. "Employment Act of 1946." in *The Oxford Companion to United States History*, edited by P. S. Boyer. Oxford: Oxford University Press.

Committee on Science Engineering and Public Policy, National Academy of Sciences, Division of Policy and Global Affairs, and Institute of Medicine. 2009. *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. Washington, DC: Natl. Academies Press.

Cook, Thomas D. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Cooper, Harris M. 1998. *Synthesizing Research: A Guide for Literature Reviews*. Thousand Oaks, Calif.: Sage Publications.

Cooper, Harris M. and Larry V. Hedges. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Counts, GS. 1925. "The Social Status of Occupations: A Problem in Vocational Guidance." *The School Review* 33:16-27.

Coxon, Anthony , P. M. Davies, and Charles L. Jones. 1978. *The Images of Occupational Prestige: A Study in Social Cognition*. New York: St. Martin's Press.

Duncan, OD and RW Hodge. 1963. "Education and Occupational Mobility a Regression Analysis." *The American Journal of Sociology* 68:629-44.

Duncan, Otis Dudley. 1961. "A Socioeconomic Index for All Occupations." Pp. 109-38 in *Occupations and Social Status*, edited by A. J. Reiss, Jr. New York: Free Press.

—. 1984. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.

England, Paula and Dana Dunn. 1988. "Evaluating Work and Comparable Worth." *Annual review of sociology* 14:227-48.

Erikson, Robert and John H. Goldthorpe. 1985. "A Model of Core Social Fluidity in Industrial Nations." Institut fur Sozialwissenschaften Universitat Mannheim.

—. 1987a. "Commonality and Variation in Social Fluidity in Industrial Nations. Part I: A Model for Evaluating the `Jfh Hypothesis'." *European Sociological Review* 3:54-77.

—. 1987b. "Commonality and Variation in Social Fluidity in Industrial Nations. Part Ii: The Model of Core Social Fluidity Applied." *European Sociological Review* 3:145-66.

—. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: The Clarendon Press.

Erikson, Robert, John H. Goldthorpe, and Lucienne Portocarero. 1982. "Social Fluidity in Industrial Nations: England, France and Sweden." *British Journal of Sociology* 33:1-34.

Featherman, David L. and Robert M. Hauser. 1978. *Opportunity and Change*. New York: Academic Press.

Featherman, David L., F. Lancaster Jones, and Robert M. Hauser. 1975. "Assumptions of Social Mobility Research in the U.S.: The Case of Occupational Status." *Social Science Research* 4:329-60.

Ganzeboom, Harry B., Paul M. De Graaf, and Donald J. Treiman. 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21:1-56.

Glaser, Robert, Robert L. Linn, and George W. Bohrnstedt. 1993a. "Setting Performance Standards for Student Achievement." National Academy of Education, Stanford, CA.

—. 1993b. "The Trial State Assessment: Prospects and Realities." National Academy of Education, Stanford, CA.

Greenhouse, Joel B., Davida Fromm, Satish Iyengar, Mary Amanda Dew, Audrey L. Holland, and Robert E. Kass. 1990. "The Making of a Meta-Analysis: A Quantitative Review of the Aphasia Treatment Literature." Pp. 29-46 in *The Future of Meta-Analysis*, edited by K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.

Grol-Prokopczyk, Hanna, Jeremy Freese, and Robert M. Hauser. 2009. "Anchors—a Way? Using Anchoring Vignettes to Assess Group Differences in Self-Rated Health." Center for Demography and Ecology, University of Wisconsin-Madison. Madison, Wisconsin.

Halaby, Charles N. 1993. "Reply to Wright." *American Sociological Review* 58:35-36.

Halaby, Charles N. and David L. Weakliem. 1993. "Ownership and Authority in the Earnings Function: Alternative Specifications." *American Sociological Review* 58:16-30.

Haller, AO and DB Bills. 1979. "Occupational Prestige Hierarchies: Theory and Evidence." *Contemporary Sociology* 8:721-34.

Haller, AO and DB Holsinger. 1972. "Variations in Occupational Prestige Hierarchies: Brazilian Data." *The American Journal of Sociology* 77:941-56.

Haller, AO and DM Lewis. 1966. "The Hypothesis of Intersocietal Similarity in Occupational Prestige Hierarchies." *American Journal of Sociology*:210-16.

Hamblin, RL. 1971. "Ratio Measurement for the Social Sciences." *Social Forces* 50:191-206.

—. 1974. "Social Attitudes: Magnitude Measurement and Theory." *Measurement in the social sciences: Theories and strategies*:61-120.

Hartmann, Heidi I. 1985. *Comparable Worth: New Directions for Research*. Washington, D.C.: National Academy Press.

Haug, Marie R. and Marvin B. Sussman. 1971. "The Indiscriminate State of Social Class Measurement." *Social Forces* 49:549-63.

Hauser, Philip M. 1964. "Labor Force." Pp. 160-90 in *Handbook of Modern Sociology*, edited by R. E. L. Faris. Chicago: Rand McNally & Co.

Hauser, Robert M. 1982. "Occupational Status in the 19th and 20th Centuries." *Historical Methods* 15:111-26.

—. 1987. "Sharing Data: It's Time for Asa Journals to Follow the Folkways of a Scientific Sociology." *American Sociological Review* 52.

—. 2005. "K.L. Alexander, D.R. Entwisle, S.L. Dauber, on the Success of Failure, a Reassessment of the Effects of Retention in the Primary School Grades, 2nd Edition." *Journal of School Psychology* 43:87-94.

Hauser, Robert M. and David L. Featherman. 1977. *The Process of Stratification: Trends and Analyses*. New York: Academic Press.

Hauser, Robert M. and John Allen Logan. 1992. "How Not to Measure Intergenerational Occupational Persistence." *American Journal of Sociology* 97:1689-711.

Hauser, Robert M. and John Robert Warren. 1997. "Socioeconomic Indexes for Occupations: A Review, Update, and Critique." Pp. 177-298 in *Sociological Methodology 1997*, edited by A. E. Raftery. Cambridge: Basil Blackwell.

Hauser, Robert M., John Robert Warren, Min-Hsiung Huang, and Wendy Y. Carter. 2000. "Occupational Status, Education, and Social Mobility in the Meritocracy." Pp. 179-229 in *Meritocracy and Economic Inequality*, edited by K. Arrow, S. Bowles, and S. Durlauf. Princeton: Princeton University Press.

Heckman, James and Solomon Polachek. 1974. "Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship." *Journal of the American Statistical Association* 69:350-54.

Hershberg, Theodore, Michael Katz, Lawrence Glasco, Stuart Blumin, and Clyde Griffen. 1974. "Occupation and Ethnicity in Five Nineteenth Century Cities: A Collaborative Inquiry." *Historical Methods Newsletter* 7:174-216.

Hodge, Robert W., Paul M. Siegel, and Peter H. Rossi. 1964. "Occupational Prestige in the United States, 1925-63." *American Journal of Sociology* 70:286-302.

Hollingshead, August B. 1957. "Two Factor Index of Social Position." New Haven, Connecticut: Yale University.

—. 1971. "Commentary on 'the Indiscriminate State of Social Class Measurement'." *Social Forces* 49:563-67.

Hollingshead, August B. and F. C. Redlich. 1958. *Social Class and Mental Illness*. New York: Wiley.

Hout, Michael and Robert M. Hauser. 1992. "Symmetry and Hierarchy in Social Mobility: A Methodological Analysis of the Casmin Model of Class Mobility." *European-Sociological-Review; 1992, 8, 3, Dec, 239-266: European-Sociological-Review*.

Hunter, John E. and Frank L. Schmidt. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, Calif.: Sage.

Inkeles, A and PH Rossi. 1956. "National Comparisons of Occupational Prestige." *The American Journal of Sociology* 61:329-39.

Jencks, Christopher S., Lauri Perman, and Lee Rainwater. 1988. "What Is a Good Job? A New Measure of Labor Market Success." *American Journal of Sociology* 93:1322-57.

Jones, Lyle V. and Ingram Olkin. 2004. "The Nation's Report Card: Evolution and Perspectives." Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

Kemp, S. 1991. "Magnitude Estimation of the Utility of Public Goods." *Journal of applied Psychology* 76:533-40.

King, Gary, Cristopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Survey Research." *American Political Science Review* 98:191-207.

Kirsch, Irwin S., Educational Testing Service, and National Center for Education Statistics. 1993. *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Washington, D.C.: Office of Educational Research and Improvement [Supt. of Docs., U.S. G.P.O., distributor.

Kirsch, Irwin S. and Andrew J. Kolstad. 2001. *Technical Report and Data File User's Manual for the 1992 National Adult Literacy Survey*. Jessup, MD: U.S. Dept. of Education, Office of Educational Research and Improvement.

Klatzky, Sheila R. and Robert W. Hodge. 1971. "A Canonical Correlation Analysis of Occupational Mobility." *Journal of the American Statistical Association* 66:16-22.

Kraus, Vered, E. O. Schild, and Robert W. Hodge. 1978. "Occupational Prestige in the Collective Conscience." *Social Forces* 56:900-18.

Krieger, Nancy, David R. Williams, and Nancy E. Moss. 1997. "Measuring Social Class in U.S. Public Health Research: Concepts, Methodologies, and Guidelines." Pp. 341-78, vol. 18, edited by J. E. Fielding, L. B. Lave, and B. Starfield. Palo Alto, California: Annual Reviews.

Laird, Nan. 1990. "A Discussion of the Aphasia Study." Pp. 47-52 in *The Future of Meta-Analysis*, edited by K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.

Liberatos, Penny, Bruce G. Link, and Jennifer L. Kelsey. 1988. "The Measurement of Social Class in Epidemiology." *Epidemiological Reviews* 10:87-121.

Linn, RL. 2003. "Accountability: Responsibility and Reasonable Expectations." *Educational Researcher* 32:3.

McEwen, B. 2000. "Allostasis and Allostatic Load Implications for Neuropsychopharmacology." *Neuropsychopharmacology* 22:108-24.

McEwen, Bruce S. and Elizabeth Norton Lasley. 2002. *The End of Stress as We Know It*. Washington, D.C.: Joseph Henry Press.

McEwen, BS. 1998. "Protective and Damaging Effects of Stress Mediators." *The New England Journal of Medicine* 338:171.

McEwen, BS and T Seeman. 1999. "Protective and Damaging Effects of Mediators of Stress: Elaborating and Testing the Concepts of Allostasis and Allostatic Load." *ANNALS-NEW YORK ACADEMY OF SCIENCES* 896:30-47.

McEwen, BS and E Stellar. 1993. "Stress and the Individual: Mechanisms Leading to Disease." *Archives of Internal Medicine* 153:2093.

Miech, Richard A. and Robert M. Hauser. 2001. "Socioeconomic Status (Ses) and Health at Midlife: A Comparison of Educational Attainment with Occupation-Based Indicators." *Annals of Epidemiology* 11:75-84.

Mosteller, Frederick and John W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, Massachusetts: Addison-Wesley.

Nakao, Keiko and Judith Treas. 1994. "Updating Occupational Prestige and Socioeconomic Scores: How the New Measures Measure Up." Pp. 1-72 in *Sociological Methodology 1994*, edited by P. V. Marsden. Washington, D.C.: American Sociological Association.

Nam, Charles B. 1963. *Methodology and Scores of Socioeconomic Status*. Washington, D.C.: U.S. Bureau of the Census.

Nam, Charles B. and Mary G. Powers. 1983. *The Socioeconomic Approach to Status Measurement (with a Guide to Occupational and Socioeconomic Status Scores).* Houston: Cap and Gown Press.

Nam, Charles B. and E. Walter Terrie. 1982. "Measurement of Socioeconomic Status from United States Census Data." Pp. 29-42, edited by M. G. Powers. Boulder, Colorado: Westview Press.

—. 1988. "1980-Based Nam-Powers Occupational Status Scores." Tallahassee, Florida: Center for The Study of Population Florida State University.

Nam, Charles B., E. Walter Terrie, and Carl P. Schmertmann. 1994. "Comparison of the 1980 Updated Duncan and Nam-Powers Occupational Scores." Tallahassee, Florida: Center for The Study of Population Florida State University.

National Research Council. 2005a. "Measuring Literacy: Performance Levels for Adults." edited by R. M. Hauser, C. Edley, J. A. Koenig, and S. Elliot. Washington, DC: National Academies Press.

National Research Council, Committee on National Statistics, and Commission on Behavioral and Social Sciences and Education. 2000. *Improving Access to and Confidentiality of Research Data: Report of a Workshop*, Edited by C. D. Mackie and N. M. Bradburn. Washington, DC: National Academy Press.

National Research Council, Committee on Embedding Common Test Items in State and District Assessments, Board on Testing and Assessment. 1999a. *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests*, Edited by D. M. Koretz, M. W. Bertenthal, and B. F. Green. Washington, DC: National Academy Press.

National Research Council, Committee on Equivalency and Linkage of Educational Tests, Board on Testing and Assessment. 1999b. *Uncommon Measures: Equivalence and Linkage among Educational Tests*, Edited by M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, and F. C. Hemphill. Washington, DC: National Academy Press.

National Research Council, Committee on Evaluation of National and State Assessments of Education Progress, Board on Testing and Assessment. 1999c. *Grading the Nation's Report Card Evaluating Naep and Transforming the Assessment of Educational Progress*, Edited by J. W. Pellegrino, L. R. Jones, and K. J. Mitchell. Washington, D.C.: National Academy Press.

National Research Council, Committee on National Statistics. 1985. *Sharing Research Data*, Edited by S. E. Fienberg, M. E. Martin, and M. L. Straf. Washington, D.C.: National Academy Press.

National Research Council, Panel on Data Access for Research Purposes, Committee on National Statistics. 2005b. *Expanding Access to Research Data : Reconciling Risks and Opportunities*. Washington, DC: National Academies Press.

National Research Council, Panel on Poverty and Family Assistance. 1995. *Measuring Poverty: A New Approach*. Edited by C. F. Citro and R. T. Michael. Washington, D.C.: National Academy Press.

Olkin, Ingram. 1990. "History and Goals." Pp. 3-10 in *The Future of Meta-Analysis*, edited by K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.

Organisation for Economic Co-operation and Development. 2007a. *Pisa 2006: Science Competencies for Tomorrow's World*, Vol. 2, Data. Paris: OECD.

—. 2007b. *Pisa 2006: Science Competencies for Tomorrow's World*, Vol. 1, Analysis. Paris: OECD.

—. 2007c. "Quality and Equity in the Performance of Students and Schools." Pp. 169-212 in *Pisa 2006: Science Competencies for Tomorrow's World*, vol. 1, Analysis. Paris: OECD.

Reiss, Albert J., Jr. 1961. *Occupations and Social Status*. New York: Free Press of Glencoe.

Rogers, Richard G., Robert A. Hummer, and Charles B. Nam. 2000. *Living and Dying in the USA: Behavioral, Health, and Social Differentials of Adult Mortality*. San Diego: Academic Press.

Rosenthal, Robert. 1994. "Parametric Measures of Effect Size." Pp. 231-44 in *The Handbook of Research Synthesis*, edited by H. M. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Rytina, Steve. 1989. "Life Chances and the Continuity of Rank: An Alternative Interpretation of Mobility Magnitudes over the Life Cycle." *American Sociological Review* 54:910-28.

—. 1992a. "Response to Hauser and Logan and Grusky and Van Rompaey." *American Journal of Sociology* 97:1729-48.

—. 1992b. "Scaling the Intergenerational Continuity of Occupation: Is Occupational Inheritance Ascriptive after All?" *American Journal of Sociology* 97:1658-88.

Salomon, Joshua A., Ajay Tandon, and Christopher J. L. Murray. 2004. "Comparability of Self Rated Health: Cross Sectional Multi-Country Survey Using Anchoring Vignettes." *BMJ* 328:258-0.

Schumpeter, Joseph Alois. 1942. *Capitalism, Socialism, and Democracy*. New York, London,: Harper & Brothers.

Sewell, William H., Robert M. Hauser, Kristen W. Springer, and Taissa S. Hauser. 2004. "As We Age: The Wisconsin Longitudinal Study, 1957-2001." Pp. 3-111 in *Research in Social Stratification and Mobility*, vol. 20, edited by K. Leicht. London: Elsevier.

Shepard, Lorrie A., Mary Lee Smith, and Scott F. Marion. 1996. "Failed Evidence on Grade Retention." *Psychology in the Schools* 33:251-61.

—. 1998. "On the Success of Failure: A Rejoinder to Alexander." *Psychology in the Schools* 35:404-07.

Siegel, Paul M. 1970. "Occupational Prestige in the Negro Subculture." *Sociological Inquiry* 40:156-71.

—. 1971. "Prestige in the American Occupational Structure." University of Chicago.

Singer, Burton and Carol D. Ryff. 1999. "Hierarchies of Life Histories and Associated Health Risks." *Ann NY Acad Sci* 896:96-115.

Smith, M. 1943. "An Empirical Scale of Prestige Status of Occupations." *American Sociological Review* 8:185-92.

Stevens, Gillian and David L. Featherman. 1981. "A Revised Socioeconomic Index of Occupational Status." *Social Science Research* 10:364-95.

Stevens, SS. 1971. "Issues in Psychophysical Measurement." *Psychological review* 78:426-50.

Treiman, DJ. 1976. "A Standard Occupational Prestige Scale for Use with Historical Data." *Journal of Interdisciplinary History* 7:283-304.

Treiman, Donald J. 1975. "Problems of Concept and Measurement in the Comparative Study of Occupational Mobility." *Social Science Research* 4:183-230.

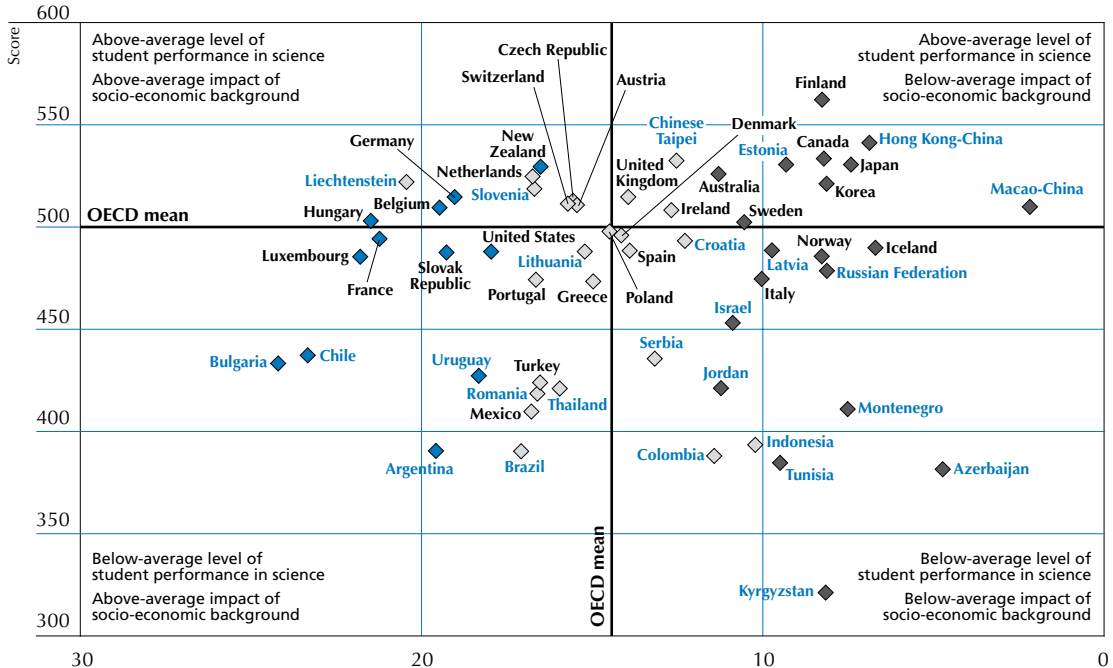—. 1977. *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

United States. National Commission on Employment and Unemployment Statistics. 1979.

    Counting the Labor Force. Washington, D.C.: Supt. of Docs., U.S. Govt. Print. Off.

Wachter, Kenneth W. and Miron L. Straf. 1990. *The Future of Meta-Analysis*. New York:

    Russell Sage Foundation.

Warren, John Robert, Jennifer T. Sheridan, and Robert M. Hauser. 1998. "Choosing a Measure

    of Occupational Standing: How Useful Are Composite Measures in Analyses of Gender

    Inequality in Occupational Attainment?" *Sociological Methods and Research* 27:3-76.

Wright, Erik Olin. 1993. "Typologies, Scales, and Class Analysis: A Comment on Halaby and

    Weakliem." *American Sociological Review* 58:31-34.

—. 1997. *Class Counts: Comparative Studies in Class Analysis*. Cambridge: Cambridge

    University Press.

Figure 4.10

Performance in science and the impact of socio-economic background

Average performance of countries on the PISA science scale and the relationship between performance and the PISA index of economic, social and cultural status

◆ Strength of the relationship between performance and socio-economic background **above** the OECD average impact

◇ Strength of the relationship between performance and socio-economic background **not statistically significantly different from** the OECD average impact

◆ Strength of the relationship between performance and socio-economic background **below** the OECD average impact

Note: OECD mean used in this figure is the arithmetic average of all OECD countries.

Source: OECD PISA 2006 database, Table 4.4a.

StatLink ᗘᖴᔌ http://dx.doi.org/10.1787/141848881750

Figure 1. Error Variance in Science Achievement by Percentage of Variance Explained
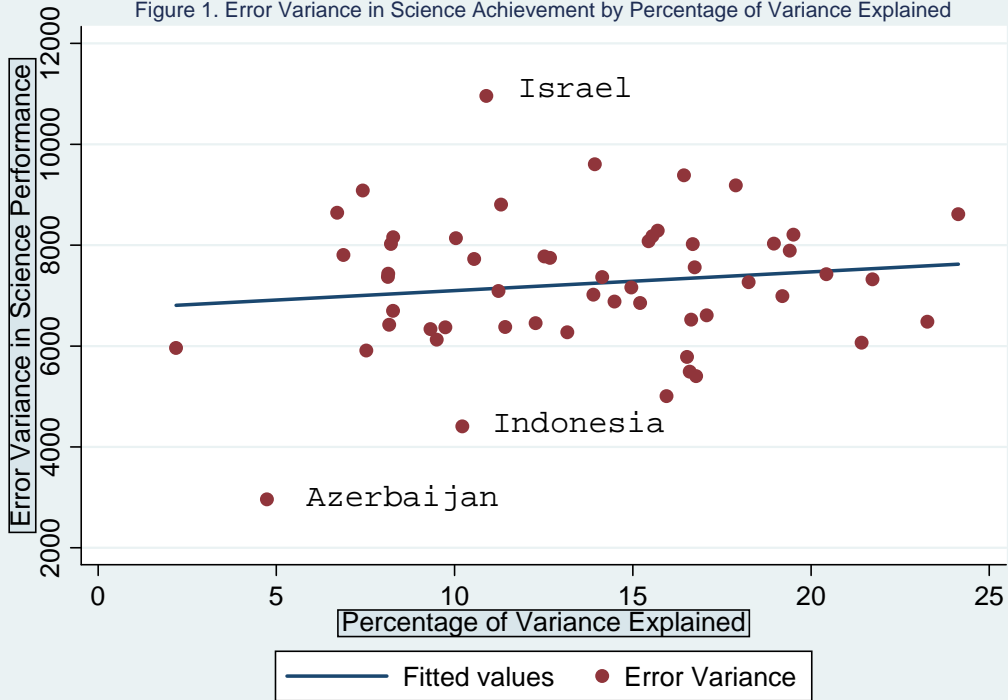
Figure 2. Performance in Science (SP) by Error Variance in Background Regression