MEASURING VARIABLES

Suppose we have formulated some research problem in terms of a *proposition* (or *expectation* or *hypothesis*) explicitly stating a *relationship* or *association* between two *variables* that pertain and some *unit of analysis*. Consider propositions #1 and #16 in Problem Set #3, which we can state more formally as follows:

#1 DEGREE OF SENIORITY is associated with DEGREE OF PRAGMATISM among members of Congress.

#16 DEGREE OF PROPORTIONALITY is associated with NUMBER OF POLITICAL PARTIES among the electoral systems of the world. [Duverger's Law]

In order to test such hypotheses, we need to collect empirical data (using a survey, historical records (of election, Congressional roll calls, etc.) reference books, or whatever). But before we can start collecting data we need to devise some way of using data to actually *measure* the variables that we have conceptually identified. This requires us to establish some kind of *linkage* or *correspondence* between the data that we will collect and the "conceptual" (or "abstract") variables in the hypothesis. For some variables, the measurement problem is quite straightforward, but for other variables, it can be very difficult. In social science and political science especially, appropriate data often comes from surveys but, as previously noted, it may well come from other sources as well, e.g., documents, historical records, direct (participant or non-participant) observation, etc. (In doing Problem Set #4, students too often assume that data used to measure variables must come from surveys.)

The problem of measuring variables entails two distinct though interrelated problems that may be referred to as *coding* and *operationalization*. The coding problem is relatively straightforward; it entails determining what the variable will "look like" in a codebook, i.e., its name, description, and (in particular) its range of possible values (in the manner of Problem Set #3A). The operationalization problem is more difficult — and often *much* more difficult; it entails specifying the *practical operations* that will be used to "observe" or "measure" the actual value of the variable in each case, so that data is appropriately collected and (if necessary) coded.

In sentences #2 and #11 of Problem Set #3A, LEVEL OF EDUCATION (pertaining to individuals) is one of the variables that has been identified. Let us consider how this variable might be measured.

Coding the Variable LEVEL OF EDUCATION

With respect to coding, we must decide what the *range of possible values* will look like (and whether these possible values will result in a *nominal*, *ordinal*, *interval*, or *ratio* variable). Here are some possibilities:

(1) LEVEL OF EDUCATION (*dichotomous*):
- (A) Low
- (B) High

(2) LEVEL OF EDUCATION (*qualitative/ordinal but "imprecise"*):
- (C) Low
- (D) Medium
- (E) High

(3) LEVEL OF EDUCATION: HIGHEST LEVEL ATTAINED (*qualitative/ordinal and more "precise"*)
- (A) Grade school
- (B) Middle school
- (C) Some highschool
- (F) Highschool graduate
- (G) Some college
- (H) College graduate
- (I) Some graduate/professional school
- (J) Graduate/professional degree

(4) LEVEL OF EDUCATION: NUMBER OF YEARS OF FORMAL EDUCATION (*quantitative/interval*):

ACTUAL NUMBER OF YEARS OF FORMAL EDUCATION

(5) LEVEL OF EDUCATION (*quantitative/interval*):

[SOME KIND OF NUMERICAL SCALE DERIVED FROM A TEST (perhaps 0-100)]

Operationalizing the Variable LEVEL OF EDUCATION

With respect to operationalization, we must decide what practical operations we will use to assign a particular value of LEVEL OF EDUCATION to a particular case. If we are doing survey research, we presumably will determine a respondent's level of education simply by asking an appropriately phrased question and recording the response (hoping that the responses are more or less truthful). (See variable V62 in the SETUPS/NES Codebook; the full-scale NES uses this basic question plus several follow-up questions.) But in other circumstances — for example if we were looking at student testing data — we might use documentary records. Note that if we use simple LO, MED, HI categories, we must decide on the cutoff points separating these categories in terms of years or grade levels. (Probably we should not use such crude categories at the outset but rather record the more *precise* categories of grade level or years. We might subsequently have reason to *recode* the variable into LO, MID, HI categories. See below for a discussion of *precision* and *recoding*.)

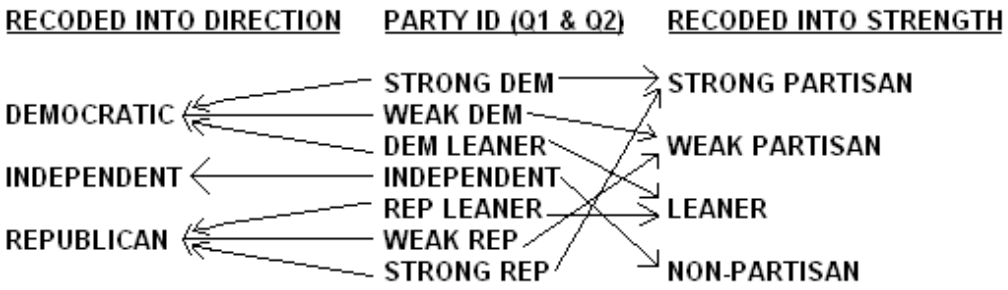
Other Examples for Class Discussion **[from PS #3A]**

LEVEL OF SENIORITY (#1)
DEGREE OF PRAGMATISM (#1)
DEGREE OF RELIGIOSITY (#2)
DIRECTION OF IDEOLOGY or DEGREE OF LIBERALISM/CONSERVATISM (sentence #14 or Congressional study)
DEGREE OF PROPORTIONALITY (#16)
NUMBER OF POLITICAL PARTIES (#16)

Survey Data and Recoding

In the context of survey research, each question is a variable, and operationalizing such variables is a matter of designing questions and their options (for a closed-form question) or coding schemes for responses (for an open-ended question) and then asking respondents to answer the questions. However, other variables may be constructed by combining questions, combining or recoding options, forming indexes, etc.

For example, (following NES) the Student Survey had a question (Q1) and a follow-up question (Q2) pertaining to Party Identification. As previously discussed in class, they are *combined* to form the standard seven-category DIRECTION AND STRENGTH OF PARTY IDENTIFICATION measure. This then may be *recoded* back into a three-category DIRECTION (but not STRENGTH) OF PARTY IDENTIFICATION measure or into a four-category STRENGTH (but not DIRECTION) OF PARTY IDENTIFICATION measure. (Typically all this excludes "other/minor party" or "DK/apolitical" responses. SPSS allows you to recode variables; see Section VIII of the SPSS handout.)



Criteria for Measuring Variables

In general, there is no "best" way to measure a variable in social science research. (Sometimes there isn't even a "pretty good" way.) But usually there are one or two reasonably good ways and many very bad ways. Here are some considerations.

1. You need to be *conceptually clear* about what it is you are you are trying to measure. For example, you may be doing research about rank and file members of the VAP and you are interested in aspects of their partisanship. But you need to be clear about what it is that you have in mind: (1) how people ordinarily vote; (2) how people voted in the most recent election (and for what offices); (3) how people intend to vote in the upcoming election; (4) how people are registered to vote (in states that have registration by party); (5) whether people actually belong to party organizations, clubs, etc.; and (6) how people think about themselves in party terms. We have a number of different concepts of partisanship floating around here, and we need to be clear which one we are actually interested in and trying to measure. (The standard NES PARTY IDENTIFICATION measure is intended to measure the last concept of partisanship, i.e., psychological identification.) Of course, we may be interested in several of these concepts (we might be interested in knowing how frequently people who think of themselves as Republican or Democrats actually register otherwise), so we may need to design several measures for several distinct concepts of partisanship.

2. Thinking about variables — especially non-dichotomous ordinal or interval level variables — encourage us to think in *dimensional* terms (low to high, small to big, left to right, etc.). Multidimensional concepts should usually be broken up into one-dimensional variables and measures. For example, IDEOLOGY (Liberal - Conservative) might better be two or more variables, e.g., ECONOMIC LIBERALISM/CONSERVATISM, SOCIAL LIBERALISM /CONSERVATISM; likewise POLITICAL EFFICACY might better be two variables: INTERNAL POLITICAL EFFICACY (based on sense of "self) vs. EXTERNAL POLITICAL EFFICACY (based on sense of "the system"). However, sometimes a single variable and measure can represent a multidimensional concept. An example is the standard NES seven-point PARTY IDENTIFICATION scale discussed above, which combines DIRECTION of Party ID (Democrat vs. Republican) with STRENGTH of Party ID (leaner to strong). Typical survey questions on issues for which the possible responses (values) run from "strongly agree" through indifference to "strongly disagree" likewise combine direction and strength of opinion. These examples "work" as single variables only because the U.S. has a two-party (vs. multi-party) system and because most issues are "two-sided" (vs. multi-sided).

3. Much empirical research involves studying relationships between two (or more) variables (as suggested in PS #3A). This requires that each variable be *independently* measured; otherwise, a hypothesis becomes a *tautology* (a statement that is true by definition) masquerading as an empirical proposition (Recall our class discussion on operationalizing DEGREE OF RELIGIOSITY using SETUPS data, and also see point #6 immediately below; now consider the sentence "More religious people attend church (etc.) more frequently than less religious people.")

4. It generally makes sense to use *standard* or *conventional* measures when they are available. For example, if you were to do a small local survey inquiring about respondents' party identification (among other things), it would probably be desirable to use the standard NES measure of party identification because: (1) this will be generally *easier* than devising your

own measure; (2) you know the NES survey question has been extensively *pretested*; (3) you will not need to describe the measure in detail in your research report (you can simply cite NES); (4) and your results will then be *comparable* to NES (and many other) studies. (A norm is developing among quantitative political scientists that data sets used in a paper, article, or book should be made readily available to readers and other researchers, so that others can *replicate* [double-check] the analysis. The fact the data sets are now computer files that can be distributed over the internet makes such a policy much more feasible that it was a decade or more ago.)

5. Measures of variables should always be *public* and *reproducible* (by others) Every term paper, journal article, book, etc., reporting the results of empirical research *should describe clearly how the variables were measured*. This may be done in a special methodology section or chapter, an appendix, or (if you used standard measures) by appropriate citation. Unless standard questions were used, a research report based on survey data should include a verbatim transcript of the relevant questions in the survey questionnaire.

6. Sometimes you will need to use an *indirect measure* (or *indicator* or *proxy)* of a variable that has not been — and perhaps cannot be — measured directly. It may be difficult to directly measure the DEGREE OF RELIGIOSITY of individuals. But we can measure their FREQUENCY OF CHURCH (etc.) ATTENDANCE more readily. It seems plausible that these two variables are closely related. So we may choose the let the second (easier-to- measure) variable "stand in" as an indicator of or proxy for the first (harder-to-measure) variable. Of course, there is a sense in which *all* survey variables are merely indicators of the variables we are truly interested in, because all survey research takes *responses to questions* to be more or less accurate indicators of *unobserved behavior* (e.g., whether and how respondents voted) or (even more commonly) *unobservable attitudes and opinions*.

7. It is often desirable to use a *composite* measure or *index*. For example, if we want to measure LEVEL OF POLITICAL TRUST/CYNICISM in individuals by means of a survey, a single question is unlikely to be helpful. Certainly a question like "How cynical are you about the political process in this country?" is unlikely to produce useful data. A single question like Q21 in the Student Survey is likely to be *unreliable* (see point #2 below), since people may respond on the basis of idiosyncratic circumstances of the moment, rather than on the basis of their more general and enduring dispositions. But if respondents answer in consistently cynical or trusting ways over a variety of related questions (like Student Survey Q20-22), we are likely to be more confident that the overall pattern of responses indicates something meaningful about their LEVEL OF POLITICAL CYNICISM. If you measure a variable by means of an index, it is necessary not only to describe how the individual measures are constructed but also to specify the rule of composition by which they are combined into a single overall measure. Because it is a combination of several or many measures, an index is often not expressed in any particular *unit of measurement* but in terms of some arbitrary *index number* or *score* (e.g., 1-5 [usually deemed to be an interval measure]).

Accuracy in Measurement

Of course, we want our measures of variables to be as accurate as possible — that is, we want a close correspondence between our measures and the concepts of interest. Accuracy in measurement is usefully subdivided into a number of logically distinct components.

1. *Precision* refers to how "refined" the categories of a discrete variable are. For example, are the values of the variable RELIGIOUS AFFILIATION to be (i) "Christian," etc. (least precise), (ii) "Protestant," etc. (a bit more precise), (iii) "Mainline Protestant," "Evangelical Protestant," etc. (still more precise), or (iii) "Episcopalian," "Lutheran," "Pentecostal," etc. (most precise). For another example, how precisely do we record the observed values of a continuous quantitative variable, e.g., do we record LEVEL OF TURNOUT in the 1996 election as 49%, 49.4%, 49.3946%, etc.?. About 25 years ago the precision of reported SAT scores was reduced from the nearest point to the nearest ten points (e.g., scores that were previously reported as 634 or 628 are now reported simply as 630). To *recode* a discrete variable (or to create *class intervals* for a continuous variable) is to reduce the precision of the measure for analytical purposes. If we were to recode LEVEL OF EDUCATION (3) or (4) into LEVEL OF EDUCATION (1) or (2) in some fashion, we would be reducing the precision of measurement.

   Precision relates only to the coding problem. Other components of accuracy pertain to the problem of operationalization.

2. A measure is *unreliable* to the extent that, when it is applied repeatedly *to the same case* (whose true value remains *constant*), it gives *different observed (measured) values*. All measures (especially of continuous variables with an infinite number of possible values) are unreliable to some degree. (This includes measures used in the "exact" physical sciences.) We have seen that when we attempt to *measure a population parameter by using a sample statistic*, the results are somewhat unreliable because of sampling error. Perhaps measures of "scholastic aptitude" and "general intelligence" such as SAT, IQ, etc., scores should not be rounded off but reported with a "margin of error" that reflects to their inherent unreliability. One major advantage of using an index to measure a variable is that an index is more reliable than its individual components (the SAT is as reliable as it is because it includes a wide variety questions). (Another advantage of an index is that it probably has greater precision — as defined above in #1 — than its individual components.)

   While a measure whose only accuracy problem is unreliability gives the right answer *on average*, this is not true of a measure that is *biased* or *invalid*.

3. A measure is *biased* to the extent that it tends *consistently* to produce measured values that are *too high* or *too low* relative to independently determined "true values." Bias can be either upwards or downwards, and the magnitude of the bias may a constant interval, a constant ratio, or may vary with the value of the variable being measured. In any event, bias in measurement, once recognized, is often (fairly) easy to correct for. We have already

discussed bias in relation to samples; we saw that non-random sampling may produce biased samples and therefore biased sample statistics (the accuracy of which cannot be improved by increasing sample size). One advantage of random samples is that they produce sample statistics (for percentages or averages) that are unbiased (though not fully reliable). Some social science measures are known to be biased. For example, survey measures of voting turnout are consistently higher than "true" turnout figures as determined by vote totals and census estimates of the VAP. There have been claims that the Consumer Price Index is biased upwards (i.e., that it consistently tends to overstate inflation) and longer standing claims that the Unemployment Index is biased downwards (i.e., that it consistently tends to understate unemployment), because it understates the size of the "labor force").

SAT is often said to be biased, but the complaints here actually pertain to its *validity*. (SAT, and many other measures, cannot be said to be biased in the sense defined here because there is no independent way of determining the "true value" of the variable in each case.)

4. A measure is *invalid* to the extent it is actually measuring variables *other than* the variable it is intended to measure. Some people say SAT really measures "test-taking ability" rather than true scholastic aptitude. Others say that SAT (in part) measures "middle-classness" or socially privileged status, rather than (or, more plausibly, in addition to) scholastic aptitude (though it should be noted that the SAT was introduced some 65 years ago as an attempt to do just the opposite). Another way of stating this claim is to say that SAT is "biased against" students from poorer, lower status, minority, or immigrant backgrounds. But notice that this is *not* a problem of "bias" as defined in #3 above but of validity — the claim is not that SAT is giving everyone scores that are too low (or too high) but that SAT is measuring students' social status instead of, or as well as, their scholastic aptitude. A clear-cut example of an *invalid test of scholastic aptitude* would be one administered in English to a group of children some of whom are "native Americans" and others of whom are (children of) recent immigrants to the U.S. who speak a language other than English at home. The test probably measures scholastic aptitude in part but clearly it also measures English language fluency (which is logically distinct from scholastic aptitude) in important part. If one is dealing with variables that pertain to aggregates such as states, nations, or other jurisdictions, measures that are based on *totals* like murders per year (see sentence #3 in PS #3) or traffic fatalities per year are obviously invalid, because they reflect not only how much mayhem there is, or how dangerous it is to drive, in those jurisdictions but also their total populations; the measures need to be put on a per capita (or per passenger-mile) basis to be valid.

In general, there is no simple or straightforward way for assessing the validity of a measure (and you are not responsible for all the terms and distinctions in Weisberg et al., pp. 94-96), but a couple of general considerations can be mentioned. Suppose we want to construct a measure of DEGREE OF RACIAL PREJUDICE among whites in the U.S. (It would be more difficult to construct a *single measure* of racial prejudice that would be valid *across racial groups* in the U.S., let alone across different cultures.) We surely would want to use an index rather than a single question. The measure has *internal* (or *face*) *validity* to the extent the items composing it all pertain overtly to race relations. (Items in many

psychological measures do not have such face validity.) Then we could test the *external validity* of the measure by applying it to "known groups." For example, we might administer the questionnaire to members of neo-Nazi group (in the unlikely event that university researchers could secure the cooperation of such respondents) on the one hand and to (white) members of (say) the American Civil Liberties Union on the other hand. If our measure did not pretty cleanly separate members of the two sets of groups (giving almost all respondents in the first group relatively HI scores and almost all respondents in the second group relatively LO scores), we could be pretty sure that it was not valid.