MEASURING VARIABLES: ANSWERS AND DISCUSSION

Some General Comments (This A&D is especially detailed, supplementing the discussion in Handout #4 and the PowerPoints, and should be read carefully.):

1.   In problems #5-7, a fairly common problem is that students address only the problem of *coding*, i.e., specifying what the values or categories of the variable will be, and not the more difficult problem of *operationalization*, i.e., specifying a *practical procedure* that will be used to collect data to determined that a case belongs in a particular category or has a particular value.

   For example, in "operationalizing" the variable DEGREE OF COMPETITIVENESS (of districts), some answers would say something like

            (A)   Very competitive
            (B)   Somewhat competitive
            (C)   Not competitive

   without giving any suggestion as to what kind of data would be collected and how it would be operationally determined whether a given district falls in category (A), (B), or (C).

2.   On the other hand, having proposed a perfectly reasonable and possibly quite *precise* measure of some variable — e.g., of DEGREE OF COMPETITIVENESS in terms (say) INCUMBENT'S MARGIN OF VICTORY IN LAST ELECTION (an essentially continuous and quantitative measure), many students (in effect) suggested "throwing away" that precision by *recoding* MARGIN into a few discrete categories such as those noted above. There is no reason to lose such precision at the data collection stage. (There may be reason to give up precision at the data analysis stage.)

3.   A related problem (anticipated in the second paragraph of Handout #4) is that many students seem to assume that all data must be collected, and all variables operationalized, through survey research — by asking questions of samples of people. Furthermore, the population to be surveyed was not always identified. A variable such as DEGREE OF COMPETITIVENESS (of districts) is probably much better (and certainly much more readily) operationalized on the basis of documentary research (e.g., election records) available in reference collections or elsewhere.

4.   A confusion worth mentioning is that the *empirical truth or falsity* of a proposition is logically unrelated to the *accuracy of the measures* of the variables used in the research aimed at assessing that truth or falsity. For example, some answers noted that some students may get good grades without studying hard while others may get poor grades despite much hard studying, contrary to the statement referred to in problem #7. It was also noted that outside jobs, etc., may interfere with studies. Both points may well be true, but they have nothing to do with the *accuracy of our measures* for DEGREE OF STUDY EFFORT and LEVEL OF GRADES. Indeed, we can't possibly find exceptions to the general proposition that "hard studying makes for good grades" (or disconfirm it entirely) without (reasonably accurate) measures of the two variables. More generally, we need to have (reasonably accurate) measures before we can gather and analyze *any* empirical data that bears on the empirical truth or falsity of the proposition.

Answers and discussion for the specific problems begin on the next page.

1.   Here is the recoding for cases 1-10 plus several other cases that present "problems." Remember that the value labels for Q2 depends on the respondent's answer to Q1.

[*Data from Spreadsheet*]			[*Decoded Data*]		[*Q1 & Q2 Combined into New Variable*]	
Value Codes			***Value Labels***		***Codes and Labels***	
ID	Q1	Q2	Q1	Q2	7-CATEGORY PARTY ID	
1	2	3	IND	NEITHER	4	PURE IND
2	1	2	DEM	NOT VERY STRONG	2	WEAK DEM
3	1	1	DEM	STRONG	1	STRONG DEM
4	1	1	DEM	STRONG	1	STRONG DEM
5	1	2	DEM	NOT VERY STRONG	2	WEAK DEM
6	2	2	IND	REP	5	REP LEANER
7	3	2	REP	REP	4	WEAK REP
8	2	1	IND	DEM	3	DEM LEANER
9	2	1	IND	DEM	3	DEM LEANER
10	5	9	DK	NA	9	NA
13	3	3	REP	???	6,7,9?	S/W REP, NA?
22	4	9	OTHR	NA	9	NA [Minor Party]
25	3	9	REP	NA	6,7	S/W REP?

[*Data*]			[*First Recoded Variable*]		[*Second Recoded Variable*]	
ID	Q1	Q2	DIRECTION OF PARTY ID		STRENGTH OF PARTY ID	
1	2	3	2	INDEPENDENT	1	NON-PARTISAN
2	1	2	1	DEMOCRATIC	3	WEAK PARTISAN
3	1	1	1	DEMOCRATIC	4	STRONG PARTISAN
4	1	1	1	DEMOCRATIC	4	STRONG PARTISAN
5	1	2	1	DEMOCRATIC	3	WEAK PARTISAN
6	2	2	3	REPUBLICAN	2	LEANER
7	3	2	3	REPUBLICAN	3	WEAK PARTISAN
8	2	1	1	DEMOCRATIC	2	LEANER
9	2	1	1	DEMOCRATIC	2	LEANER
10	5	9	9	NA	9	NA
13	3	3	3	REPUBLICAN	3,4	W/S PARTISAN?
22	4	9	9	NA	9	NA
25	3	9	3	REPUBLICAN	3,4	W/S PARTISAN?

Note.   Here I made the numerical codes for 7-CATEGORY PARTY ID run from 1 = STRONG DEM to 7 = STRONG REP (as in the SETUPS/NES Codebook); for DIRECTION OF PARTY ID run 1 = DEM, 2 = IND, and 3 = REP (as suggested on p. 3 of

Handout #4); and for STRENGTH OF PARTY ID run from 1 = NON-PARTISAN to 4 = STRONG PARTISAN (as also suggested on p. 3 of Handout #4); with 9 = MISSING DATA. The numerical order of the (non-missing) categories could just as well be reversed (but the coding should reflect the natural ordering from STRONG DEM to STRONG REP and from STRONG PARTISAN to NON-PARTISAN). Note that we can characterize the 4-CATEGORY STRENGTH OF PARTY ID variable the 7-CATEGORY PARTY ID variable "folded over" at its middle category.

2.   Since each individual variable is *dichotomous*, an obvious procedure is to add up (or average) the number of "cynical" (or "trusting") responses of each respondent to produce an overall INDEX OF POLITICAL CYNICISM (or TRUST) score. In fact, since all questions are coded with the same "polarity" ["cynical" always beings the lower code value and "trusting" the higher code value], we can just add up the three *code values* (apart from missing data) for each respondent. This will produce a scale that runs from a score of 3 (Most Cynical/Least Trusting) to a score of 6 (Least cynical/Most trusting).   We might want then subtract 2 points from each score to produce an adjusted scale that runs from 1 to 4. (The effect of this is merely cosmetic, since the scale has no *ratio* property.)

[*Data from Spreeadsheet*]   DEGREE OF CYNICISM or TRUST

ID	Q18	Q19	Q20	Index Score	Adj. Score	Value Labels	Value Labels
28	1	1	1	3	1	Most Cynical	Least Trusting
29	2	1	2	5	3	Less Cynical	More Trusting
30	2	2	2	6	4	Least Cynical	Most Trusting
31	1	2	2	5	3	Less Cynical	More Trusting
32	2	2	2	6	4	Least Cynical	Most Trusting
33	2	2	3	7	5	??? [but Quite Trusting?]	
34	2	2	1	5	3	Less Cynical	More Trusting
35	1	2	1	4	2	More Cynical	Less Trusting

Notice that some students (e.g., #31 and #34) may get the *same* (intermediate) index score (of 5 [or adjusted 3]) even though they gave *different* answers to two of the three questions. *Guttman scaling* is a technique by which this potential problem can be addressed (see Weisberg et. al., pp. 212-216).

In general, we will have to provide for cases that have *missing data* on one or more of the three items making up the index. Obviously we must take care that the missing data code (e.g., 9) does get added along with the valid codes 1 and 2. And clearly respondents who are missing on *every* individual item must likewise be coded as missing on the overall index, but it is less clear what to do with cases (e.g., #33) that have missing (or invalid) responses on one or two items. (Do you think case #33 should be thrown out as missing data or scored as a 5 or 6 [adjusted 3 or 4]?)

There is also the question of what to do if we form an index that includes questions with a third "neutral" option like Q23-25. (We might form an INDEX OF POLITICAL EFFICACY based on these questions.) We could proceed in this way: assign one "point" for each "more efficacious" response (always the last option); subtract one "point" for each "less efficacious" response (always the first option); and assign zero "points" for missing data or for a "neutral" response. What results is an index that runs from –3 (LEAST EFFICACIOUS) to +3 (MOST EFFICACIOUS). We could then add 4 points to each score to get rid of non-positive numbers and thereby to produce an INDEX OF POLITICAL CYNICISM with scores that run from 1 (LEAST EFFICACIOUS) to 7 (MOST EFFICACIOUS).

Note 1. Having produced something like this, some students suggested "throwing away" this degree of *precision* by grouping index scores into dichotomous LOW and HIGH categories. There is no reason to do this (at this stage, at least). See General Comment #2.

Note 2. Some students simply declared the respondents to be more or less cynical, without explaining how they did this. Remember: *measures must public, so others know what you have done*.

Note 3. SPSS can recode variables, compute new variables (including indexes), and perform other data transforming operations automatically. Click on Transform on the menu bar and select operations from the resulting menu, and see section VIII of *Using SETUPS 1972-2004 NES Data and SPSS for Windows* [Handout #1D].

3.   Most of these choices are straightforward and presented no problems. A few are less straightforward and required some thought. There was a tendency for some students to (evidently) skim quickly through the Codebook, find a variable whose label includes a word that seems relevant (e.g., "religion," "incumbent," "time"), and write that down without thinking the problem through.

   #2   V62 [for AMOUNT OF EDUCATION] *and* V68 [*indicator* for DEGREE OF RELIGIOSITY]

   Note. V67 (RELIGIOUS AFFILIATION) won't do, unless it is recoded into two categories 1-5 → "religious" and 6 → "non-religious", which produces a very crude (imprecise, perhaps invalid, and undiscriminating [in that about 90% of all Americans would fall in the "religious" category]) measure of RELIGIOSITY. Some students suggested V45 (ABORTION OPINION), but different religious denominations have different views on this.

   #11 V62 [for AMOUNT OF EDUCATION] *and* V65 (or possibly V64 or V66) [for LEVEL OF SUCCESS] [V63 really won't do; when your mother tells you to stay in school, she is hoping, not just that you get a job, but that you get a *good*, probably *high-paying*, job.] [*We will examine relevant SETUPS data later in the semester*.]

   #13 V10 (or perhaps V13 or V15 as *indicators* of interest) or an *index* formed from V10-V16) [for LEVEL OF POLITICAL INTEREST] *and* V03 [for WHETHER OR NOT VOTED]

#14   V34 [for DEGREE OF LIBERALISM] *and* V04 (or V06, V07, or possibly V09) [for VOTING HABITS]

#15   V32 or V33 (but note that these variables are substantively quite distinct) [for LEVEL OF SATISFACTION WITH ECONOMY] *and* V04 (or possibly V06 or V07) [for VOTE CHOICE — where the DEMOCRAT and REPUBLICAN categories would have to *recoded* (differently for each election year) into INCUMBENT and CHALLENGER categories]

Note. V8 (TYPE OF HOUSE RACE) will *not* do. This was a *very* common error, coming about I suppose because the variable description has the word "incumbent" in it. V8 is a "contextual" variables that provides objective information (not coming from the respondent) concerning the competitive situation that existed in the respondent's Congressional District, and it has nothing to do with the respondent's House vote or opinion concerning the House candidates. However, it is possible to construct a appropriate measure of VOTE CHOICE using V07 and V08 combined in this fashion:

V07	V08	VOTE (INCUMBENT OR CHALLENGER?)
1	1	INCUMBENT
1	2	NA
1	3	CHALLENGER
2	1	CHALLENGER
2	2	NA
2	3	INCUMBENT
Otherwise		NA

#17   V09 [*recoded* into STRENGTH OF PARTY IDENTIFICATION as in problem #1 above] and V01 [for (passage of) TIME]

Note 1. V60 [AGE] cannot be used for TIME, since AGE is cross-sectional data recorded at varying points in time (for example, potential respondents who were 30 in 1972 were 62 in 2004.) V05 (TIME OF VOTE DECISION) won't do either (think about it).

Note 2. Question #3 in Problem Set #5B asks you to assess the empirical truth of this claim using the SETUPS/NES 1972-2004 data.

4.   V68 ("How often do you go to religious services?") may be taken as a behavioral *indicator* of (or *proxy* for) DEGREE OF RELIGIOSITY, since — *other things being equal* — we would expect more religious people to attend religious services more frequently than less religious people. It is an indicator, because we are taking some *overt behavior* that is "objective" (and relatively easy to measure) as indicative of an *underlying attitude or orientation* that is "subjective" (and hard to measure). (In a comparable fashion, we might take VOTED IN ELECTION (Y/N) as an [imprecise] indicator for LEVEL OF INTEREST IN

ELECTION.) As a measure, V68 is probably more reliable than most other SETUPS variables, because the question is fairly straightforward and unambiguous. It is probably somewhat biased (like V03), in that reported church attendance is likely greater than actual church attendance, because regular church attendance (again like voting) is perhaps expected of "good members of the community" (*social desirability bias*). But the main problem with this indicator (and many others) is that of *validity*, as is suggested by the proviso "other things being equal" emphasized above. While frequency of church attendance no doubt is a function of DEGREE OF RELIGIOSITY, it is a function of many other things as well, e.g., proximity of the nearest place or worship (of your preferred religion or denomination), whether or not other people in your family are religious (think of a quite non-religious person who attends church regularly to accompany — and please — a religious spouse), enjoyment of the social interaction that church attendance provides, desire to appear devout, etc. Frequency of church attendance reflects all these other variables, as well as what we want to measure. (Note also that we could *not* measure DEGREE OF RELIGIOSITY by V68 to test the following perfectly sensible empirical proposition: "Highly religious people attend church more frequently than less religious people." Such an operationalization would turn this proposition into a *tautology* [i.e., something that appears to be an empirical statement but is actually true by operational definition], since we would be operationalizing *both* (the "independent" and "dependent" — as will be discussed in Handout #9) variables in the statement the same way.)

Note. Some students said that V68 was problematic because only Christians attend "church." But note that, while "Church" appears in the *name* of V68, respondents never see this and the actual *question* asked of respondents refers to "religious services" — phrasing that is appropriate for (most) non-Christians.

Given the opportunity to frame your own survey questions, you could devise questions that would more directly measure LEVEL OF RELIGIOSITY. By way of illustration, one study (Herbert McClosky and John Zaller, *The American Ethos: Public Attitudes Towards Capitalism and Democracy*) used an INDEX OF RELIGIOSITY based on the following questions:

> "How religious are you?"
> - (A)   Deeply
> - (B)   Fairly
> - (C)   Not very
> - (D)   Not at all
>
> "The best hope for the future of mankind lies in
> - (A)   science and human reason."
> - (B)   faith in God."
> - (C)   Decline to choose
>
> "What happens to man will depend mostly on
> - (A)   mankind's own efforts."
> - (B)   the will of God."
> - (C)   Decline to choose

"I believe that religion
- (A)   is largely old-fashioned and out of date."
- (B)   can still answer most of today's problems."
- (C)   Decline to choose

Recent ANES surveys have actually included questions along similar lines, but they are not included in the (much abbreviated) SETUPS Codebook and data. The relevant portion of an original questionnaire is reproduced below. (These questions appear in the questionnaire after the basic question for RELIGION [i.e., V67] and would not all be asked of non-Christians.)

Do you consider religion to be an important part of your life, or not?

Would you say your religion provides some guidance in your day-to-day living, quite a bit of guidance, or a great deal of guidance in your day-to-day living?

Some people have had deep religious experiences which have transformed their lives. I'm thinking of experiences sometimes described as "being born again in one's faith" or "discovering Jesus Christ in one's life." There are deeply religious people who have not had an experience of this sort. How about you; have you had such an experience?

Here are four statements about the bible, and I'd like you to tell me which is closest to your own view.

1. The bible is god's word and all it says is true.
2. The bible was written by men inspired by god but it contains some human errors.
3. The bible is a good book because it was written by wise men, but God had nothing to do with it.
4. The bible was written by men who lived so long ago that it is worth very little today.

However, in my judgment it will always be very difficult (and perhaps impossible) to design a *single measure* of RELIGIOSITY that has much validity when applied across *very different religious traditions*.

5.   DEGREE OF COMPETITIVENESS. A very simple (and easy to obtain) measure would be REPRESENTATIVE'S WINNING PERCENTAGE MARGIN IN LAST ELECTION (bigger margin => less competitive). This is not entirely *valid*, however, because this statistic depends on other variables, e.g., the quality and financing of the opposition candidate in the last election. (Think of a member of Congress in what is really a very competitive district who had the good fortune to run against an opponent who was a total idiot or was caught in a scandal just before the election, so the member won by an unusually large margin.) A better measure therefore might be REPRESENTATIVE'S AVERAGE WINNING MARGIN (over the past several elections or for as long as he or she has held office). Alternative measures might be based on the Presidential vote in the district in the last election (or averaged over several recent elections or on the balance of Democratic to Republican voter registration in the district (usable only in those states with voter registration by party). The larger the support for a Presidential candidate of, or the proportion of voters registered with, the party opposite from the member's party, the more competitive the district. Still another approach would be to measure competition in terms of the balance of campaign expenditures in the last

(several) elections(s). This data must be reported to the Federal Election Commission and is available to the public. (Also review the General Comments on the first page that pertain particularly to measuring this variable.)

DEGREE OF RESPONSIVENESS. First, we need to be clear about what *concept* of "responsiveness" we have in mind. Do we mean that the extent to which a member's roll-call votes are generally congruent with preponderant opinion in his or her district? This would suggest one approach to operationalization that would probably be based on roll-call data and constituency data of some sort. Alternatively (and perhaps more likely), do we mean the extent to which a member is "responsive" to the requests and concerns of constituents (in the manner of an "ombudsman"). A common measure (an indicator) used in political science research is (something like) NUMBER OF DAYS PER YEAR SPENT IN DISTRICT (available from congressional records). This is not entirely valid, because to some extent it is a measure of the proximity of the district to Washington, D.C. (and whether the member has moved his family home to the D.C. area), as well as other variables. In any case, this measures only one aspect of DEGREE OF RESPONSIVENESS. It would be better to form an *index* of responsiveness based on several items such as NUMBER OF STAFF WORKERS DEDICATED TO CONSTITUENCY (as opposed to legislative) BUSINESS, NUMBER OF MAILING TO CONSTITUENTS, NUMBER OF DISTRICT OFFICES, etc.

Some students suggested using the approval ratings of members of Congress (supposing they are available) as indicators of their responsiveness (on the plausible supposition that responsiveness causes their approval rating to go up). The problem is that members with low approval rating have an especially large incentive to be responsive. Moreover, members elected from more competitive districts are likely to have relatively low approval ratings, since many of their constituents will have voted against them in the previous election.

6.   NUMBER OF DOCTOR VISITS is a valid (and presumably negative, i.e., more doctor visits indicating poorer health) indicator/proxy for LEVEL OF HEALTH if NUMBER OF DOCTOR VISITS is influenced *only* (or at least mostly) by LEVEL OF HEALTH and not (much) by any other variables. But NUMBER OF DOCTOR VISITS will likely be influenced by INCOME (or HEALTH INSURANCE STATUS) of prospective patients (uninsured poor people won't visit the doctors so often, even if they are quite sick). People differ in how much trust they put in doctors or how inclined they are, if sick, to "tough it out" without medical attention. Others have hypochondriac tendencies and visit a lot of doctors even when there is nothing wrong with them. Given such considerations, it seems likely that this indicator has serious validity problems. Finally, doctors (and others) think that if you go to them for treatment, they can make you somewhat healthier so, while more doctor visits might be indicative of poor health, they can also produce better health.

Note.   Whether or not apple eating promotes health is irrelevant to the question of whether FREQUENCY OF DOCTOR VISITS is a valid indicator of LEVEL OF HEALTH.

7.   "HARDNESS" OF STUDYING. In practice, we probably would have to transform this into AMOUNT OF TIME STUDYING. Then the operational question becomes how to determine this in each case. The most feasible approach would be to conduct a student survey and include a question asking respondents how many hours they study a week for each of their courses (and average over courses). The responses probably would not be very reliable and would likely be biased upwards. Another approach would be to ask a sample of students to keep a log over a week or two, recording how they are spending their time (in the manner of some TV rating surveys). Again there is the problem of whether these would be kept accurately. A still less feasible approach would be to have observers follow students in a sample around for a week or so, with the observers keeping the log. A problem that arises here is the problem of *obtrusive measurement* (or *observation*): knowing that they are being observed and that their activities are being logged, the student subjects would likely behave differently from the way they otherwise would. (A similar problem might arise if students were asked to keep their own logs.)

LEVEL OF GRADES. This is a good deal more straightforward. Ask students about their current or cumulative GPAs or, better, verify this information by checking their transcripts or similar documentary records. (Note that a GPA is an *index* composed out of course grades that is both more *precise* and more *reliable* (but no more *valid*) than its individual components.)

8.   Taking NUMBER OF RETURNS as a measure (indicator) of LEVEL OF CUSTOMER SATISFACTION is invalid if different stores have different sales volumes. RATE OF RETURNS would be a more valid measure of SATISFACTION.

        for Sears: 36/1100 =   3.3%
        for La Boutique: 12/200   =   6%

But if the stores have different return policies, even this rate is not a valid measure of LEVEL OF CUSTOMER SATISFACTION.

9.   To calculate the rate of executions per million in Texas:
        144 / 1.986 million ≈ 8.5 executions per million population

State	Executions (1977-1998) per Million (in 1990)	As decimal fraction or %	
FL	3.0	.0000030	= .00030%
AL	3.9	.0000039	= .00039%
NV	5.0	.0000050	= .00050%
MO	5.7	.0000057	= .00057%
AR	6.8	.0000068	= .00068%
VA	7.4	.0000074	= .00074%
TX	8.5	.0000085	= .00085%

So, with respects to rates, Florida and Texas are actually quite dissimilar. (By comparison, the nationwide *rate of murders per year per 100,000 of population* is about 5.5 [15-20 years year ago it was about 10], ranging from about 1 to 12 across states and much more widely among smaller jurisdictions.)

10.   These questions test people's *knowledge of particular subjects*, not their *general intelligence*. The first might appear as an item in an index measuring "cultural literacy," the second in an index measuring of sports knowledge.

11.   The fact that the BBI is not involved in most criminal cases is not the issue; the FBI simply assembles reports from local police department. But some police departments may not report their figures at all, and others may not keep complete and accurate records. The legal definitions of crimes, or police norms as to how crimes are "coded," vary from jurisdiction to jurisdiction. Sometimes police have an incentive to fudge their figures (downward to show what a good job they are doing, or upward to justify budget increases). Probably most important, victims of minor crimes (such as petty theft, vandalism, etc.), and of some major crimes (especially sexual assault) may not bother to report them to the police (knowing there is little the police can do or because of embarrassment or fear). Such crimes may be under-reported in the National Crime Victimization Survey as well but probably to a lesser extent. A more general problem may be that any "overall crime rate" counts more and less serious crimes equally.

12.   Among many other possibilities, indicators might include: student/faculty ratio [but how are part-time students and part-time faculty to be factored in?], proportion of introductory courses taught by regular faculty, average class size*, student retention and (four-year) graduation rates [but these depends a lot on the mix of students a college enrolls in the first place, rather than what the college does for these students], percent of students who are out-of-state, faculty research productivity [but this may come at the expense of undergraduate instruction]. Bear in mind that the measures must make valid comparisons across institutions of greatly different size, so indicators must be stated in terms of rates or percents, not absolute numbers, dollars, or whatever.

    * There is a statistical phenomenon commonly called the *class size paradox*. Consider the following highly simplified example. 1000 students are enrolled in one course each: three lecture classes with 300 students each and 10 seminar classes with 10 students each. Average class size as reckoned by administrators is about 77 (a total enrollment of 1000 divided among 13 classes); but average class size *as observed by students* is about 270 (900 students observe a class size of 300, while 100 observe a class size of 10). It follows that college students and administrators have fundamentally different perceptions of class size, which are both accurate but in different ways. The same "paradox" occurs in other contexts. The typical airline "load factor" (percent of seats occupied, averaged over all flights) is about 70%, but most airline passengers would swear it is much higher, because they mostly experience flights that are (virtually) full (that's why they're full). In supplementary non-political surveys given in some POLI 300 classes, the average number of siblings of reported by students was 1.85, so on average students grew up in households with 2.85 children. But even if the POLI 300 student sample were representative of the entire population aged 18-25 or so (which it isn't), this wouldn't mean that the average U.S. family has almost 3 children. [Over the past generation or so, U.S. families have had just about 2 children on average.] Clearly childless families were entirely unrepresented in the POLI 300 sample, while the children of large families were over-represented in proportion to their size. These (not really so paradoxical) discrepancies arise because (in effect) the *unit of analysis* shifts from *classes*, *flights*, and *families* to *students*, *passengers*, and *siblings* respectively.