Missing Data and Scale Building: Some Examples

April 12, 2011

Ashley Jardina
Kenneth M. Coleman
University of Michigan

An Outline

- Why bother with multiple measures of Economic Status when we have ES05 and ES06, two seemingly good measures of household income, in the Omnibus Survey?
  - Review of discussion of May 2010. Reasons for missing data:
    - Unwillingness to share data
    - Gender relations in the household
    - Lack of knowledge of household income
  - Solution adopted in Mexico, “the light bulb scale” of family income, number of light bulbs in the home as a proxy of household economic status. Works well in a poorer country. Would not work in Qatar.

Outline Continued

- Thought problem considered in May 2010, led to intriguing suggestions from attendees for a survey in Qatar, many of which *are included* in the 2010 Omnibus Survey:
  - ES01: Household Employees [Qataris only]
    - ES011: Number of Maids
    - ES012: Number of Nannies
    - ES013: Number of Drivers
    - ES014: Number of Gardeners
    - ES015: Number of Cooks
    - ES016: Number of Other Household Employees

Outline Continued

- ES02: Luxury Living Quarters [Qataris only]
  - ES021: Palace
  - ES022: Vacation Home
  - ES023: Yacht
  - ES014: Chalet
  - ES015: Farmhouse
- ES02a: Size of TV
  - Owns TV larger than 46”
  - Does not own TV larger than 46”
- ES03: Swimming pool [shared pools not counted]
  - Residence has private swimming pool
  - Residence does not have private swimming pool

Outline Continued

- ES04: Number of bedrooms of dwelling [in which interview conducted]
- ES04a: Number of vehicles owned
  - Car/Saloons
  - SUVs
  - Pickup/Trucks

- These items in the 2010 Omnibus survey may give us:
  - Fewer missing data responses
  - An opportunity to tap other dimensions of economic status.

Possible Components of a Scaled Measure of Economic Status: Focusing on Qataris Only

Considerations for Counts, Indexes and Scale Construction of Possible Use in Assessing Economic Status

Var nam	Content	Stratum	N	Valid Val	Missing Values	Missing Value Ns	Impression of Skewness
ES05	HH Income	Qatari	689	0 - Qr 150,000+	8,9, System	Ns = 54, 22,1450	Very Str: < QR 50,000 = 511 of 613 valid resp.
ES05	HH Income	Ex-Pats	768	0 - Qr 150,000+	8,9, System	Ns = 11, 11,1371	Extr. Str: < QR 50,000 = 721 of 746 valid resp.
ES011	Maids employed	Qatari	689	0-9	98, 99, System	Ns = 2, 1, 1450	Moderate: 0=46, 1=283, 2=249, 3+=111
ES012	Nannies employed	Qatari	689	0-10	98, 99, System	Ns = 15, 5, 1450	Very Str: 0=575, 1=66, 2=17, 3+=10
ES013	Drivers employed	Qatari	689	0-9	98, 99, System	Ns= 6, 1, 1450	Strong: 0 =241, 1=324, 2=94, 3+=21
ES014	Gardeners empl	Qatari	689	0-9	98, 99, System	Ns=18, 5, 1450	Very Str: 0=578, 1=82, 2=4, 3+=2
ES015	Cooks employed	Qatari	689	0-9	98, 99, System	Ns=17, 5, 1450	Very Str: 0=605, 1=53, 2=4, 3+=6
ES016	Others employed	Qatari	689	0-11	98, 99, System	Ns=17, 5, 1450	Very Str: 0=646, 1=14, 2=3, 3+=3
ES021	Own palace	Qatari	689	1, 2	8,9, System	Ns=3,1, 1450	Extremely Str: Yes (1)=19, No (2)=666
ES022	Own vacation home	Qatari	689	1, 2	8,9, System	Ns=3,2, 1450	Very Str: Yes (1)=79, No (2)=605
ES023	Own yacht	Qatari	689	1, 2	8,9, System	Ns=3,1, 1450	Extremely Str: Yes (1)=23, No (2)=662
ES024	Own chalet	Qatari	689	1, 2	8,9, System	Ns=3,1, 1450	Extremely Str: Yes (1)=14, No (2)=671
ES025	Own farm house	Qatari	689	1, 2	8,9, System	Ns=3,1, 1450	Extremely Str: Yes (1)=56, No (2)=629
ES02a	TV > 46 inches	Qatari	689	1,2	8,9, System	Ns=16, 0,1450	Strong: Yes (1)=256, No (2)=417
ES03	Swimming pool?	Qatari	689	1,2	8,9, System	Ns=0,1,1450	Extremely Str: Yes (1) = 33, No (2)=654

Fundamental Issues in Scale Construction, I

- What do the distributions of each potential item in the scale look like? Do certain items give a wider distribution on answers?
  - Remember that the purpose of analysis is to explain variation or co-variation in ***variables***, i.e., measured concepts that actually vary.
- How much do the variables co-vary; how strong is the intercorrelation?
  - Use of cross-tabs to explore at the first level of analysis.

Fundamental Issues in Scale Construction, II

- Is there a preferred simple item, such as ES05 [Qataris and White Collar Ex-Pats] or ES06 [Blue Collar Guest Workers], but one which has excessive missing data?
  - Could another, highly correlated item, simply be substituted?
  - Which respondents are "missing"? Can we characterize those who are missing on the preferred variable?
    - High education? Specific age grouping? Females?
- Which items seem to have "face validity" as plausible measures of the same underlying concept?
- The benefits of multiple indicators.
  - Psychometric theory
    - True variation plus an error component in each measure.
  - Multiple indicators need to have some degree of correlation [co-variation], but not too much. Otherwise, additional measures cannot compensate for any defects of existing measures.

Missing Data on Income: Comparing Strata

- On ES05, among Qataris missing data reaches 11.0% [DK=7.8%; REF=3.2%].
- Among White Collar Ex-Patriots the percentage of DK and Ref on ES05 is only 3.2% [DK=1.6%; REF=1.6%].
- And only 1 of 682 blue collar guest workers [0.1%] did not know or refused to reveal income on ES06.
- The order of the severity of missing data on income is:
  - Qataris = greatest challenge, with over one in ten interviewees generating missing data.
  - White Collar Ex Pats = approximately one in thirty cases have missing data.
  - Blue Collar Guest Workers: Fewer than one in thirty cases exhibit missing data on any kind on the income question.

One Possible Solution: Substitute ES04 [Number of BR in HH] for ES05 [HH Income]

number of bedrooms in hh

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1	.1	.2	.2
	1.00	6	.3	.9	1.1
	2.00	31	1.5	4.5	5.6
	3.00	83	3.9	12.1	17.7
	4.00	156	7.3	22.7	40.4
	5.00	165	7.7	24.0	64.3
	6.00	119	5.6	17.3	81.6
	7.00	50	2.3	7.2	88.8
	8.00	36	1.7	5.2	94.0
	9.00	18	.8	2.6	96.6
	10.00	11	.5	1.6	98.2
	11.00	2	.1	.3	98.5
	12.00	6	.3	.9	99.4
	14.00	2	.1	.3	99.7
	15.00	1	.0	.1	99.8
	20.00	1	.1	.2	100.0
	Total	689	32.2	100.0	
Missing	System	1450	67.8		
Total		2139	100.0		

Would it make sense simply to substitute a variable with greater variation , lower skewness, and no missing data [ES04] for HH Income [ES05]?

<u>ES04 [# of Bedrooms]</u>

Mean: 5.16
Standard Deviation: 2.10

Skewness: 1.59 on ES04 versus 2.39 on ES05.

An Easy First Step

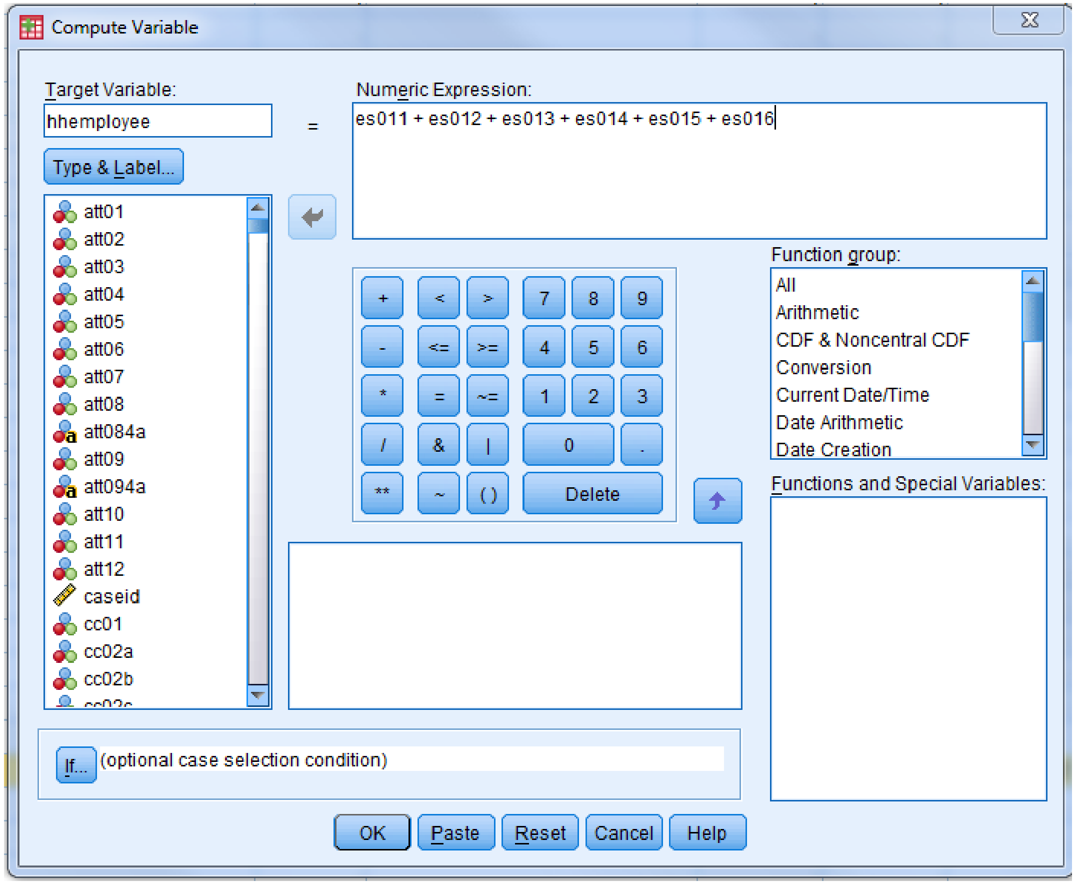
COUNT VARIABLES

Count Variables: Two Examples

- One way to combine variables is simply to count cases of similar phenomena. In the 2010 Omnibus data set one might do that with two variables ES01 [household employees] and ES04a [number of vehicles].
  - In doing a COUNT, the analyst does make assumptions, such as assuming that a cook is comparable to a gardener, or that an SUV is comparable to a pickup. Not exactly true, but each represents an "investment" closer in value to each other than other possible investments, such as employing an orchestra or owning a jet airplane.
  - Counting number of residences might be more troublesome if a palace ≠ farm house ≠ vacation home.
- The following slides illustrate how to do a COUNT in SPSS, using ES01 and ES04a.

Count Variables Continued

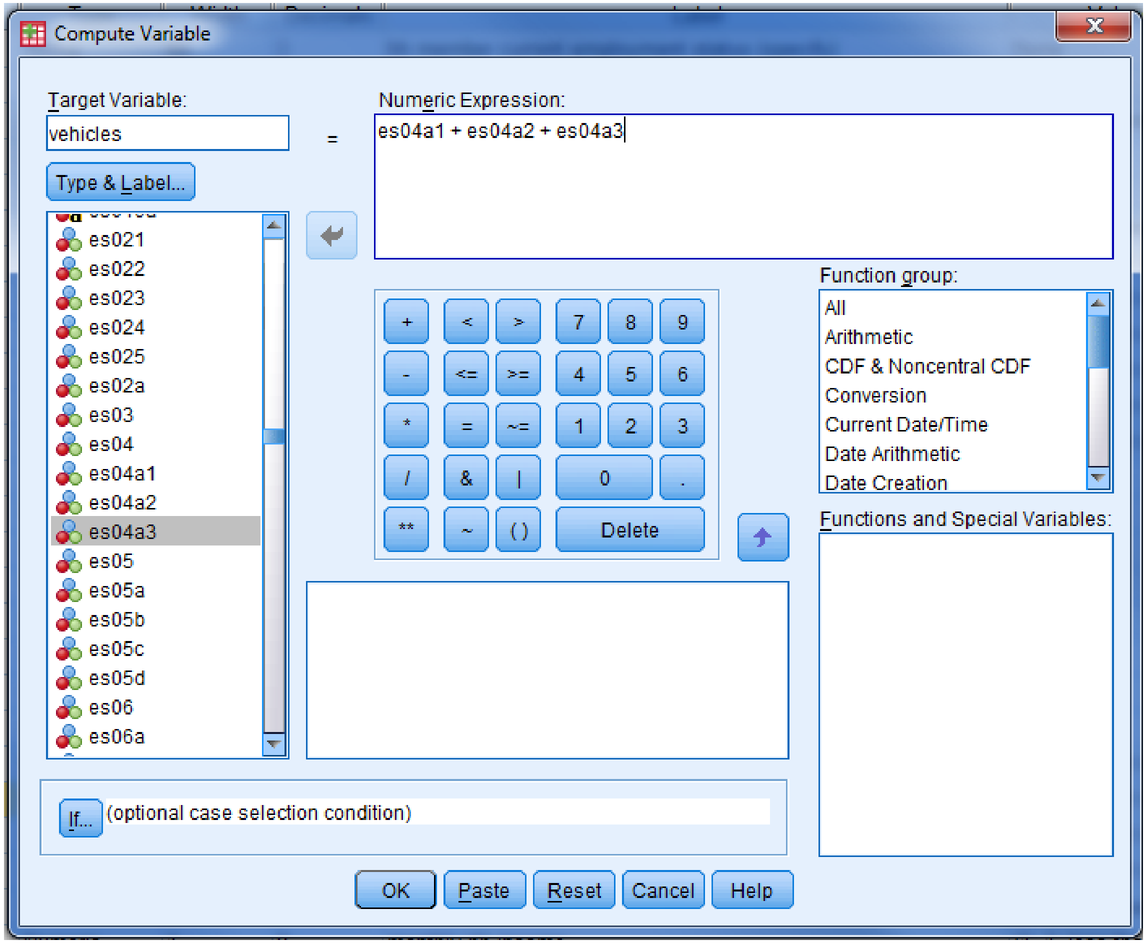
- Compute a new variable equal to the sum of the relevant variables (i.e., number of maids + number of nannies + number of drivers + etc.)



The new variable, "hhemployee" is equal to the total number of household employees for each survey respondent.

Count Variables Continued

- We can follow the same procedure to create a variable equal to the total number of vehicles (cars, suvs, and trucks) in each respondent's household



Count Variables Continued

Total number of household employees

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	33	1.5	4.9	4.9
	1.00	123	5.7	18.5	23.4
	2.00	147	6.9	22.1	45.5
	3.00	166	7.7	24.9	70.4
	4.00	95	4.4	14.3	84.7
	5.00	45	2.1	6.8	91.5
	6.00	20	.9	3.0	94.5
	7.00	8	.4	1.2	95.7
	8.00	6	.3	.9	96.6
	9.00	3	.2	.5	97.1
	10.00	4	.2	.5	97.7
	11.00	3	.1	.4	98.1
	13.00	3	.1	.4	98.5
	14.00	5	.2	.7	99.2
	15.00	2	.1	.3	99.6
	22.00	1	.1	.2	99.8
	34.00	1	.1	.2	100.0
	Total	665	31.1	100.0	
Missing	System	1474	68.9		
Total		2139	100.0		

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Total number of household employees	665	.00	34.00	3.0978	2.81818
Valid N (listwise)	665				

Count Variables Continued

Total number of household vehicles

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	15	.7	2.2	2.2
	1.00	64	3.0	9.6	11.8
	2.00	140	6.5	20.8	32.6
	3.00	169	7.9	25.2	57.7
	4.00	114	5.3	16.9	74.6
	5.00	85	4.0	12.6	87.2
	6.00	32	1.5	4.8	92.0
	7.00	19	.9	2.8	94.8
	8.00	16	.8	2.4	97.2
	9.00	9	.4	1.4	98.6
	10.00	1	.1	.2	98.8
	12.00	1	.1	.2	99.0
	13.00	3	.1	.4	99.5
	23.00	2	.1	.3	99.8
	37.00	1	.1	.2	100.0
	Total	673	31.4	100.0	
Missing	System	1466	68.6		
Total		2139	100.0		

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Total number of household vehicles	673	.00	37.00	3.6318	2.75805
Valid N (listwise)	673				

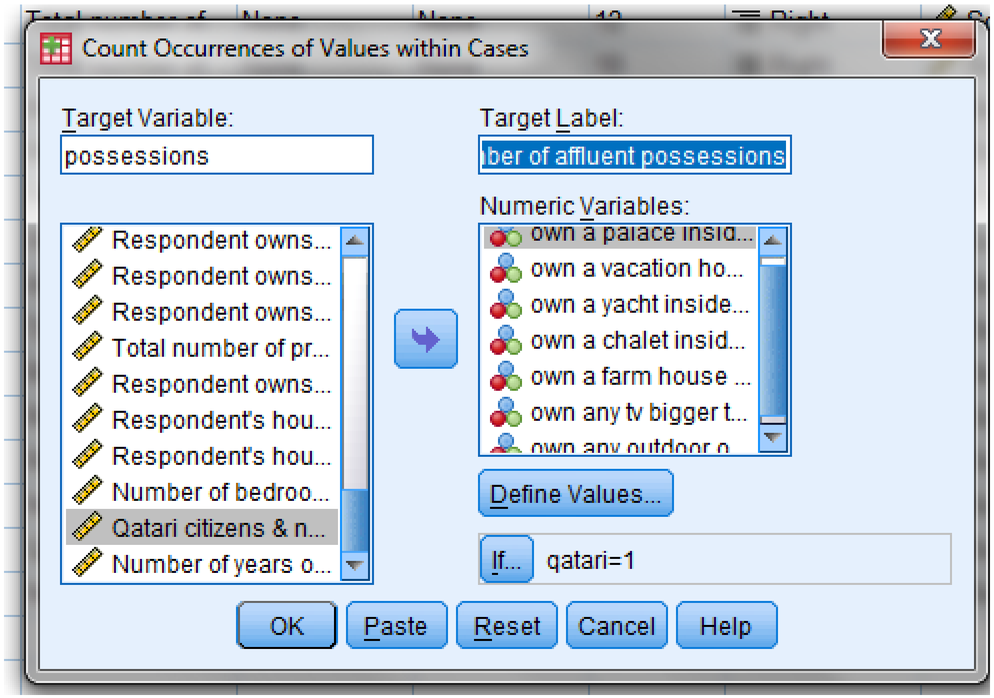
Count Variables: Another Method

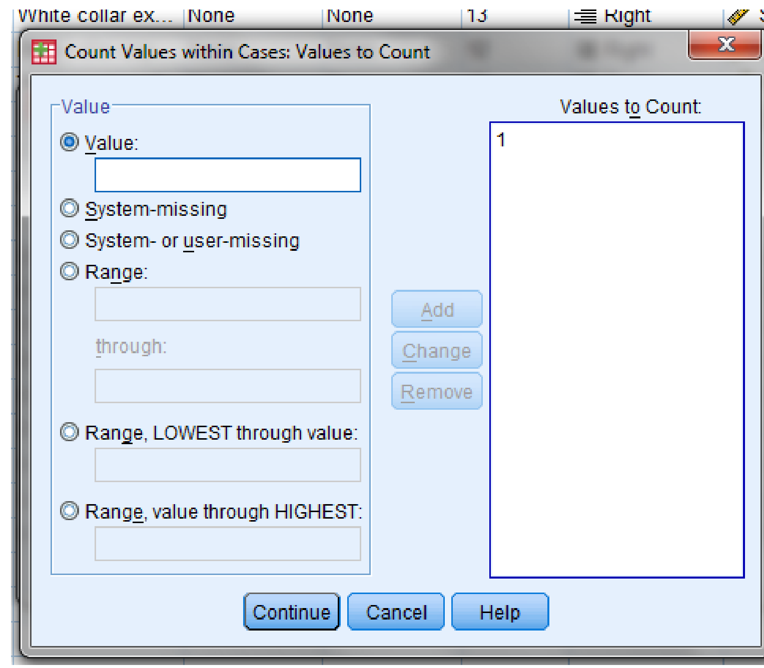
We can also use the COUNT function to count the *number of times* a value occurs, rather than *adding together* the values of variables.

This recode function can be used to construct simple summary indices of how many (or how often) certain responses are provided.

For example, we can take the questions in which respondents indicated only “yes” or “no” rather than “how many” (.e.g, “do you own a palace” vs. “how many cars do you own”) and create an index of the number of affluent possessions for each Qatari respondent.

Count Variables Continued





In the dataset, a value of 1 indicates that a respondent said “yes” to whether they own a palace, vacation home, chalet, farmhouse, big tv, or outdoor pool. Therefore, we want to tell SPSS to count the number of 1’s.

Count Variables Continued

Statistics

number of affluent possessions

N	Valid	689
	Missing	0

number of affluent possessions

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	368	53.4	53.4	53.4
	1.00	231	33.5	33.5	86.9
	2.00	50	7.3	7.3	94.2
	3.00	24	3.6	3.6	97.8
	4.00	8	1.1	1.1	98.9
	5.00	3	.4	.4	99.4
	6.00	3	.4	.4	99.8
	7.00	1	.2	.2	100.0
	Total	689	100.0	100.0	

We can look at the frequencies of the new index we created to see what the distribution of affluent possessions is among Qataris. We see, for example, that 33.5% (N=231) of Qataris in the sample have 1 of these possessions.

A Second Step

EXPLORING COVARIATION AMONG POSSIBLE INDICATORS

Extent of Co-Variation in ES indicators?

- Case of Number of BR in HH and Number of HH Employees.

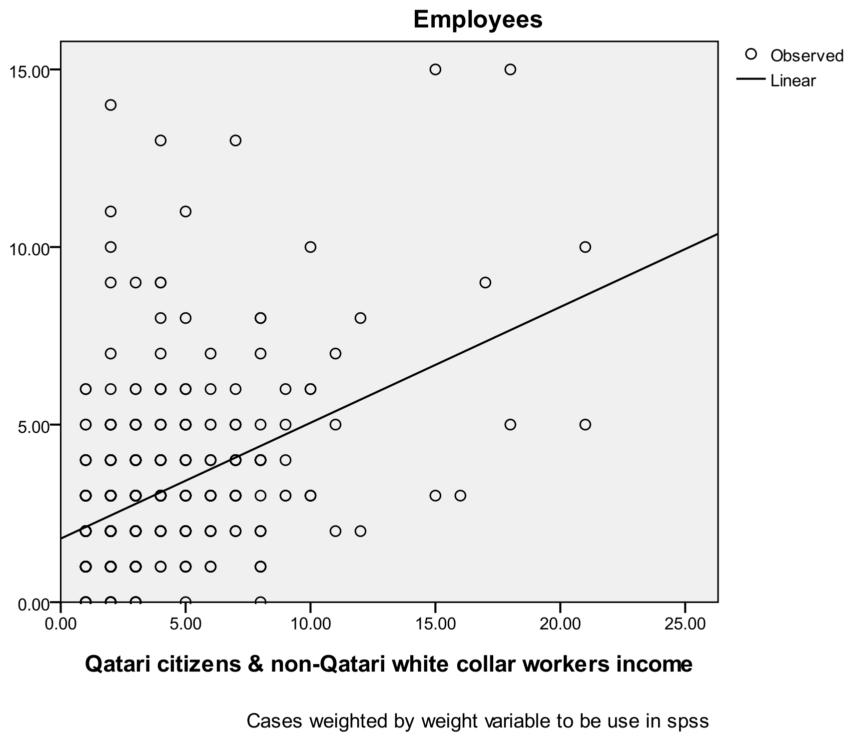
While far from a perfect relationship, in this output from SPSS we can see that there is a tendency for HH with more bedrooms to be HH with more employees. The Pearson correlation coefficient Is +.365 [on a scale of -1.0 to +1.0]. If the correlation were +1.0 or -1.0, all the data points would be on the regression line.

Total Bedrooms in Qatari households

Note: these graphs can be made in SPSS using the "curve estimation" option under regression analysis.

Extent of Co-Variation in ES indicators?

- Case of Unfolded Income Scale [ES05_INC, to be defined later] and Number of HH Employees.

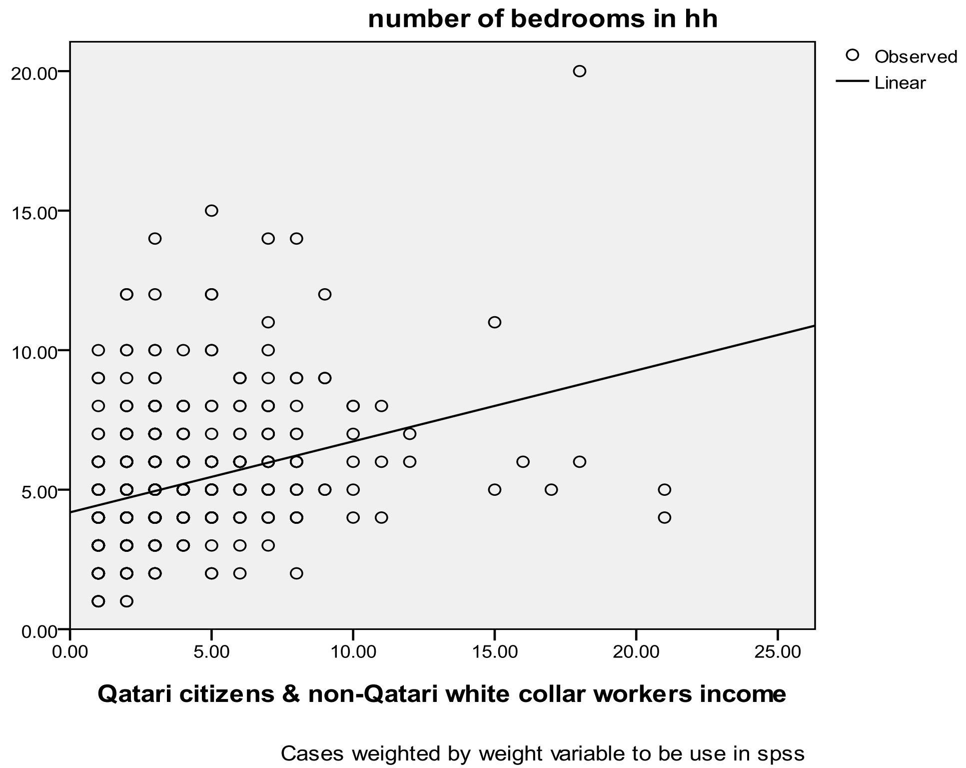


While far from a perfect relationship, in this output from SPSS we can see that there is a tendency for Qatari HH with higher incomes to employ more HH staff. The Pearson correlation coefficient Is +.409 [on a scale of -1.0 to +1.0]. If the correlation were ± 1.0, all the data points would be on the regression line.

Qatari HH Income in Increments of QR 10,000

Extent of Co-Variation in ES indicators?

- Case of Unfolded Income Scale [ES05_INC, to be defined later] and Number of Bedrooms.



Qatari HH Income in Increments of QR 10,000

Again, while far from a perfect relationship, in this output from SPSS we can see that there is a tendency for Qatari HH with higher incomes to have houses with more bedrooms. The Pearson correlation coefficient Is +.339 [on a scale of -1.0 to +1.0]. If the correlation were ± 1.0, all the data points would be on the regression line.

Would Substitution of Mean Income for Missing Data on ES05 Make Sense?

Descriptive Statistics

	N	Mean	Std. Deviation
number of bedrooms in hh	689	5.1557	2.09448
Valid N (listwise)	689		

Descriptive Statistics

total monthly income of all hh members		N	Mean
1. less than qr50,000	number of bedrooms in hh	511	4.8407
	Valid N (listwise)	511	
2. qr50,000 to less than qr100,000	number of bedrooms in hh	87	6.3582
	Valid N (listwise)	87	
3. qr100,000 to less than qr150,000	number of bedrooms in hh	8	6.2948
	Valid N (listwise)	8	
4. qr150,000 or more	number of bedrooms in hh	7	8.0707
	Valid N (listwise)	7	
8. don't know	number of bedrooms in hh	54	5.5321
	Valid N (listwise)	54	
9. refused	number of bedrooms in hh	22	5.4019
	Valid N (listwise)	22	

Note similarity of overall mean number of bedrooms among all interviewees to the mean number of bedrooms in HH where interviewees either did not know HH income or refused to reveal it. All are in the range of 5.16 to 5.53 bedrooms.

Some Observations Based on Relationship Between ES04 and ES05 Pertinent to Inferences Regarding Missing Data

- * Overall, the number of bedrooms in the HH [ES04] is strongly associated with HH income [ES05].
  - HH with incomes under QR 50,000 have, on average, 4.84 bedrooms, while those with incomes of QR 200,000 or more have, on average, 8.07 bedrooms.
  - If one assumes that ES04 could serve as a proxy for ES05, observations relevant to the 11.0% cases of missing data on ES05 are possible.
    - The mean number of bedrooms in the whole sample is 5.15, while the mean number of bedrooms among DK respondents if 5.53 and among Ref respondents is 5.41, both closer to 5.11, the overall mean, than to the number of bedrooms in any other income category.
    - Is this indirect evidence that substitution of a mean value on ES05 would make sense? But what about the fact that ES05 is highly skewed and has only four categories?

Thought Exercise: The Art of Addressing Missing Data:

- There are some relatively "easy choices" that we could make pertaining to missing data on ES05. What are the consequences of using each?
  - Should we accept 11.0% of cases as missing among Qataris? What are the consequences of doing that?
    - Hint: What if another variable that we want to run income against has another 10% missing values, and the missing values on Variable XYZ do not overlap with those on ES05?
    - Hint: What percentage of missing data on income might one find in Western Europe or in the United States?
  - Should we *accept some error, but seemingly a modest amount,* by substituting the mean income value on ES05, thereby losing fewer cases?
  - Or should we simply substitute ES04 for ES05 in subsequent analyses, since ES04 [number of BR in HH] has no missing data at all and is another measure of ES.

Another Approach:
Unfolding ES05a to ES05b

BUILDING A NEW HH INCOME ITEM

Unfolding Household Income: Qataris and White Collar Ex-Patriots

- ES05 in the data set places respondents in wide categories, while items ES05a – ES05d "unfold" those categories.
- Constructing a more detailed scale is possible using ES05a – ES05d.
- In this case both ES05 and the more detailed scale [ES05a-ES05d] are "bottom-heavy" scales, with many cases falling toward the lower end of the income spectrum.
- Given that this is an initial national survey, it was hard to foresee the distribution of reported income.
  - In the future, one might wish to have more categories at the lower end of the scale.

Unfolding ES05

- Note that ES05 has answers in terms of categories that encompass ranges of QR 50,000.
- However, ES05a – ES05d break those down into further QR 10,000 increments, until reaching QR 200,000 + QR.
- ES05a – ES05d can be combined into a new and more detailed scale. See Appendix for code.
- The benefits for doing so are to reach a finer degree of measurement of income categories. In this case, it leads to *a somewhat less skewed* distribution of values on HH income, but a distribution that remains skewed.

ES05_INC: A Variable Created to "Unfold" Larger Income Groupings

Unfolded Scale ES05a-ES05d Both		*Unfolded Scale* ES05a-ES05d Qataris		*Unfolded Scale* ES05a-ES05d Ex-Pats	
QR < 10,000	369	QR < 10,000	86	QR < 10,000	283
QR 10,000 - 19,999	451	QR 10,000 - 19,999	170	QR 10,000 - 19,999	281
QR 20,000 - 29,999	211	QR 20,000 - 29,999	124	QR 20,000 - 29,999	87
QR 30,000 - 39,999	107	QR 30,000 - 39,999	64	QR 30,000 - 39,999	43
QR 40,000 - 49,999	68	QR 40,000 - 49,999	56	QR 40,000 - 49,999	12
QR 50,000 - 59,999	41	QR 50,000 - 59,999	30	QR 50,000 - 59,999	11
QR 60,000 - 69,999	28	QR 60,000 - 69,999	26	QR 60,000 - 69,999	2
QR 70,000 - 79, 999	18	QR 70,000 - 79, 999	15	QR 70,000 - 79, 999	3
QR 80,000 - 89,999	6	QR 80,000 - 89,999	5	QR 80,000 - 89,999	1
QR 90,000 - 99,999	7	QR 90,000 - 99,999	7	QR 90,000 - 99,999	**0**
QR 100,000-109,999	4	QR 100,000-109,999	4	QR 100,000-109,999	**0**
QR 110,000-119,999	2	QR 110,000-119,999	2	QR 110,000-119,999	**0**
QR 120,000-129,999	**0**	QR 120,000-129,999	**0**	QR 120,000-129,999	**0**
QR 130,000-139,999	1	QR 130,000-139,999	**0**	QR 130,000-139,999	1
QR 140,000-149,999	2	QR 140,000-149,999	2	QR 140,000-149,999	**0**
QR 150,000-159,999	1	QR 150,000-159,999	1	QR 150,000-159,999	**0**
QR 160,000-169,999	1	QR 160,000-169,999	1	QR 160,000-169,999	**0**
QR 170,000-179,999	1	QR 170,000-179,999	**0**	QR 170,000-179,999	1
QR 180,000-189,999	1	QR 180,000-189,999	0	QR 180,000-189,999	1
QR 190,000-199,999	**0**	QR 190,000-199,999	**0**	QR 190,000-199,999	**0**
QR 200,000 +	4	QR 200,000 +	3	QR 200,000 +	1
Others [Unable to Specify}	4	Others [Unable to Specify]	0	Others [Unable to Specify]	4
Missing Data: DK: 65 Ref: 33		Missing Data: DK:54 Ref:22		Missing Data: DK: 11 Ref: 11	

Distribution still highly skewed after unfolding

Key Concepts in Scale Construction

RELIABILITY AND VALIDITY

RELIABILITY AND VALIDITY ARE TWO RELATED CONCEPTS THAT REFER TO POSSIBLE MEASUREMENT ERRORS

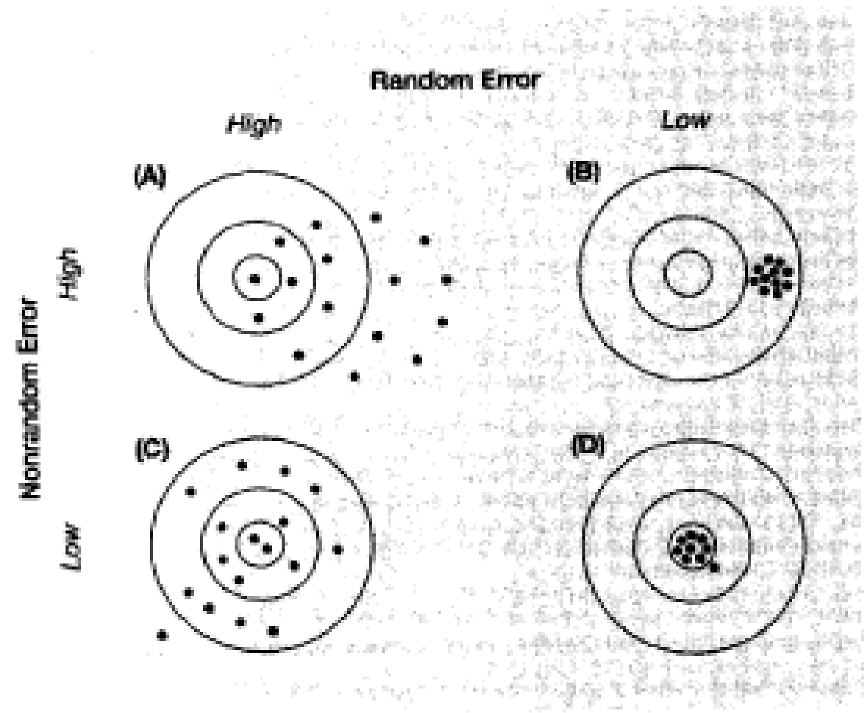
Reliability refers to how consistent or precise the measurement is

Validity refers to whether we are measuring what we think we are (the concept)

Validity and Reliability

In the May 2010 presentations, we defined these terms by referring to non-random and random measurement error.

Figure 4–3 Random and Nonrandom Error



Neither Reliable Nor Valid

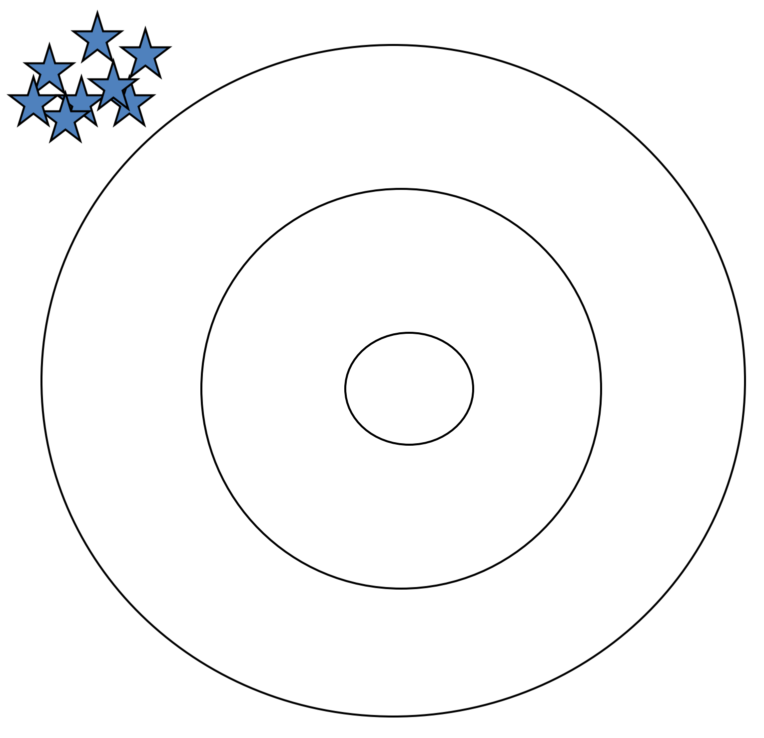
Reliable, But Not Valid

Valid, But Not Reliable

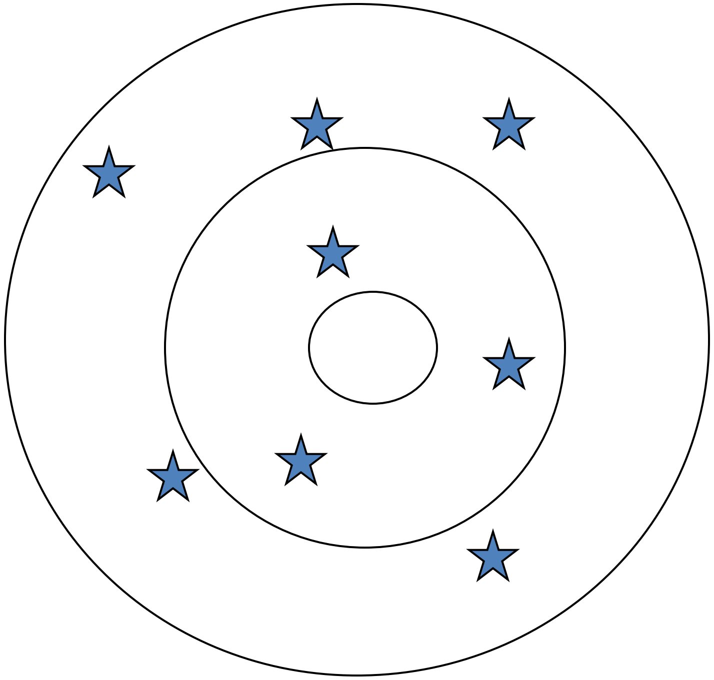
Reliable and Valid

Illustration from Shively, The Craft of Political Research, 6th edition, Pearson Prentice Hall, 2005, p. 49.

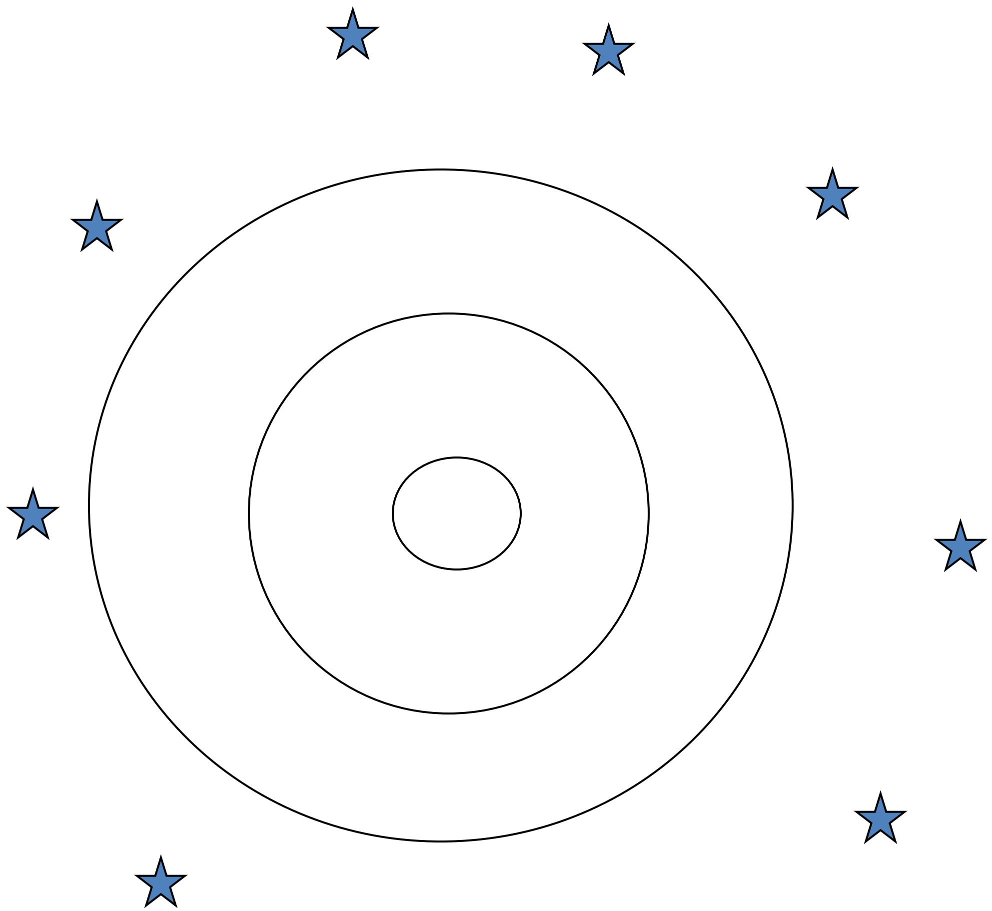
Reliable, Not Valid



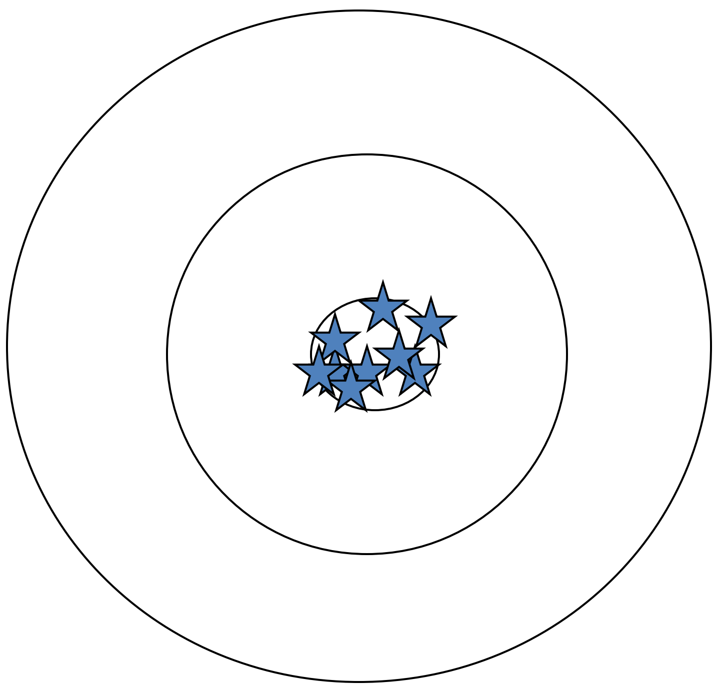
Valid, Not Reliable



Not Valid, Not Reliable



Valid and Reliable



"Assessing Validity"

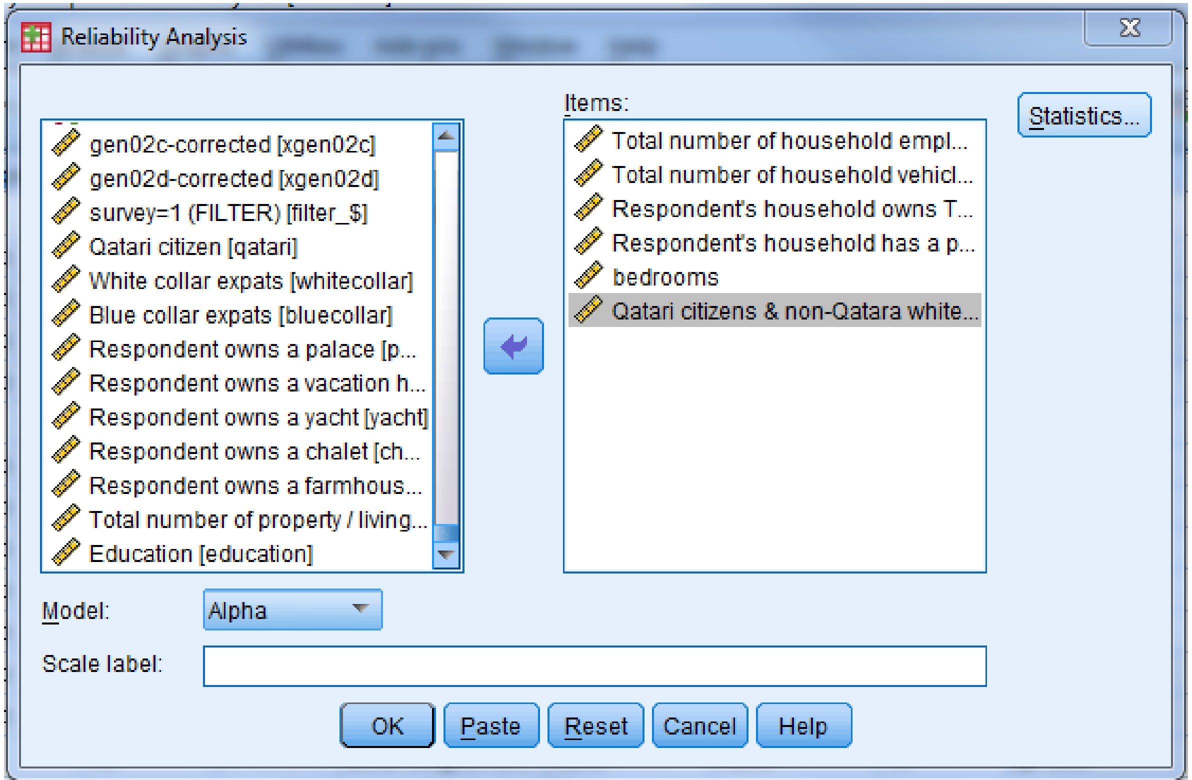
- One way we attempt to assess validity in scale construction is whether the scale we construct is related to other measureable constructs *in a theoretically expected way*.
- Example:
  - In some societies, one would have doubts about one's measure of economic status if it were NOT positively correlated with the status of occupations.
    - People with higher economic status would presumably occupy jobs of higher occupational status, e.g., jobs that are highly respected.
    - Would that be the case in Qatar?
  - If not, what might be a variable – *independent of economic status* – to which one might expect economic status to be related?
    - Could this be used to assess the validity of any measures of economic status that we construct?
- Another way to assess validity is to ask where measures of a concept are differentiated empirically from measures of related concepts, i.e., do these measures exhibit "*discriminant validity*." More about that when we discuss factor scaling.

Assessing Reliability

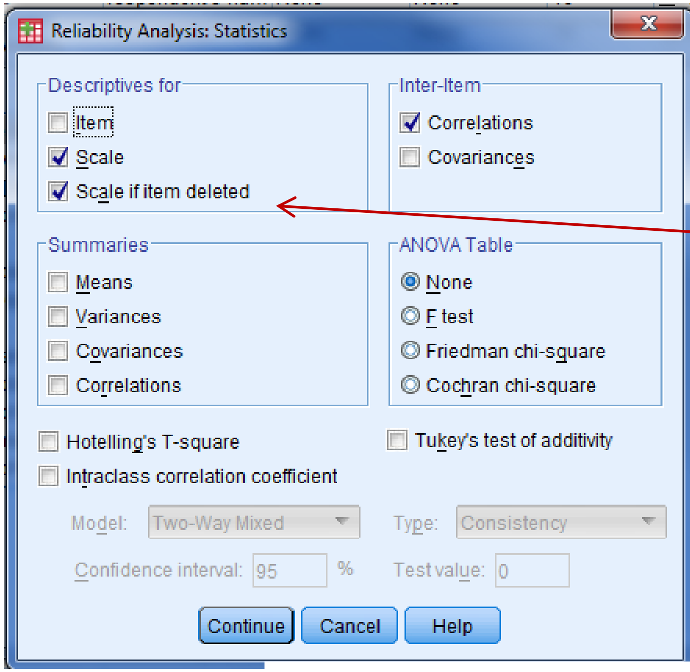
- A scale is considered reliable when the items we use to construct it are closely related. In other words, the scale has <u>*internal consistency.*</u>
- Examining correlations between items can give us one sense of how items are related.
- One way we can measure internal consistency among all the items we may want to scale is by calculating a **Cronbach's Alpha.** This method provides an estimate of reliability.
  - The method generates a coefficient based on the average inter-correlation among the items you may want to scale
  - It produces coefficients that range between 0 and 1. Higher values indicate greater internal consistency.
  - There is some disagreement over what constitutes "good" or acceptable reliability. Generally, coefficients between 0.6 and 0.7 are considered acceptable.

Assessing Reliability

- SPSS provides an option for generating a Cronbach's Alpha in the "Analyze" Menu
- An example: Can we use the measures of income, number of household employees, number of vehicles, and TV possession to construct a scale measuring material wealth among Qataris?
- The Cronbach's Alpha will give us a sense of whether it is appropriate to combine these individual items into a single measure



Assessing Reliability



When calculating a Cronbach's Alpha, you can choose to produce a table displaying the subsequent alpha if each individual variable were deleted. This option can tell us whether an item may not belong in a scale.

Sometimes items may be related, but it is not always appropriate to combine these items into a single measure. Use the information in the last column and your own intuition about the items to make this judgment.

Item-Total Statistics

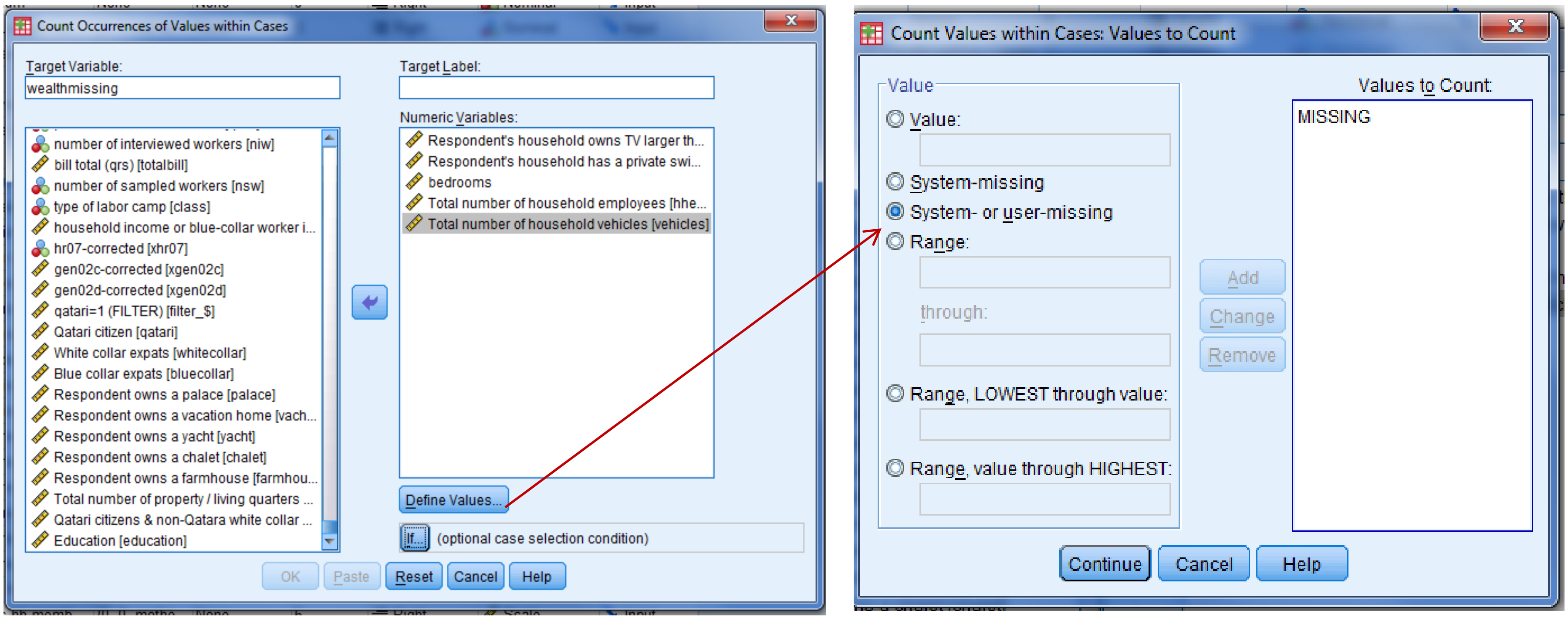
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Total number of household employees	12.2974	33.687	.572	.329	.563
Total number of household vehicles	11.7530	30.117	.535	.304	.577
Respondent's household owns TV larger than 46 inches	14.7148	50.406	.171	.039	.690
Respondent's household has a private swimming pool	15.0487	51.200	.207	.071	.694
bedrooms	10.0226	34.702	.513	.270	.587
Qatari citizens & non-Qatara white collar workers income	11.6070	29.995	.498	.283	.599

Constructing the Scale

- How do we actually combine the items into a single variable?
- We could simply add them and divide by the number of items.
  - The problem with this method is that SPSS will delete cases in which a respondent is coded as missing for at least one of the available items.
- A better method is to create a variable consisting of the mean of the available items
  - So if the scale consists of 4 variables, the new item will be the mean of all 4 items for those who have valid codes for all 4 items. If a respondent has valid codes for only 3 of the items, then the value of the scale for that respondent will be the mean of the available 3, and so forth.

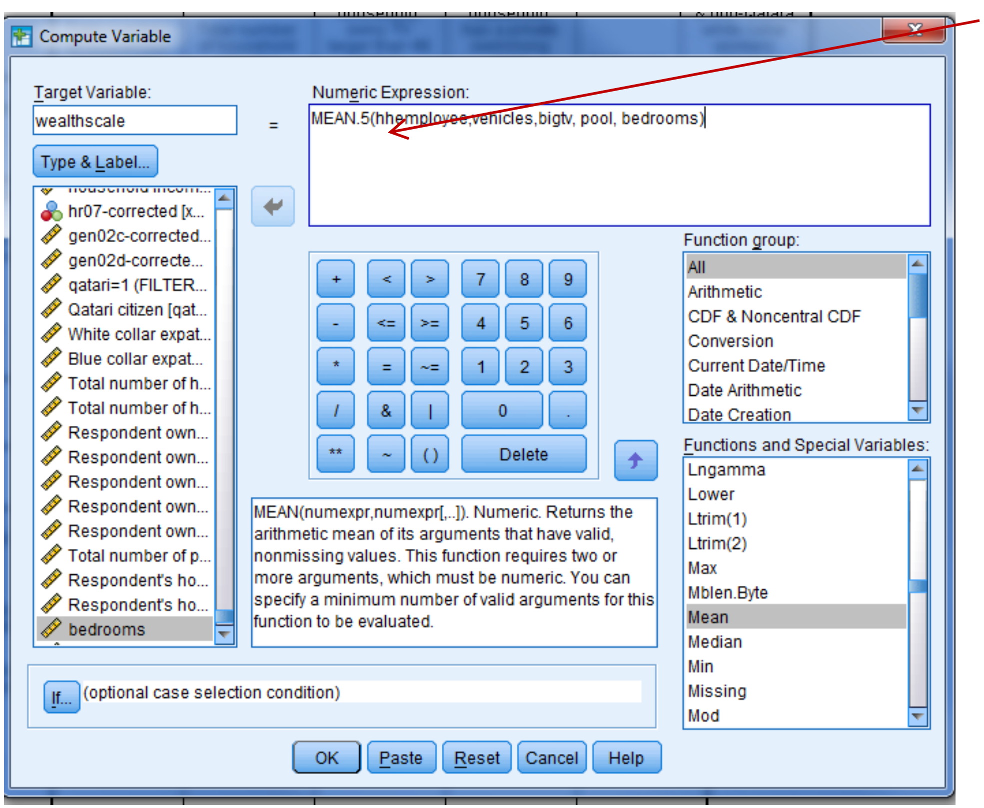
Constructing the Scale

- How do we create a new variable consisting of the mean of the available variables?
- First, we count the number of missing variables and save this information in a new variable.

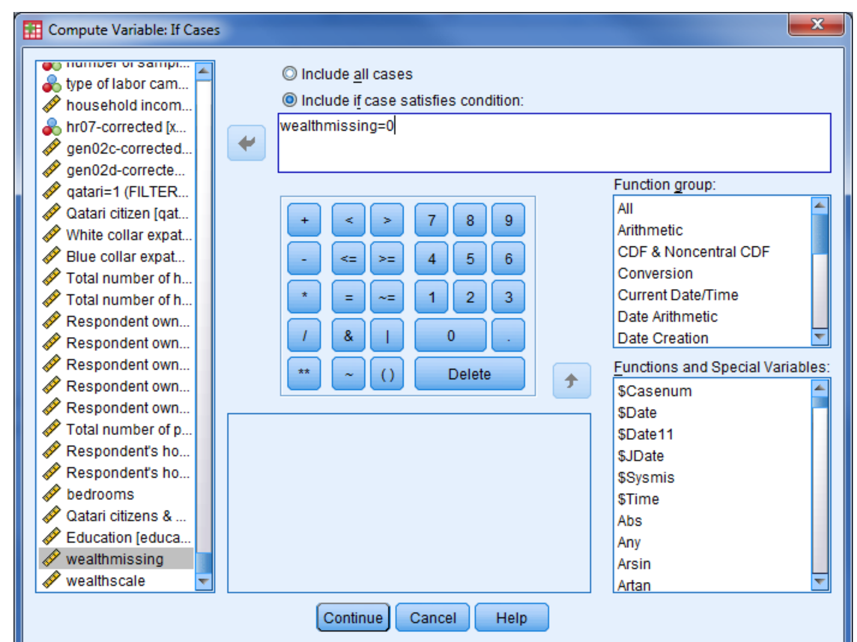


Constructing the Scale

- When creating our scale, we use the generated count variable in a series of “If statements” to tell SPSS how many variables it should use to calculate a mean value.
- The scale in the example consists of 5 items. If a respondent answered all five, the new variable (our scale) will consist of the mean of all five variables. If a respondent only has non-missing responses for 4 of the items, the new variable will consist of the mean of the available 4, and so forth.



“Mean.5(var1, var2, var3, var4, var5)” tells SPSS to take the mean of a total of five variables.



Constructing the Scale

The SPSS Syntax:

```
COUNT
wealthmissing= hhemployee vehicles bigtv pool bedrooms (missing).
EXECUTE .

Compute wealthscale=999.
if (wealthmissing=0) wealthscale=MEAN.5(hhemployee, vehicles, bigtv, pool, bedrooms).
if (wealthmissing=1) wealthscale=MEAN.4(hhemployee, vehicles, bigtv, pool, bedrooms).
if (wealthmissing=2) wealthscale=MEAN.3(hhemployee, vehicles, bigtv, pool, bedrooms).
if (wealthmissing=3) wealthscale=MEAN.2(hhemployee, vehicles, bigtv, pool, bedrooms).
if (wealthmissing=4) wealthscale=MEAN.1(hhemployee, vehicles, bigtv, pool, bedrooms).
if (wealthmissing=5) wealthscale=999.
Missing values wealthscale (999).
```

Another Approach to Scaling:

FACTOR SCALING

Another Approach to Scale Construction, I

- An-often unnoticed feature of the techniques for assessing reliability, which is a common practice in “scale construction,” is that we initially treat each item equally, as if it were an “equally good” measure of the underlying concept.
- Then we perform procedures to “test” that assumption.
- As a result of those procedures, we throw out the measures that don’t seem to fit with the other measures. If a threshold condition is not met, an item will be discarded
- However, there is another way to go about scale construction – one could *weight the various questions unequally*, admitting that all items do not necessarily warrant equal treatment - perhaps not all are equally good measures of the underlying construct.

Another Approach To Scale Construction, II

- Factor scaling addresses the issue of the utility of specific measures in a different way, by assuming two things:
  - One can identify items that co-vary sufficiently strongly to represent that same underlying dimension or factor.
  - But some of those items are “more central” to an underlying structure of co-variation.
  - Items should be weighted proportionately to their participation in the underlying structure of co-variation.
  - One can address discriminant validity via the procedure. Do the same items “load” on the same factor? If not, discard items that do not fit.

Another Approach… III

- One runs varimax factor analysis, extracting factor score coefficients.
- Then one uses those coefficients in a formula like this, assuming that we have three indicators of an underlying concept:
  - Scaled Variable = Factor Score Coefficient Var01 (Var01 – Mean of Var001)/Standard Deviation of Var01 +[or -] Factor Score Coefficient Var02 (Var02 – Mean of Var02)/Standard Deviation of Var02 +[or -] Factor Score Coefficient Var03 (Var023– Mean of Var03)/Standard Deviation of Var03.
- This gives one a variable:
  - That approximates a normal distribution [the subtraction of the mean of each variable from the specific values of the variable, divided by the standard deviation of the variable does this, a procedure known as “standardization”.
  - But the factor score coefficients “weight” the specific items by the extent to which they “define” the underlying factor.

Another Approach…IV

- In the current example, a factor analysis (varimax rotation) was run on six variables: ES05_INC [unfolded income], Total Employees, Total Vehicles in the HH, Number of Bedrooms, Swimming Pool and 46"+ TV. We can see that owning a 46"+ TV is the variable least strongly related to the others.

Correlation Matrix

		Employees	Total of Vehicles in HH	Qatari citizens & non-Qatari white collar workers income	number of bedrooms in hh	ES03 Dummy [Pool]	ES02a Dummy [46" TV]
Correlation	Employees	1.000	.498	.409	.365	.304	.175
	Total of Vehicles in HH	.498	1.000	.372	.416	.168	.075
	Qatari citizens & non-Qatari white collar workers income	.409	.372	1.000	.339	.320	.140
	number of bedrooms in hh	.365	.416	.339	1.000	.137	.128
	ES03 Dummy [Pool]	.304	.168	.320	.137	1.000	.175
	ES02a Dummy [46" TV]	.175	.075	.140	.128	.175	1.000
Sig. (1-tailed)	Employees		.000	.000	.000	.000	.000
	Total of Vehicles in HH	.000		.000	.000	.000	.027
	Qatari citizens & non-Qatari white collar workers income	.000	.000		.000	.000	.000
	number of bedrooms in hh	.000	.000	.000		.000	.000
	ES03 Dummy [Pool]	.000	.000	.000	.000		.000
	ES02a Dummy [46" TV]	.000	.027	.000	.000	.000	

Another Approach… V

- This can also be seen in the factor loadings, in which four variables load on the first factor, while two variables define a second factor.

Note that these four variables load strongly on Factor 1.

Factor 1 exhibits some discriminant validity from Factor 2 by virtue of being a separate factor. However, note that there is a weak loading for Income and for Employees on Factor 2.

Rotated Component Matrix[a]

	Component	
	1	2
Employees	.725	.277
Total of Vehicles in HH	.812	-.020
Qatari citizens & non-Qatari white collar workers income	.633	.327
number of bedrooms in hh	.721	.013
ES03 Dummy [Pool]	.240	.685
ES02a Dummy [46" TV]	-.022	.795

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

While owning 46" TVs and having a swimming pool define second factor.

Another Approach… VI

- To build a factor scale, one would use the Component Score Coefficients [generated by SPSS], as well as the mean and standard deviation of each variable, to create a standardized, but weighted, variable.
- One could build a scale for each factor, but let us focus on factor 1.
- The four included variables would be "weighted" by their overall participation in the structure of co-variation that Factor 1 represents. Hence, each variable is not treated as an exact equal. The weighting happens via the multiplication term.

ES_NEW = .314*(Employees – 3.0978)/2.81818 + .425*(Vehicles-3.6318)/2.75805 + .256*(ES05_INC-2.8365)/2.4441 + . 371*(Bedrooms-5.1557)/2.09448.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N	Missing N
Employees	3.0978	2.81818	665	769
Total of Vehicles in HH	3.6318	2.75805	673	761
Qatari citizens & non-Qatari white collar workers income	2.8365	2.44441	1357	76
number of bedrooms in hh	5.1557	2.09448	689	745
ES03 Dummy [Pool]	.0486	.21509	688	746
ES02a Dummy [46" TV]	.3799	.48572	673	761

Component Score Coefficient Matrix

	Component	
	1	2
Employees	.314	.081
Total of Vehicles in HH	.425	-.197
Qatari citizens & non-Qatari white collar workers income	.256	.146
number of bedrooms in hh	.371	-.148
ES03 Dummy [Pool]	-.027	.544
ES02a Dummy [46" TV]	-.187	.699

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Factor Scales Represent a Standardized and Weighted Scale

- Factor Scales are standardized such that the mean approaches zero [in this case, the mean of ES_NEW is .0527], while the standard deviation approximates 1.0 [for ES_NEW it is 1.00686].

- The other feature of factor scaling worthy of note is that the variables are not weighted equally. Recall the weights:

  - Income [ES05_INC] = .256
  - Total Vehicles in HH = .425
  - Bedrooms [ES04] = .371
  - Total HH Employees = .314

- While not so in this example, there could be negatively weighted items in the scale.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ES_NEW	589	-1.72	8.53	.0527	1.00686
Valid N (listwise)	589				

Possible Class Exercise

- Validity thought exercise: What should our measurement be related to and in which direction? What should economic status predict? What should predict economic status?
- Ultimately, scaling consists of art as well as science. There are some mathematical tools we employ. But we are called upon to make judgments that are "more than mathematical." They include a sense of face validity, and a theoretical logic for why these indicators should plausibly be construed as "measuring the same thing," and a sense of how the scale ought to be related to other known measures [or how it can be distinguished conceptually and empirically from other similar, but measurable, concepts."

The Art of Scale Construction

- In some social sciences, such as psychology, there are long established scales that scholars have come to accept, and their efforts at scale building are essentially “work at the margins,” enhancing or adding to that which most scholars accept.
- In other social sciences, there is much less consensus on scale construction. One is almost starting from scratch in every study.
- In Qatar, SESRI has both the advantage of developing a scaling tradition based, in part, on annual Omnibus surveys, but the disadvantage of sometimes not knowing what one will find. Example: Income distribution in 2010 Omnibus survey.
- One learns and builds over time – from one’s own experience and from that of others.

Summary Questions for SESRI [or users of the SESRI data set] Regarding ES Series

- Is 11% missing data [on ES05] too much to tolerate among Qataris?
- Could we “sell” ES04 [Bedrooms] to consumers of our research as equivalent to ES05. ES04 has no missing data.
- Can we really add anything important by using ES01-ES04a to build a more comprehensive scale ?
  - If we add something, are there good quantitative bases for creating a combined indicator?
- If 11% missing data is too much, can we build a scale that compensates for those missing data?
  - What scale should we construct?

Appendix A: Code for ES05_INC

```
• Compute ES05_INC=999.
• IF (ES05A=1) ES05_INC=1.
• IF (ES05A=2) ES05_INC=2.
• IF (ES05A=3) ES05_INC=3.
• IF (ES05A=4) ES05_INC=4.
• IF (ES05A=5) ES05_INC=5.
• IF (ES05B=1) ES05_INC=6.
• IF (ES05B=2) ES05_INC=7.
• IF (ES05B=3) ES05_INC=8.
• IF (ES05B=4) ES05_INC=9.
• IF (ES05B=5) ES05_INC=10.
• IF (ES05C=1) ES05_INC=11.
• IF (ES05C=2) ES05_INC=12.
• IF (ES05C=3) ES05_INC=13.
• IF (ES05C=4) ES05_INC=14.
• IF (ES05C=5) ES05_INC=15.
• IF (ES05D=1) ES05_INC=16.
• IF (ES05D=2) ES05_INC=17.
• IF (ES05D=3) ES05_INC=18.
• IF (ES05D=4) ES05_INC=19.
• IF (ES05D=5) ES05_INC=20.
• IF (ES05D=6) ES05_INC=21.
• IF (ES05=8) ES05_INC=-8.
• IF (ES05=9) ES05_INC=-9.
• IF (ES05=1 & (ES05A=8 or ES05A=9)) ES05_INC=3.
• EXECUTE.
• IF (ES05=2 & (ES05B=8 or ES05B=9)) ES05_INC=8.
• EXECUTE.
• IF (ES05=3 & (ES05C=8 or ES05C=9)) ES05_INC=13.
• EXECUTE.
• IF (ES05=4 & (ES05D=8 or ES05D=9)) ES05_INC=18.
• EXECUTE.
• Missing val ES05_INC (999,-8,-9)
```

```
• VARIABLE LABEL ES05_INC 'Qatari citizens & non-Qatari white collar workers income'.
• VALUE LABELS ES05_INC
• 1 Less than QR10,000'
• 2 'QR10,000 to less than QR20,000'
• 3 'QR20,000 to less than QR30,000'
• 4 'QR30,000 to less than QR40,000'
• 5 'QR40,000 to less than QR50,000'
• 6 'QR50,000 to less than QR60,000'
• 7 'QR60,000 to less than QR70,000'
• 8 'QR70,000 to less than QR80,000'
• 9 'QR80,000 to less than QR90,000'
• 10 'QR90,000 to less than QR100,000'
• 11 'QR100,000 to less than QR110,000'
• 12 'QR110,000 to less than QR120,000'
• 13 'QR120,000 to less than QR130,000'
• 14 'QR130,000 to less than QR140,000'
• 15 'QR140,000 to less than QR150,000'
• 16 'QR150,000 to less than QR160,000'
• 17 'QR160,000 to less than QR170,000'
• 18 'QR170,000 to less than QR180,000'
• 19 'QR180,000 to less than QR190,000'
• 20 'QR190,000 to less than QR200,000'
• 21 'QR200,000 or more'
• -8 'DON'T KNOW'
• -9 'REFUSED'
• 31 'Less than QR50,000'
• 32 'QR50,000 to less than QR100,000'
• 33 'QR100,000 to less than QR150,000'
• 34 'QR150,000 or more'.
• FREQUENCIES VARIABLES=ES05_INC
•   /ORDER=ANALYSIS.
```

Appendix A [continued]:
ES05_INC Among Qataris and White Collar Ex-Pats

Descriptive Statistics[a]

household type		N	Mean	Std. Deviation
1. qatari	Income	688	2.9326	6.32757
	Valid N (listwise)	688		
2. white collar	Income	767	2.4459	5.00229
	Valid N (listwise)	767		

a. No statistics are computed for one or more split files because there are no valid cases.

Appendix B: Missing Income Data in Other Surveys

- In the 2008 American National Election Study, 2.76% respondents were coded as “refused” and 3.14% were coded as “don’t know.”

- In the 1990 American National Election Study, 5.76% of respondents were coded as “refused” and 3.64% were coded as “don’t know.”

Appendix B [Missing Data on Family Income in the Americas, 2010]

National Sample	N	N Offering Fam. Income Data	Missing%
Mexico	1,562	1,393	11
Guatemala	1,504	1,344	11
El Salvador	1,550	1,464	6
Honduras	1,596	1,504	6
Nicaragua	1,540	1,451	6
Costa Rica	1,500	1,170	22
Panama	1,536	1,488	3
Colombia	1,506	1,350	10
Ecuador	3,000	2,818	6
Bolivia	3,018	2,554	15
Peru	1,500	1,371	9
Paraguay	1,502	1,181	21
Chile	1,965	1,676	15
Uruguay	1,500	1,402	7
Brazil	2,482	2,363	5
Venezuela	1,500	1,360	9
Argentina	1,410	1,132	20
Dominican Republic	1,500	1,333	11
Haiti	1,752	1,629	7
Jamaica	1,504	1,222	19
Guyana	1,540	1,314	15
Trinidad & Tobago	1,503	1,151	23
Belize	1,504	1,353	10
Suriname	1,516	1,342	11
United States	1,500	1,463	2
Canada	1,500	1,485	1

Data from Latin American Public Opinion Project, Vanderbilt University, Barometer of the Americas, 2010.
Face to face national surveys, except for shorter telephone surveys in the US and Canada.

Appendix C: Class Exercise

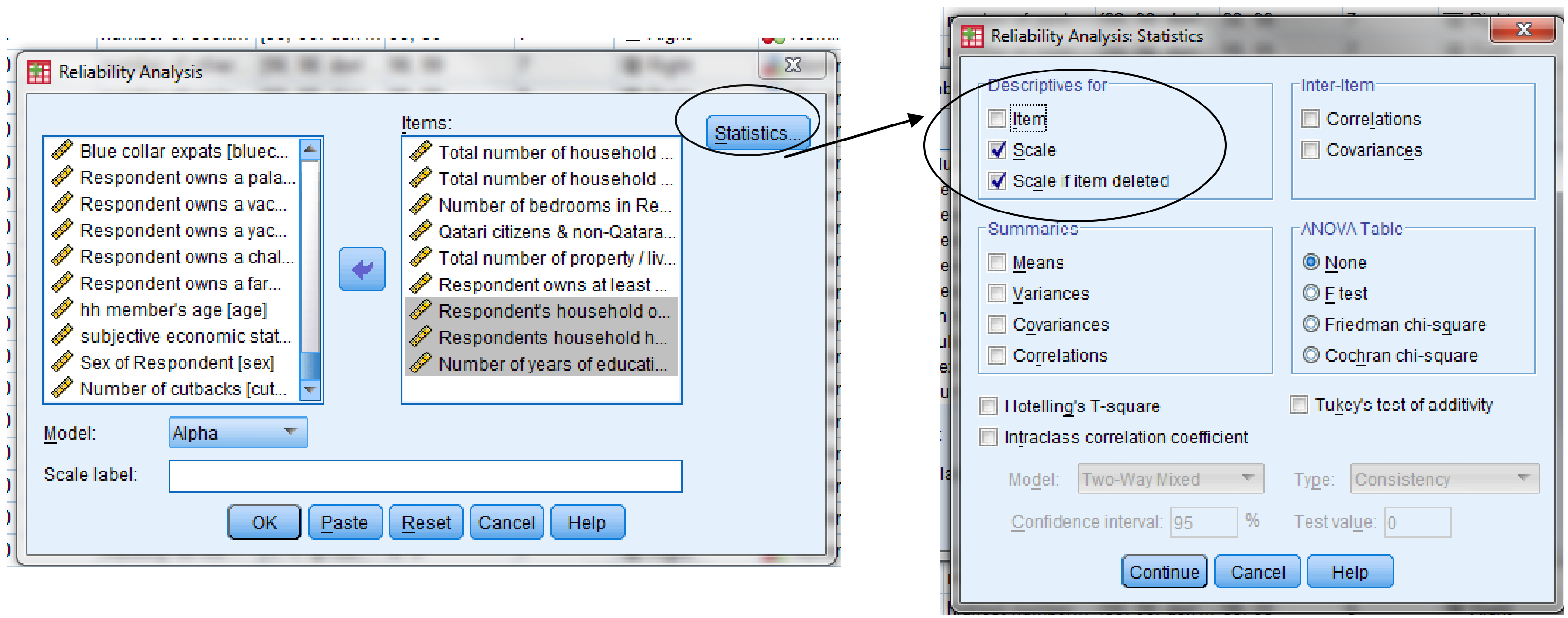
Practicing Reliability Analysis

There are several measures of economic status in the SESRI Omnibus survey. Let us say that you wanted to choose from the following items coded in **DATASET 2** to construct a scale of *socioeconomic status* among Qataris.

Variable Name	Description
hhemployee	Total number of household employees
Vehicles	Total number of vehicles
property	Respondent owns either a palace, vacation home, yacht, chalet, or farmhouse
propertycount	Number of additional properties (as listed above) owned
bigtv	Respondent owns a TV bigger than 46 inches
pool	Respondent's household has a private swimming pool
bedrooms	Total number of bedrooms in Respondent's household
ES05_inc	Household income
education	Number of years of education

Appendix C: Class Exercise

We can use reliability analysis to help determine which of the items should go into a single measure of socioeconomic status. We can conduct a reliability analysis from the **Analyze / Scale / Reliability analysis** menu:



Appendix C: Class Exercise

If we conduct the Cronbach's Alpha analysis with all nine of the above variables, we get the following:

Reliability Statistics

Cronbach's Alpha	N of Items
.588	9

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Total number of household employees	25.8380	55.023	.472	.503
Total number of household vehicles	25.3624	55.141	.482	.501
Respondent owns at least one additional property	28.5072	71.499	.145	.591
Total number of property / living quarters owned by respondent's household	28.4358	70.809	.133	.590
Respondent's household owns TV larger than 46 inches	28.3146	70.840	.187	.588
Respondents household has a private swimming pool	28.6528	72.029	.157	.594
Number of bedrooms in Respondent's household	23.5069	56.912	.379	.529
Qatari citizens & non-Qatara white collar workers income	25.0684	44.057	.577	.439
Number of years of education.	15.8759	46.611	.217	.652

Out of the above nine variables, select the first three you would eliminate from the scale. *Remember that an alpha between .6 and .7 (or higher) is generally considered acceptable*. Choose the variables that if deleted, will most improve the alpha level of the scale.

Appendix C: Class Exercise

Question 1: Which three variables did you delete?

Run the Reliability Analysis yourself, but instead of replicating what's above, eliminate the three variables you decided should be eliminated.

Question 2: What is the resulting Cronbach's Alpha?

Question 3: Now that you've eliminated three of the variables, are there anymore you can remove to subsequently improve the alpha level? If so, which variables?

Run the analysis again, this time deleting the selected variables from the scale.
Question 4: What is the resulting Cronbach's Alpha?

Question 5a: Can we improve the alpha level by further removing variables from the scale? If we could, which variables would we delete?

Question 5b: If we can't improve the reliability statistics, why not?

Appendix C: Class Exercise

The Cronbach's Analysis with all 9 potential measures of socioeconomic status

Reliability Statistics

Cronbach's Alpha	N of Items
.588	9

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Total number of household employees	25.8380	55.023	.472	.503
Total number of household vehicles	25.3624	55.141	.482	.501
Total number of property / living quarters owned by respondent's household	28.4358	70.809	.133	.590
Respondent owns at least one additional property	28.5072	71.499	.145	.591
Respondent's household owns TV larger than 46 inches	28.3146	70.840	.187	.588
Respondents household has a private swimming pool	28.6528	72.029	.157	.594
Number of bedrooms in Respondent's household	23.5069	56.912	.379	.529
Qatari citizens & non-Qatara white collar workers income	25.0684	44.057	.577	.439
Number of years of education.	15.8759	46.611	.217	.652

Appendix C: Class Exercise

The Cronbach's Analysis removing property, pool, and education

Reliability Statistics

Cronbach's Alpha	N of Items
.638	6

The alpha with these variables is .638, so we want to consider removing items that will raise the alpha above that level.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	
Total number of household employees	12.7633	31.324	.549	.515	
Total number of household vehicles	12.2315	29.566	.464	.556	
Total number of property / living quarters owned by respondent's household	15.4418	46.768	.271	.641	✖
Respondent's household owns TV larger than 46 inches	15.3103	48.286	.140	.656	✖
Number of bedrooms in Respondent's household	10.6024	35.711	.414	.577	
Qatari citizens & non-Qatara white collar workers income	12.1345	29.409	.449	.566	

Appendix C: Class Exercise

The Cronbach's Analysis removing propertycount and bigtv

Reliability Statistics

Cronbach's Alpha	N of Items
.672	4

Here we see that deleting none of the remaining variables will improve the alpha level above .672.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Total number of household employees	12.1331	28.695	.525	.564
Total number of household vehicles	11.6170	26.491	.470	.595
Number of bedrooms in Respondent's household	10.0049	32.483	.413	.634
Qatari citizens & non-Qatara white collar workers income	11.4904	26.339	.431	.628