Paper SP01

AN ANIMATED GUIDE: MATCHING TEST AND CONTROL SUBJECTS

Russell Lavery, Independent Consultant

ABSTRACT

Matching of similar test and control subjects is a part of the experimental process in many disciplines. It is done to reduce sum-square error and selection bias; a threat to internal validity of the experiment.

The correct experimental process is to match subjects and then randomly assign one of each pair to be a test subject. Occasionally a researcher will be given a group of test subjects and told to find matching controls. While the results are not as logically valid as those generated by the preferred method, this request is made with some regularity. Accordingly, this paper concentrates on this less than optimal situation and the CD includes SAS® code designed to handle this circumstance. The code can easily be modified to handle the more valid procedure.

INTRODUCTION

The program that does this matching (complete with sample data sets) and two handouts on threats to validity are in appendices.

Matching, especially without randomization, does not assure that the test and control groups are equivalent and has been strongly criticized.

The strongest justification for matching is when you are investigating a set of variables and you want to eliminate the effect of other variables that have unknown form. A well-formulated ANCOVA is extremely effective in reducing sum-square error but requires that the form of the relationships between the Xs and Y be known. Matching can, if well done, let you eliminate the effects of variables without knowing their form (Figure 1).

Matching involves making many judgments that cannot be checked for correctness. You cannot *know* if you have found the best possible set of matches. Examining and justifying decisions at each step can achieve a reasonable matching, and that is all that can be done.

The algorithm in this paper allows for a great flexibility in the definition of best pairings and incorporates some techniques that make the program fast enough to be used on very large data sets.

The matching algorithm presented here creates "best pairings" based on three types of criteria:

1) Absolute criteria (e.g. control and test must have the same gender and marital status)
2) Closeness criteria (e.g. closest on age, income and years of schooling)
3) Qualification criteria (e.g. must have used product in last three months and subscribe to magazine X)

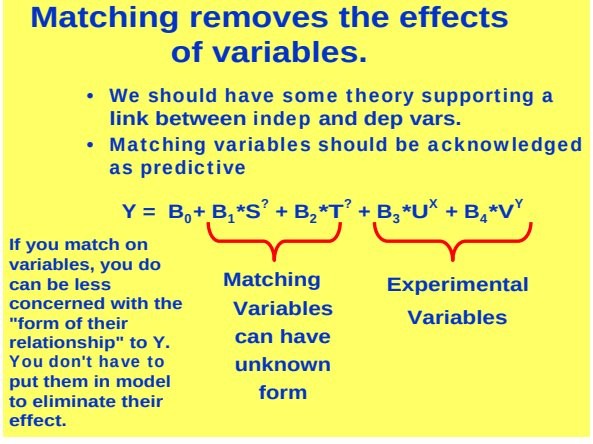

FIGURE 1

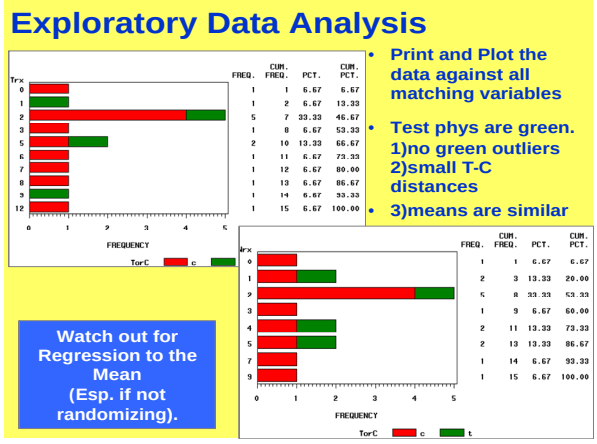

FIGURE 2

METHOD

This paper will describe the steps in a matching process and digress to explain issues that are critical to a particular step. The steps in the matching process are:

1) Examine the theory from the discipline for candidate X variables and links between X and Y variables
2) Flowchart the experimental process
3) Look for threats to the validity of the experiment
4) Do Exploratory Data Analysis (EDA) to investigate the data
5) Select matching variables, specifying business filters, matching criteria, distance measurement and distance filter
6) Specify best matching logic
7) Modify and run the program

The example used in this paper will be the hypothetical matching of test physicians with their best match from a group of possible control physicians. The experimental issue is the effectiveness of a new drug sample pack to be given to the physicians. The business goal for this campaign is to convert high volume competitor loyal physicians to your product.

Assume you have a moderately sized list of such competitor loyal physicians that will get the new drug sample pack. Management considers the business (convert these people to our drug) need to be so strong that they insist that all of these physicians get the new sample pack. Of course, they also insist that you still measure the effectiveness of the pack. The reason the client gives for not splitting this competitor loyal group into test and control groups is that the new sample pack can't hurt sales and might well help. Additionally, these physicians are thought to be so important to the business that the sales force must try to convert all of these people to the new product immediately.

This is the situation where you are not allowed to randomize to test and control but are forced to match controls to a pre-existing group of test subjects. Your controls will have to come from outside the list of important competitor loyal physicians but must resemble them as closely as possible.

THE PROCESS

STFEP 1: EXAMINE THEORY FROM THE DISCIPLINE

Matching should not be done on convenience variables -- variables that happen to be in your data set. X variables should be known to have a predictive effect on the Y variable. You can know that an X variable is predictive because of a) your own research, b) the knowledge of experts c) scholarly research or d) a combination of the above. The more theoretical input at the beginning of the process, the better the business results.

STEP 2: FLOWCHART THE PROCESS

You need to decide when and where you will measure your data to insure that your sample/measurement point will provide valid measurements of the business process under study. You want to function with independence as you program and do not want to be constantly asking questions. Understanding the experimental process – and where it can go astray – will allow you to deliver a high quality product with minimal time spent asking your client questions.

If you knew just what the client was planning to do, you could ask a small list of focused questions and get all the information you need to do your job. However, until you know the experimental process, you cannot ask those focused questions. If the client knew what you require to do your job, he could tell you just the information you needed to know and warn you of the pitfalls you might encounter. However, it is an unusual client who understands what it is that you need to do your job.

Since neither of the parties in the project can, a-priori, instruct the other, you need to get a general understanding of the process and use that understanding to ask for the information you need. The best way to get a general understanding is to flowchart the process. After creating the flowchart, you should be able to ask questions whose answers you need to do your job. A flowchart is an excellent document for reviewing the experimental plan and for revealing those threats to validity for which proper matching provides no cure.

STEP 3: LOOK FOR THREATS TO VALIDITY

Generally, a valid experiment is a good experiment. A valid experiment will be reproducible and give accurate results.

Nothing, not even correct coding, is more critical to the business success of the experiment than is the careful analysis of threats to experimental validity. You must consider several kinds of experimental validity and the various threats to each. Cook & Campbell and Campbell & Stanley are the authors cited most often on issues of threats to experimental validity.

STEP 4: DO EDA TO INVESTIGATE THE DATA

Do some Exploratory Data Analysis (EDA) to see if the group of test physicians should even be matched to the controls. Check to see that the means for the two groups are about the same. If they are not, consider not matching at all and using another statistical technique.

If the tests and controls are not "well mixed" in the EDA plots, the success of the project is in doubt because you will likely have regression to the mean and get erroneous statistical results. Statistically speaking, the EDA checks if the two groups appear to have come from the same parent population. In addition, you should look for outlier test subjects that cannot be matched by any control.

Finally, in an effort to reduce the size of the program's SQL Cartesian product, see if you are willing to make an estimate of what constitutes an unacceptably distant match between subject pairs. THE EDA will let you get a feel for the data and make this guess. This estimate will be put into the WHERE clause of a SQL join and is intended only to reduce processing time.

Looking at Figure 2, you might conclude that the controls and the tests are fairly well mixed and that there are several controls near every test. Because the tests have many options for matching and you want to keep your files small, you might eliminate, in the SQL join, any pair where the distance between pairs is greater than, rather arbitrarily, seven. Note that this rough distance filter might be modified in the following step as the distance measurement is formalized. This code runs quickly and can be re-run if the value for an "unacceptably distant match" does not let the algorithm find enough matches.

STEP 5: SELECT MATCHING VARIABLES

One important characteristic of a physician is her/his prescription volume. Volume is measured by counting the number of prescriptions she writes -- information that can be purchased from suppliers. Volume is a theoretically justified variable that is known to be correlated with many physician behaviors, including new product adoption. Common physician matching variables are NRx (number/volume of new prescriptions filled at reporting pharmacies that month) and TRx (Total number/volume prescriptions filled at reporting pharmacies that month, where TRx =NRx + renewals).

Specify Business Filters

The business problem might require that you filter out some of the tests and controls as unusable. As an example, experts might specify that the controls must have some non-zero volume on both NRx and TRx for the last three months. This means that any physician who has not prescribed in the last three months should be filtered out of the data set. Those physicians might be sick or on an extended vacation and could bring unwanted variability into the data set.

This process of exclusion should have some business reason. It is a matching pre-process and can be implemented in an early data step.

Specify Matching Exactly Criteria

Physician specialization affects prescribing patterns. Orthopedic physicians prescribe drugs for earaches less frequently than do family practice physicians. Because specialty affects behavior, physicians are generally paired within their own specialty.

Board certification is evidence of study past the awarding of the medical license and indicates interest in keeping current on new therapies. Physicians are often paired with someone who has their level of board certification.

There are regional differences in the practice of medicine due to training differences and the influence of managed care. In addition, diseases tend to have a geographic pattern that also can impact prescribing patterns. Accordingly, physicians are often matched by Metropolitan Statistical Area (MSA).

This condition of absolute equality has implications. If we are matching on gender, any male test physician would match with all male control physicians. Additionally, there is no ability to rank matches based on quality of match using this type of criteria. All exact matches are equally good. Exactly matching criteria reduce the number of rows that the SQL matching returns, but does not allow us to judge the "quality of a match".

Specify Matching Closely Criteria
There is a geometric aspect to the problem of matching people who are similar on continuous or nearly continuous variables. When a client says similar, he is talking about people who will plot close to each other in the "problem space".

The problem space is a space defined by the continuous variables that pertain to the problem. In this example we are matching on TRx and NRx and they are the two axis used in Figure 3.

Subjects who are similar, plot close together in the problem space. Plotting subjects in problem space is often very helpful

Specify Distance Measurement
Calculating distance is more complex than is commonly assumed. There are many ways to measure distance and a review of the clustering literature will turn up a large number of measures from which to choose. Fortunately, simple and familiar measures are often applicable.

Figure 3 shows a plot of some test and control physicians. A commonly specified measure of similarity is Euclidian distance. Euclidian distance is the distance measure we all learned in high school:
Distance= $(\Delta Nrx^2 + \Delta Trx^2)^{1/2}$

Euclidean distance often works well, or at least cannot be proven to work poorly.

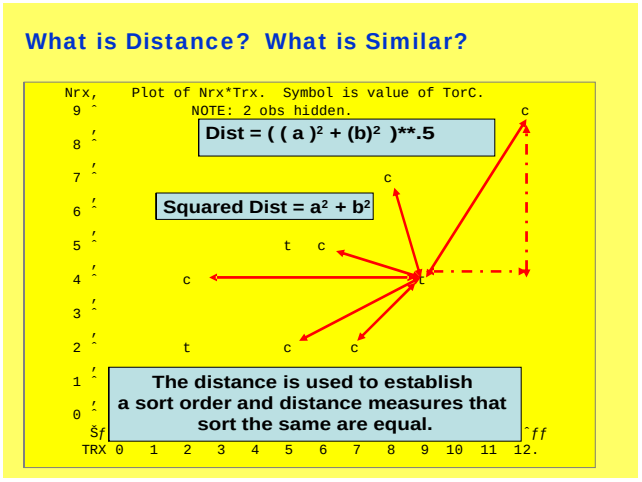

FIGURE 3

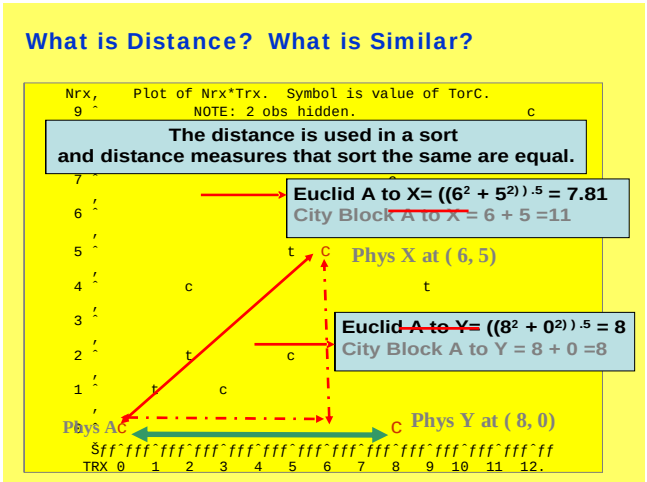

FIGURE 4

Specify Distance Filter
Finally, we should use the EDA to create a bad match filter -- physicians more than a certain distance apart should not be considered for matching.

An SQL merge in the program will consider all combinations of test and control pairs. If we require that each such pair must be within some distance of each other, the computer files will be smaller and the program will run in less time.

In Step 4 above we proposed that seven units would be a reasonable distance limit for a good match. If we use the Euclidean distance we do not need to adjust that number. If we were to use one of the distance measures mentioned in the following digressions, we might have to change that value.
The following digressions discuss some of the problems you must face when deciding on variables and a distance measure.

DIGRESSION: What is similar?
Measuring the similarity of subjects is a more complex problem than is usually assumed. Usually we do not calculate

similarity. Instead we calculate how different things (dis-similarity) are, then declare that things that are not very different are similar.

DIGRESSION: What is contention for a test or for a control?

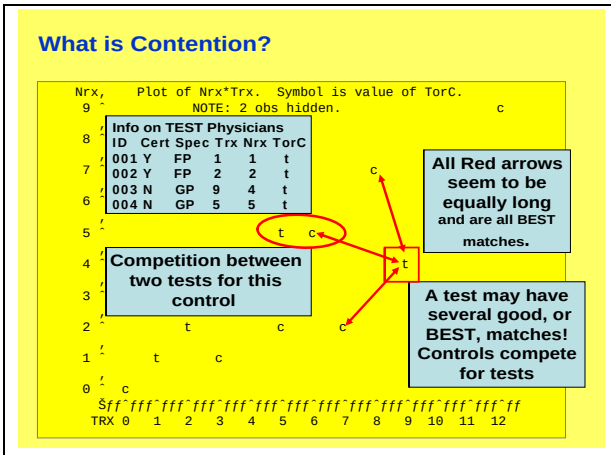

Figure 5

In figure 5 there are three controls that seem to be equally distant from the test in the red square. These controls are contending to be selected as a control for that test.

It should be noted that the control in the red oval is the best control for the test in the red oval. The tests in the red square and red oval are contending for the control in the red oval. While the control in the red oval is the best control for two different tests, it is a much better match for the test that is also in the red oval.

Any algorithm should recognize this contention and match up the test and control inside the redo oval. The test in the red square has other (visually) equal good controls to which it can be matched..

DIGRESSION: What is psychologically similar?
We often use easily obtained physical measurements (TRx, NRx or market share) as proxies for difficult to obtain psychological measurements like brand loyalty.

Consider the practice of using a brands's market share (count of a physicians prescriptions to a brand / count of total prescriptions) as a measure of physician brand loyalty. Many variables, other than a physician's personal preference for a drug, can influence what she prescribes. Behaviors are very imperfect proxy variables for psychological states. Many more men would want to own a Ferrari than actually buy, for example.

DIGRESSION: The research influences the matching process.
The nature of the research can have a tremendous impact on the matching process and a story will illustrate this. The author ran cross-country in school and was the slowest runner on his school's very good team. In a third of the meets, the author, while last on his team, finished before the first runner on the opposing team.

For social research (measuring the sociological effects of team membership), the author would likely have been paired with the worst runner on the opposing team. Physically, the author's closest match would have been the best runner on the opposing team, and there would have been no good "physical matches" for any of the remaining runners on the author's team.

DIGRESSION: Distorting the problem space to weight variables by their importance
Sometimes a client will say that one variable is "twice as important" as another variable. If this is truly the case, a simple multiplication will do. If it is twice as important to be as close on NRx as on TRx, for example you could make

$$\text{Distance} = ((2*\Delta Nrx)^2 + \Delta Trx^2)^{\frac{1}{2}}$$

However, since the distance we are using is a psychological distance, it is actually very difficult to determine proper weighting in these distance calculations. It's unlikely that your client can tell if the formula for psychological distance should truely be

$$\text{Distance} = ((2*\Delta NRx)^2 + \Delta TRx^2)^{\frac{1}{2}}$$

or

$$\text{Distance} = ((2*\Delta NRx^2) + \Delta TRx^2)^{\frac{1}{2}}$$

While it is usually not possible to validate the correctness of a distance formula -- researching scholarly journals for validated scales offers some hope -- this is a critical step in the matching process. The use of an improper distance formula can result in significant changes in the business solution that results from the process.

If your client says that one dimension is twice as important as another, it's often helpful to transform the variables and make plots of data (actual data or dummy) in both the old and new problem spaces. It's often easier to have a meeting and have clients select the “better, more realistic plot” than to try to describe the effects of these space-distorting transformations.

DIGRESSION: How the distance between points is used in sorting
It should be noted that the algorithm in this paper does not use distance directly. It sorts the paired observations by the distance between each pair.

All formulas for calculating distance between physician pairs that sort physician pairs in the same order will give the same business answer.

In Figure 4, we could use
Distance= $(\Delta NRx^2 + \Delta TRx^2)^{\frac{1}{2}}$
or
Distance= $\Delta NRx^2 + \Delta TRx^2$
and get very different numbers for the distances between pairs.

However, if we used both of these formulas and calculated the distances between each point in the data set with every other point (the SQL Cartesian product), then sorted the data set from smallest distance to largest, they would sort in the same order. In this case, it does not matter which of the two formulas you use. The order in which the values sort is what is important to the business solution, not the absolute distance between pairings.

DIGRESSION: Not all distances are the same
Not all distances sort in the same order. In Figure 4 we compare Euclidean distance with another common distance measure – City Block distance. City Block distance measures distance the way you do when you give directions in a city. If you have to walk three blocks east and five blocks south to get somewhere, you say it is eight (3 + 5) blocks away. The business logic of a problem can, sometimes, make this the most appropriate distance measurement.

In Figure 4 we calculate the distance between physician A and physicians X and Y. If we use Euclidean distance, X is closer to A. If we use City Block distance, Y is closer to A. These distance measures would not sort the test-control pairings in the same order, so switching between Euclidean and City Block might well change the results.

There are many additional distance measures that your client might request. Knowing that measures that sort the data into the same order will result in the same results can save a lot of time. Should a client ask you to use a new distance measure and you can quickly tell that it will not change the sort order, you can simply report that the new measure will not change the results already derived.

Clustering research is concerned with the grouping together of similar objects and has developed many specialized distance measures. Romesberg's book explains many distance measures and is very easy to read.

DIGRESSION: Drawing pictures is useful
Try graphing points and calculating differences for a few made-up pairings and see if your distance formula makes sense. Create pairings with small distances, large distances and, very importantly, zero distances. Pair points manually and see if the results make sense. Associate names of people with some of the data points and try to tell stories about how the people behave. If you can't make up a believable story that incorporates your points and your distance measure, something likely does not reflect reality and needs additional work.

DIGRESSION: Standardizations and normalizations
Judge which of the following pairs you would consider to be closer (here just on one variable, TRx):

Pair 1: TestTrx=55 Control Trx=60 Diff.=5
Pair 2: TestTrx=2 Control Trx=5 Diff.=3

Obviously, Pair 1 would plot farther apart than Pair 2 because five is a larger distance than three. But maybe a physician writing sixty prescriptions a month is, in a business sense, more similar to the physician writing fifty-five prescriptions than are the two physicians writing two and five prescriptions respectively.

It is common to normalize distances by dividing the distance between the test and control by the value of the test. (Since you will be matching many controls to a test, it is common to use the test value in the divisor of the formula). You could standardize with (control –test) / test as:

Standardization Method 1.
distance-pair 1 = (60-55) / 55 = .091
distance-pair 2 = (5-2) / 2 =1.5

or with a variation like this one:

Standardization Method 2.
distance-pair 1 = 60 / 55 = 1.091
distance-pair 2 = 5 / 2 =2.5

Either of these normalizations makes Pair 1 much closer together than Pair 2.

DIGRESSION: Look for logical discontinuities

Frequently, a distance formula will be mathematically defined over a large set of conditions but only makes business sense in a subset of those conditions. Often a formula should only be applied when a component of the distance formula is within a certain range. The next two figures will illustrate a few problems caused by logical discontinuities.

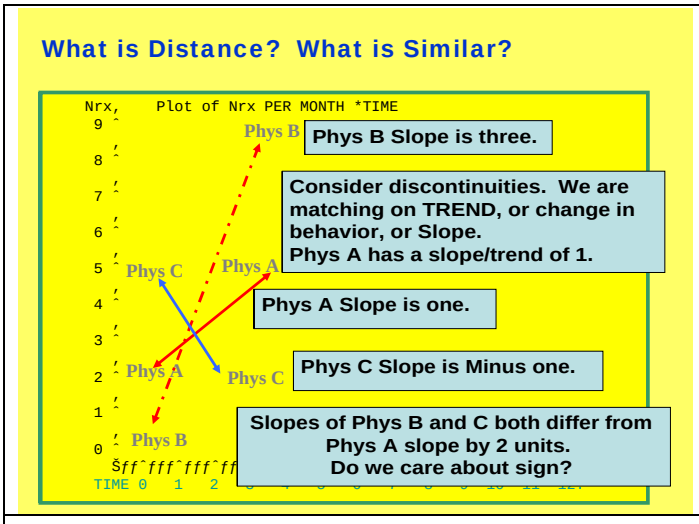

FIGURE 6

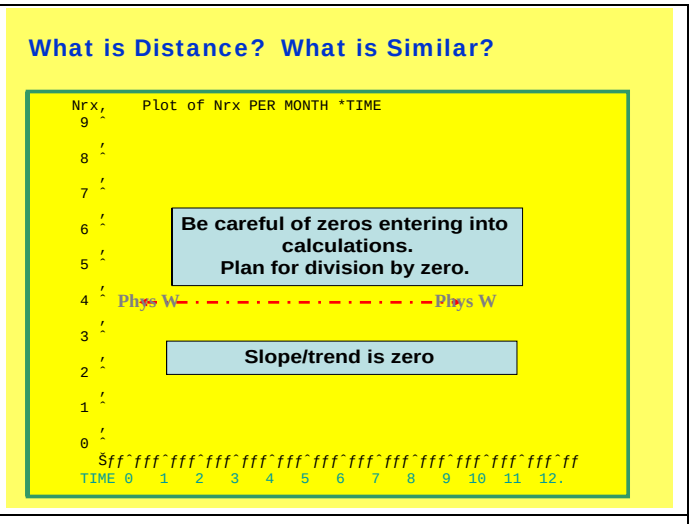

FIGURE 7

In Figure 6 the client has asked that physicians be matched on trend. We know trend is measured as a slope.

The variable of interest is adoption rate (slope) for the product, not the current usage. Figure 6 plots NRx vs. Time. Imagine that physician A is the test and physicians B and C are potential controls. Physician A has a slope of 1. Physician B has a slope of three and the trends (physician B slope – physician A slope) differ by 2. Physician C has a slope of –1 and the slopes (physician A slope - physician C slope) differ by 2. Since the differences in trends are equal in magnitude, you might assume that physicians B and C are equally good matches for physician A.

However, physicians A and B are increasing their usage and physician C is decreasing her usage. For business reasons, you might want to declare that there is a logical discontinuity at zero slope and decide to never match physicians with slopes of different signs. Plotting helps uncover these issues.

DIGRESSION Watch for zeros and missing values

Since zeros and missing values can both cause problems in calculations, you should look for ways that zeros can creep into your work. Zero and missing value problems can occur even when all physicians have non-zero values for the matching variables. Figure 7 illustrates just one of the ways zeros can creep into your calculations.

As shown, if a test physician had no change in her prescribing habits, her trend (slope) would be zero. We have seen that dividing the control value by the test value is a useful normalizing process. However, if you wanted to match on normalized slope, and physician W was a test physician, you would end up dividing by zero.

There can also be problems when the divisor is close to zero. If the slope of the test physician were .05 (a change of 1 Rx over 20 weeks), the divisor would likely be unstable over time (and also small). Sometimes, you will place maximum limits on the results of normalizing with small denominators. Controls are less likely to cause problems with formulas.

Note that standardization method 1, in the previous section, is much more likely to produce zero values, than

standardization method 2. (e.g. 55-55 / 55).

STEP 6: SPECIFY YOUR "BEST MATCH LOGIC"

You must decide on the matching logic for pairing your tests and controls. To use this program, your best match logic must be implemented with a distance calculation and a sort. There are several such methods from which to choose.

Figure 8 shows a data set of paired physicians with no distance greater than seven, sorted by distance. The prefixes T and C indicate test and control. In addition,

Id = Id Number",
Spec ="Physician Specialty",
Brd="Board Certified",
Trx="Total Rx in a Month"
and Nrx="New Rx in a Month".

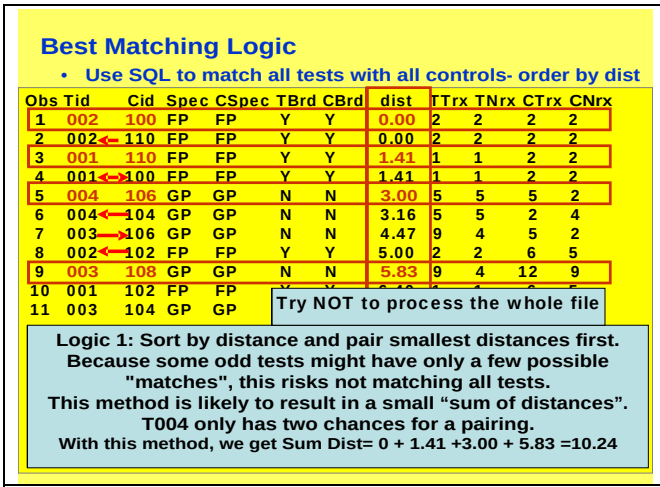

FIGURE 8

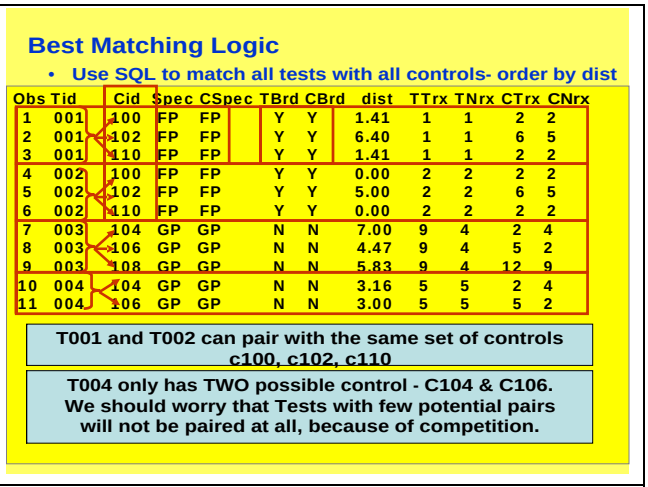

FIGURE 9

BEST MATCH LOGIC #1: Create a filtered Cartesian product and sort by distance

The program logic works, through the file presented in Figure 8, from top to bottom. It selects pairings when it finds a test-control pair where both the test and the control have not been previously assigned.

Accordingly, the program selects Obs 1. It does not select Obs 2 because Tid=002 has already been assigned. It selects Obs 3 but does not select Obs 4 because both the test and control physicians have already been paired. It selects Obs 5 but not Obs 6, 7, or 8. It selects Obs 9.

Finally, we knew at the start of the project that we were attempting to pair N (here n=four) test physicians. The program counts how many successful pairings it has made. It stops executing at Obs 9, when it makes the last of the four pairings.

It is thought (but not proven) that this logic will produce a list of pairs with a small sum of distances (here=10.24) but risks not finding matches for every test physician.

BEST MATCH LOGIC #2: Create a filtered Cartesian product and sort by number of matches and then distance

Figure 9 is the data in Figure 8, but sorted differently.

It can be seen that T001 and T002 are competing for the same set of control physicians. It can be seen that T004 competes with T003 for controls and T003 has more pairing options (3 potential controls) than does T004 (2 potential controls).

Test physician 004 has only two potential controls in this data set. In some data sets, tests might have only one potential control. It is thought that sorting by distance alone (as in Figure 8) might mean, in some situations, that physicians with few potential matches are not matched at all. If you sort on distance alone, that could occur when physicians that started with few potential controls go unmatched because all their potential controls have been previously paired.

A different best match logic might be to create a new variable – number of possible matches – and sort by possible

matches and then by distance. This sorting of the data set is shown in Figure 10.

The data in Figure 10 is been sorted by descending possible matches and then by ascending distance. The same selection process as was used before is applied here. The algorithm will start at the top of the data set and select the pairs where it had not previously assigned either physician. Results are presented in Figure 11.

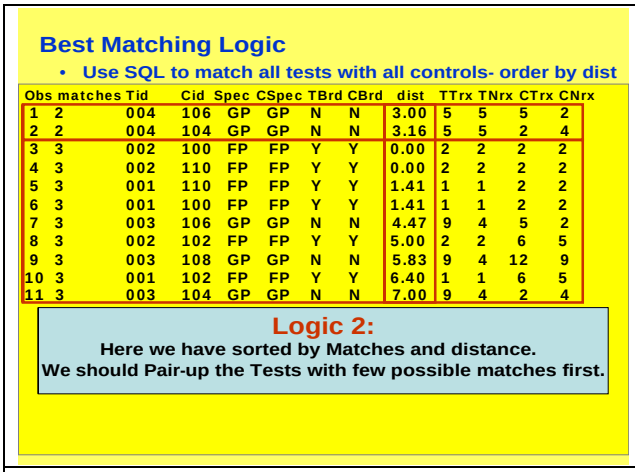

FIGURE 10

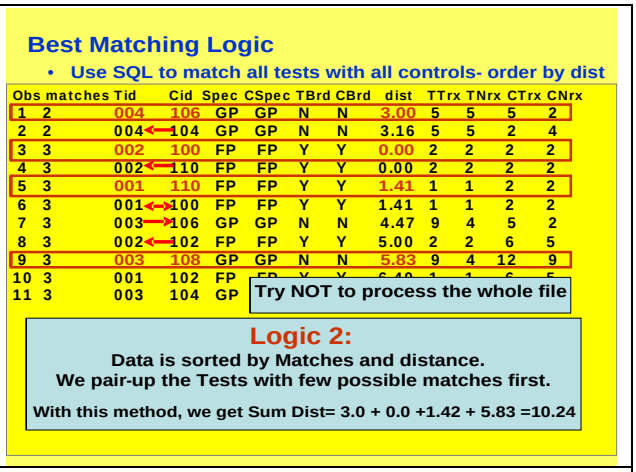

FIGURE 11

It is thought that this logic will result in more matches than Best Match Logic #1 but might also produce paired observations with larger summed distances than that method. Surprisingly, in this example, both logics had summed distances of 10.24. However, there is no assurance that the two methods will return the same sum of distances if fed different data.

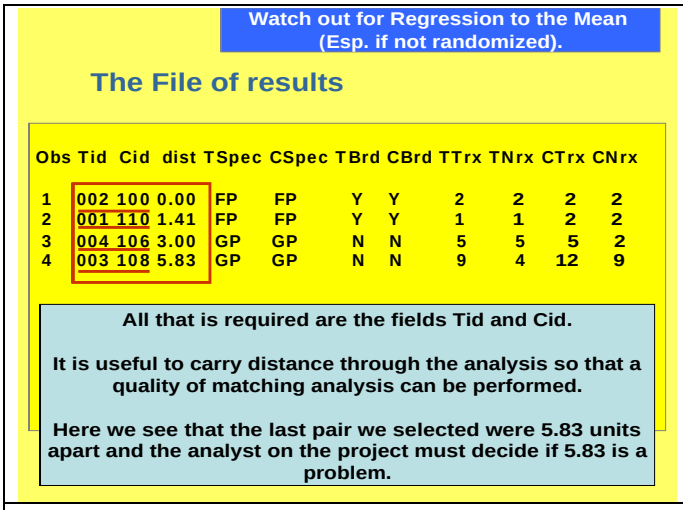

FIGURE 12

BEST MATCH LOGIC #3: Minimize penalty
The rational for this best match logic can be seen in the first four lines of Figure 10.

If physician T004 were not allowed his best pairing (distance =3.0), and had to be matched with his second choice (distance=3.16) he would suffer a penalty of .16 (.16=3.16-3.0).

If T002 were not matched with his first choice (distance=0.0) and had to settle for his second choice (distance=0.0), he would suffer a penalty of zero (0 = 0-0). Physician T001 would also suffer no penalty from being forced to take her second best pairing. A logic might be created that would put tests who will suffer large penalties first in the selection process.

This could lead to the creation of a penalty variable and a best match logic that would sort by descending penalty and then ascending distance. The selection logic would still be to select pairs where neither physician has been previously assigned

OTHER BEST MATCH LOGICS: Combine the above
Combining the penalty variable with the number of possible matches variable and distances, in one sort, could create other best match logics.

STEP 7: MODIFY AND RUN THE PROGRAM

The program (included on the CD) performs an SQL merge to create a data set that contains all possible pairings of test and control. In the merge, it removes pairs from this Cartesian product if the control does not match business criteria, if the pair do not meet the absolute matching criteria or if the pair are too far apart.

The program then sorts the data into an order that puts best matches at the top of the data set. It finishes by selecting the best pairs in a single pass through the data set. The result of this pairing is a file like that shown in Figure 12.

The algorithm implemented uses arrays in a manner similar to key-indexing and intelligently halts needless array processing, making for a fast program that can quickly process even large data sets

CONCLUSION

Matching is a complex multi-disciplinary process. Remember that the threats to validity discussed only briefly here, are critical to a successful business solution.

The SAS program, with sample data sets, is included in an appendix. This program can be used as a learning tool as well as modified to run with your data.

Additional material on threats to validity as presented by Cook & Campbell can be found in ST011_Appendix_B while material offered by Campbell & Stanley is included in ST011_Appendix_C.
The program can be speeded up even more by using array techniques explained in articles written by Paul Dorfman.

REFERENCES

Quasi-Experimentation: Design and Analysis Issues for Field Settings, Cook & Campbell, 1979, Rand McNally

Experimental and Quasi-Experimental Designs for Research, Campbell & Stanley, 1966, Rand McNally
Cluster Analysis for Researchers, H Charles Romesburg, Lifelong Learning Publications, 1984
Dorfman, Paul M, Table Look-Up by Direct Addressing: Key-Indexing , Bitmapping, Hashing. SUGI26, Paper 8

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:
Russell Lavery
Independent Contractor
Email: russell.lavery@verizon.net

```
 *Matching tests and controls;

*Deep thinking goes on before we start coding;
*what is the theory of this experiment;
*Do vars need to be transformed?;
*How should distance be measured?
*how can zeros enter the data (change of zero)?
*What happens when we observe MISSING and ZERO;


*EDA section.  Take a peek at what you were given;
*read data into three files;
*********************************************;
options ls=64;
data phys
     TPhys(rename=(id=Tid Trx=TTrx Nrx=TNrx board=TBrd Spec=tSpec) drop=torc)
     CPhys(rename=(id=Cid Trx=CTrx Nrx=CNrx board=CBrd Spec=CSpec) drop=torc);
infile datalines;
input @1 id $char3.
@5 Board $char1.
@7 Spec $CHAR2. @9 Trx 2.  @11 Nrx 2.  @14 TorC $;
output Phys;
if TorC="t" then output TPhys;
else if TorC="c" then output CPhys;
datalines;
001 Y FP 1 1 t
002 Y FP 2 2 t
003 N GP 9 4 t
004 N GP 5 5 t
100 Y FP 2 2 c
101 N FP 3 1 c
102 Y FP 6 5 c
103 N FP 8 7 c
104 N GP 2 4 c
105 Y GP 2 3 c
106 N GP 5 2 c
107 Y GP 7 2 c
108 N GP12 9 c
109 Y FP 0 0 c
110 Y FP 2 2 c
;

*Do data quality check on data;
*MAYBE print the data files;
Proc print data=Tphys;
Proc Print data=cPhys;
run;

*run freqs and tabs to check data quality;
*find bad data at this step;




rt
*Plot the data files vs all matching vars;
*Look for problems in finding matches;
Proc gchart data=phys;
```

```
Hbar Trx / subgroup=TorC discrete;
run;
Proc gchart data=phys;
Hbar Nrx / subgroup=TorC  discrete;
run;
quit;

*MAYBE (if n is small) plot t and c vs matching variables;
options ps=30;
Proc plot data=phys;
plot Nrx*Trx=TorC;
run;
quit;

*do we need to transform variables?;
*We are now happy with the data quality;


*do the big merge and calculate distance;
Proc SQL;
create table big as

select *,(  (   ((Tnrx-Cnrx)**2) +((TTrx-Ctrx)**2)  )   **.5) as dist

from TPhys as T, Cphys as c
WHERE T.TSpec=C.CSpec
     and T.TBrd=C.CBrd
     and CTrx gt 0 and CNrx gt 0
     and calculated dist le 7
order by dist;

run;
quit;
```

Distance calculated here! Many formulas could have been used.

Absolute matching

Implement conventional wisdom here to keep file small.

Keep file small by eliminating very bad matches. Use info discovered in EDA section to see how large this number should be.

```
*Use a macro to dimension the arrays;
data _null_ ;
if 0 then
   set tphys nobs=nobs;
call symput('lookfor',nobs);
run;
%let lookfor= &lookfor;
```

Nobs is available at compile time from help files. The data set is never read. Faster than SQL with into.

```
*Here is the data step that does the matching;
data bestM ;
set big;
*create the arrays and retain them;
  array TM(&lookfor) ;
  array CM(&lookfor)  ;
  array D(&lookfor)  ;
retain TM1-TM&lookfor CM1-CM&lookfor D1-D&lookfor;

*check to see if the Test physician number has been assigned;
 do i=1 to &lookfor while (TM(i) ne  . );
  if tid=TM(i)then goto escape; /* if match-don't check control*/
 end; /*We.ve looped through all the test IDs- check controls*/

*if we get to here we had not used up the test physician;
*check to see if the Control Physician number has been assigned;
 do i=1 to &lookfor while(CM(i) ne . );
  if cid=CM(i)then goto escape;
 end;  /*We've looped through all control IDs (and test Ids) */

 /*If we get here there was No Match. I points to cell with value=missing*/
 TM(i)=tid;    CM(i)=cid;      D(i) =dist
 output;

*if we loaded the last array cell (selected the last pair) – STOP;
```

Stop when the array is blank.

```
if i EQ &lookfor then STOP;
 goto oout;

*If one or both of the pair was assigned before, delete the obs;
escape:
delete;/* test/control already used - dump that test & control*/

* IF you have a good obs, jump over the delete to label=oout;
oout: /* good matches do not get deleted*/
run;
```

Validity and the threats to Validity

(A.K.A. What went wrong with my experiment)?

From: Quasi-Experimentation: Design and Analysis Issues for field settings, Cook and Campbell 1979 Rand McNally
Experimental and Quasi-Experimental Designs for Research, Campbell & Stanley 1966 Rand McNally

There is a long history of academic experiments going awry. Sometimes, expected were not seen. Sometimes results could not be reproduced. Sometimes results produced in one group of students could not be produced in-groups of slightly different students. People doing educational experiments noticed this and started creating lists of what had gone wrong in experiments. Campbell and Stanley, in 1966, collected the knowledge in the field, and added their own thoughts, to create the classic book on *how experiments go wrong and what can be done to make experiments go "right"*. In 1979, Cook and Campbell revisited the first book and added what had been learned since 1969. These books are the classics in the field of "how to design experiments" and this paper attempts to summarize the important information in the two books. Reading the books themselves is encouraged.

The three authors found that flaws/problems with experiments could be "grouped" -- based on the effect that the flaw/problem produced. A good experiment was named VALID and has four types of validity (External, Internal, Statistical & Construct). The "grouped" *problems* with past experiments were given the name "Threats to Validity". When an experiment is designed, it should be mentally checked to see if the experimental process is "protected" against the different types of threats to validity.

The authors repeatedly mention that there is **no substitution for randomization** as a method of insuring test-control equivalence. Without random assignment of subjects to test and control, there is no true "experiment". The two most important characteristics of a true experiment (an experiment that is going to work, be reproducible and allow us to make statements about causality) are:

1) *Random* assignment of subjects to test and control groups
2) The *experimenter* schedules, administers and controls the giving of the treatment (not some business manager or school administrator).

In this paper please find 1) An explanation of the four types of experimental validity and the need for "manipulation checks"

2) A listing of the commonly recognized "THREATS TO VALIDITY"

3) The 1966 table listing 16 common experimental (and quasi-experimental) designs and a rating of their resistance to the threats to validity

TYPES OF VALIDITY:

Internal Validity: Internal validity addresses the issue: Did the X we changed cause the change in observed Y? THIS IS THE KEY ISSUE FOR AN EXPERIMENT. (Imagine I changed the advertizing and sales increased… Was the increase caused by the advertizing or by the start of flue season … or by a drop in competitor advertising). If you do not have internal validity, you have a VERY poor experiment. Do not do experiments with poor internal validity.

Statistical Validity: Did the test have enough power to detect the expected change/level of shift. Consider NOT DOING an experiment if it has little power.

External Validity: To what populations, settings, treatment variables & measurement variables can this effect be generalized? (This ad copy worked for general practitioners; will it work for specialists? This ad copy increased share; will it increase total number of prescriptions as well? This ad copy worked in cities in Florida, will it work in Canadian cities as well)?

Construct Validity: Did my measurements "get at" the *thing* I was trying to measure. This is a bit philosophical. And I suggest you read use ASK.com to find some examples. Imagine I was looking for a measure of "Police effectiveness" (Police effectiveness is the *thing* I want to measure) and I counted the number of tickets written. This would measure the "activity" of the officer but also the type of job the officers' have- Evidence clerks would not have the opportunity to issue tickets. Arrests would measure, to some extent, the job *and the neighborhood* in which the officer works (officers in high-crime areas would likely have more opportunity to arrest people). If I count tickets I *am* measuring *police effectiveness* but I am also measuring and several other things at the same time.

Additionally problems can be caused by **Ineffective Manipulations.** An experimenter might want to measure the effect of fatigue on memory and might have subjects run a mile. If the group he selects has a few runners (or cyclists) in it, there might be very little fatigue produced. If the experimenter hoped to measure the effect of fatigue on memory he must be able to effectively change the level of fatigue. Experiments on the effects of "job related training" programs can go awry if training is poorly done or

ignored by the students. Subjects have been known to take long lunches, daydream during class or skip out early to play some golf. Experimenters should not expect salespersons to cooperate in an experiment that will, if successful, lower commissions or bonuses. In such a case, the experimental instructions might be subtly ignored. Checking that the change in X did happen is called a "manipulation check".

The general notation for representing an experimental structure is shown in the example below:
For an experiment with: RANDOM ASSIGNMENT…. then an Observation of Y… then a change in X … then a second Observation/Measurement of Y

The notation is: R O X O

R= random Assignment to treatment O=observation of the Y variable X=manipulation, or change, of the X variable
If there is an observed difference in Y, we HOPE it was caused by the change in X. This memo addresses that "HOPE"

The charts below are copied from Campbell and Stanley. The explain "an X represents the exposure of a group to an experimental variable or event, the effects of which are to be measured; O will refer to some process of observation or measurement; The Xs and Os in a given row are applied to the same specific persons. The left to right dimension indicates the temporal order. Xs and Os, vertical to each other, are simultaneous. To make certain important distinctions, as between Designs 2 and 6, or between Designs 4 and 10, a symbol R, indicating random assignment to separate treatment groups is necessary. This randomization is conceived to be a process occurring at a specific time and is the all-purpose procedure for achieving pretreatment equality of groups within known statistical limits. Along with this goes another graphic convention, in that the parallel rows unseparated by dashes represent comparison groups equated by randomization. Those separated by a dashed line represent comparison groups not equated by random assignment. A symbol for matching as a process for the pretreatment equating of comparison groups has not been used, because the value of this process has been greatly oversold and is more often a source of misinformation than a help to valid inference. A specific M for Materials has been used in a specific way in design 9. (Campbell & Stanley pg 6)

THREATS TO VALIDITY:

Threats to INTERNAL VALIDITY (e.g. We manipulated X & saw a change in Y- but our manipulation of X did not cause the change in Y. Something OTHER than our manipulation changed Y. The "OTHER POSSIBLE CAUSES OF THE CHANGE IN Y are the threats to internal validity.)
1) History: *Specific events* occurring between the first and second measurement of Y in addition to the change in the experimental variable. In an evaluation of a grade school reading program, maybe there was a fire alarm during the second measurement and everyone scored low. Maybe, the Olympics happened and the kids did not study.

2) Maturation: Maturation can change Y values. Maturation is a *process within the respondents operating as a function of time per se*… growing older, more tired, bored. If you give a "verbal skill" test when kids start the school year and another at the end of the school year- maturation of the children might have caused any increase in verbal skills

3) Testing: This is the effect of taking a version of the test on the ability to score well on the second test. In the picture above (O X O), the first observation can change the process and cause changes in Y. *In test/re-test, people generally do better on the second test.* Familiarity/practice seems to help in many tests.

4) Instrumentation: *Changes in calibration of an instrument, or changes in observers or scorers* may produce changes in the obtained measure of Y. Judges get bored, tired, annoyed or more expert and this changes the values they record *for the same event.*

5) Statistical Regression: This operates **whenever** groups have been selected on the basis of their extreme scores. If you pick the top five scorers on a pre-test to be controls, their scores will often drop when you give them the post test (they will regress to their average). If you match people on the basis of pretest scores you will usually find that their scores differ by a larger amount on the post-test

6) Biases in differential selection between a test and control group: If all the TEST respondents are Internet savvy and the controls are not, you can see a spurious

difference in Y. Having one group made up of high volume customer and the other regular volume customers can cause a spurious difference in Y.

7) Experimental Mortality: Subjects leaving the study can cause spurious results.

8) Interaction of Selection and Mortality Etc: It often happens that the selection process makes one group more likely to "drop out of the study" than the other group. One of the groups (test/control) might leave the experiment at different rates. In a migraine study of drug and placebo; if the placebo does not reduce pain the subjects may drop out at a higher rate than the subjects who get the drug (and presumably some pain relief). People grouped as "poor computer users" may drop out of an "internet study" at a higher rate than subjects in a "computer savvy" group- or visa versa.

9) Administration contamination: Sometimes an administrator is unwilling to withhold treatment from controls because of the good he/she thinks it might do. More subtly, the interpretation/perception/recording of the results can be influenced if the administrator knows that a subject is test or control. This is one of the reasons for the use of double blinds.

10) Ambiguity about the direction of causality. Does education cause higher salary or does "societal/family environment" cause expectations for success and cause education and cause higher salary.

11) Diffusion of treatments: Sometimes the control group gets the treatment in unplanned ways. If one classroom in a school were shown a gory video about accidents caused by drunk driving they might talk about the video with their friends in the control group.

12) Compensatory Rivalry: Imagine a test is made of the effect of discontinuing an expensive sales support tool. It can happen that a salesperson, whose bonus depends on sales and who lost "access to the tool", might work extra hard to maintain his/her bonus. Compensatory Rivalry occurs most often when participation has some financial impact on subjects and the subjects are in contact with each other. (Sometimes a sales force will anticipate a policy change and not implement a policy).

13) Resentful Demoralization: Respondents who receive the less desirable treatments can become demoralized. Imagine a program to improve reading skills in 5th grade students. Imagine the difference in motivational effect of a promised visit to the class of two different speakers- One: "Allen Iverson of the Sixers" vs. "your favorite market research professional". Resentful Demoralization occurs most often when participation has some financial impact and the subjects are in contact with each other.

Threats to STATISTICAL CONCLUSION VALIDITY: (e.g. Threats to drawing valid inferences about whether two variables co-vary)

LOW Statistical Power
Report-wise vs. Test wise error rates (not adjusting for multiple tests)
Reliability of treatment Implementation
Random Heterogeneity of Respondents
Violated Assumptions of Statistical Tests
Measures that are not reliable (Measures with a large error component)
Random Irrelevancies in the Experimental Setting

Threats to EXTERNAL VALIDITY: (e.g. We saw a change in Y in our test area, but it did not repeat when we went national.)

1) Reactive or interactive effect of testing: The pre-test might increase the sensitivity of the test group to the treatment (Imagine taking a quiz on health care issues and *THEN* seeing a 20 min documentary on AIDS in South Africa. Having the test fresh in your mind can increase/decrease the change the impact of the Video).

2) Interaction of selection biases and the X variable. If one group (test or control) is generally different (healthier, more physically active etc.) than the other group, a spurious difference in Y can result. Imagine testing a program that helps children learn to read. If one group is randomly assigned a high percentage of poor readers and

another group is mostly good readers, it is possible that the "reading program" can help the good readers more than the poor readers.

3) Reactive Elements of the Experimental Arrangements: This addresses the problem where the effect on Y can only be seen in the "highly artificial" experimental setting. Often results that can be produce in an artificial, carefully-controlled, experimental environment can not be produced in the "messy real world". People have been able to develop programs that help addicts give up drugs when administered in a highly controlled environment (Betty Ford Clinic). These programs might not work if administered to people living at home and in the environment that helped make them addicts.

Reactive elements affect compliance compliance with protocols. It is easy to achieve compliance with a complicated AIDS drug regimen if the subject is in a lab environment in a lab- harder for a group of working professional in America to be compliant– harder still for people in third world countries to be compliant. A second example would be where the subjects, knowing they are in an experiment, use the stimuli, (treatment and environmental conditions) to divine the experimenter's intent.

4) Multiple Treatment Interference: If test subjects are used over and over, they can become abnormally responsive, or unresponsive. Results seen in lab settings may not occur in other environments. People who have volunteered for many "paid drug tests" can become abnormally sensitive to chemicals.

5) Mono-operationalism bias: Mono-operationalism means that the effect was measured using ONE methodology, an often dangerous practice. It sometimes happens that an effect is only "seen" if measured with one type of instrument. Verbal (self-report) measures of effect can be unreliable. A famous study found great differences between what people said they would do and what they actually did. A researcher exposed subjects to a stimulus and measured the effect by *asking* the subjects what *they would do in a certain situation.* Another researcher got a very different result from another researcher when he exposed subjects to the same stimulus and *set up that certain situation so as to be able to observe actual behavior.*

Threats to CONSTRUCT VALIDITY: (This refers to the possibility that the operations which are meant to measure a particular cause or effect, can be construed in terms of more than one construct.) It is well known that a Physician's friendly, compassionate bedside manner, speeds recovery. Could friendly, compassionate Dr. X prescribing a pill, and not the chemicals in the pill, be the cause of the patients recoveing. This issue is intimately involved with confounding variables)

16 COMMON EXPERIMENTAL DESIGNS :

From "Experimental & Quasi-Experimental Designs for Research" By Campbell & Stanley1963, Houghton Mifflin Company

"Note: In the table, a "bad" indicates a definite weakness, a "good" indicates that the factor is controlled, a question mark indicates a possible source of concern. A "N/A" indicates that the factor is not relevant. The authors say; "It is with extreme reluctance that these summary tables are presented because they are apt to be "too helpful" and to be depended upon in place of the more complex and qualified presentation in the text. No Good or Bad indicator should be respected unless the reader comprehends why it is placed there. In particular it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs." (Campbell & Stanley pg. 8)

Not Recommended

	Internal Validity Threats								External Validity Threats			
Type of Design	History or special events other than X	Testing changes the measurement	Maturation of the subjects	Instrumentation effects	Regression to the mean	Selection Bias	Mortality (differential)	Interaction Selection & Maturation Etc	Interaction of testing & X	Interaction of Selection & X	Reactive Arrangements	Multiple X Interference
1) One Shot X 0	BAD	BAD	N/A	N/A	N/A	BAD	BAD	N/A	N/A	BAD	N/A	N/A
2) One Group Pre-post 0 X 0	BAD	BAD	BAD	BAD	?	GOOD	GOOD	BAD	BAD	BAD	?	N/A
3) Static Group 0 X 0 --- 0	GOOD	?	GOOD	GOOD	GOOD	BAD	BAD	BAD	N/A	BAD	N/A	N/A

Experimental Designs

	Internal Validity Threats								External Validity Threats			
4) Pre Post W/ Control R 0 X 0 R 0 0	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	BAD	?	?	N/A
5) Solomon R 0 X 0 R 0 0 R X 0 R 0	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	?	?	N/A
6) POSTTEST Only W/ Control **R** X O **R** O	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	?	?	N/A

Quasi- Experimental Designs

	Internal Validity Threats								External Validity Threats			
Type of Design	History or special events other than X	Testing changes the measurement	Maturation of the subjects	Instrumentation effects	Regression to the mean	Selection Bias	Mortality (differential)	Interaction Selection & Maturation Etc	Interaction of testing & X	Interaction of Selection & X	Reactive Arrangements	Multiple X Interference
7) Time series O O X O O	BAD	GOOD	GOOD	?	GOOD	GOOD	GOOD	GOOD	BAD	?	?	N/A
8) Equiv. Time Samples Design X1-O X0-O X1-O X0-O X1-O X0-O	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	BAD	?	BAD	N/A
9 Equiv. Materials MaX1O MbX0 O McX1 O MdX0 O	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	BAD	?	?	N/A
10) NON-EQUIVOLENT CONTROL GROUP O X O --- O X O	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	BAD	?	?	N/A
11) Counter Balanced designs X10 X2 0 X3 0 X4 0 X2 0 X4 0 X1 0 X3 0 X3 0 X1 0 X4 0 X2 0 X4 0 X2 0 X2 0 X1 0	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	?	?	?	?	N/A
12) Separate Sample Pre & Post R O (X) O R X O	BAD	BAD	GOOD	?	GOOD	GOOD	BAD	BAD	GOOD	GOOD	GOOD	N/A
12a) R O (X) O R X O --- R O (X) R X O	GOOD	BAD	GOOD	?	GOOD	GOOD	BAD	GOOD	GOOD	GOOD	GOOD	N/A
12b) R 01 (X) R 02 (X) R X 03	Bad	GOOD	GOOD	?	GOOD	GOOD	BAD	?	GOOD	GOOD	GOOD	N/A
12c) R 01 (X) 02 R (X) 03	BAD	BAD	GOOD	?	GOOD	GOOD	GOOD	BAD	GOOD	GOOD	GOOD	N/A
13) R O (X)	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	BAD	GOOD	GOOD	GOOD	N/A

<table>
<tr><td>R X O

R O
R O</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Type of Design</td><td>History or special events other than X</td><td>Testing changes the measurement</td><td>Maturation of the subjects</td><td>Instrumentati on effects</td><td>Regression to the mean</td><td>Selection Bias</td><td>Mortality (differential)</td><td>Interaction Selection & Maturation</td><td>Interaction of testing & X</td><td>Interaction of Selection & X</td><td>Reactive Arrangements</td><td>Multiple X Interference</td></tr>
<tr><td>13a)
|R O (X)
|R X O
R|-----------
|R O (X)
|R X O
|-----------
|R O (X)
|R X O

|R O
|R O
R|-----------
|R O
|R O
|-----------
|R O
|R O</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>N/A</td></tr>
<tr><td>14 Mult. time series
O O O X O O O

O O O O O O</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>GOOD</td><td>BAD</td><td>BAD</td><td>?</td><td>N/A</td></tr>
<tr><td>15 Institutional Cycle Design
Class A:
X O1

Class B1
R O2 X O3
Class B2
R X O4

Class C
R O5 X

Genl Pop Con. Cl.
B O6

Genl Pop Con. Cl.
C O7

O2 < O1
O5 < O4

O2 < O3
O2 < O4</td><td>GOOD
GOOD

BAD
BAD</td><td>BAD
BAD

BAD
BAD</td><td>GOOD
GOOD

BAD
GOOD</td><td>GOOD
GOOD

?
?</td><td>?
?

?
?</td><td>BAD
BAD

GOOD
GOOD</td><td>?
?

GOOD
?</td><td>N/A
N/A

N/A
N/A</td><td>GOOD
GOOD

BAD
GOOD</td><td>?
?

?
?</td><td>GOOD
GOOD

GOOD
?</td><td>N/A
N/A

N/A
N/A</td></tr>
</table>

O2 < O3 O2 < O4	N/A N/A	GOOD GOOD	N/A N/A	N/A N/A	N/A N/A	N/A N/A	N/A NA/	BAD BAD	N/A N/A	N/A N/A	N/A N/A	N/A N/A
16 Regression Discontinuity	Good	Good	Good	?	Good	Good	?	GOOD	GOOD	BAD	GOOD	GOOD