

## HOW TO GO ABOUT DATA GENERATION AND ANALYSIS (Lecture 11, 2009\_03\_08)

### Please recall my remarks in Lecture 1 about the nature of the lecture notes.

### Outlines of „lit reviews” ?

### **Steve Levitt’s (J.B.Clark medal 2003) Rules of thumb for successful empirical research**

#### **1) A paper must ask a good question**

- one that has never been asked
- you do not know the answer in advance
- no matter what answer you get it will be interesting
- others were getting a wrong answer so far

#### **2) A paper should have an “idea”**

- a clever new way of answering an old question
- a new source of identification
- uncovering a relationship nobody has thought of so far
- a new econometric method

#### **3) The simpler the execution the better**

- present data and results in as raw a form as possible
- build-in complexity later
- avoid too many assumptions when constructing more complicated estimators
- people should see where your results are coming from

#### **4) Be certain you have the right answer**

- check robustness to death
- think through all the implications of your model for the results

#### **5) Interpret your results**

- throwing regression coefficients into a table is not enough
- what’s the economic significance of your results?
- do some cost-benefit analysis, implications for policy, etc..

#### **6) Become an expert**

- learn as much as you can about the institutional background
- some first-hand, insider experience will tell you far more than the books

#### **7) When you should, fail quickly and leave the sunk costs**

- if there’s nothing in the data, dump it
- if the data goes the wrong way at the first glimpse, drop the project

#### **8) Practice makes perfect**

### ### Review

ON SUBJECT POOLS AND LEVELS OF REASONING (Lecture 10, 2009\_03\_08)

### Bosch-Domenech et al., One. Two. (three), infinity, ... : newspaper and lab beauty-contest experiments (AER 2002)

BCG (e.g., Nagel AER 1995): Players submit a number between 0 and 100. The winner is the person whose number is closest to  $p$  times the average of all submitted numbers, where  $0 < p < 1$ , and here  $2/3$ . Winners split prize.

#### Design and Implementation

- Historically, lots of lab experiments ... (Nagel 1995 and many others including new ones reported in this paper)
- Here also reported newspaper (artefactual) experiments with readers of Financial Times, Spektrum der Wissenschaft, and Expansion
- (p.p. 1693 – 7)

As mentioned, one purpose of running experiments out of the lab is to help critically assess the assumption of “parallelism.” Do we see, then, similarities or differences between Beauty-contest experiments run in labs and in newspapers?

Before entering into a detailed comparison, it is worth mentioning some of the basic differences between the two types of experiments, often due to the increased loss of control in newspaper experiments:

- (a) subjects' socio-demographic characteristics
- (b) information acquisition
- (c) coalition formation

Results (see also the various facts extracted in the lecture notes from the article):

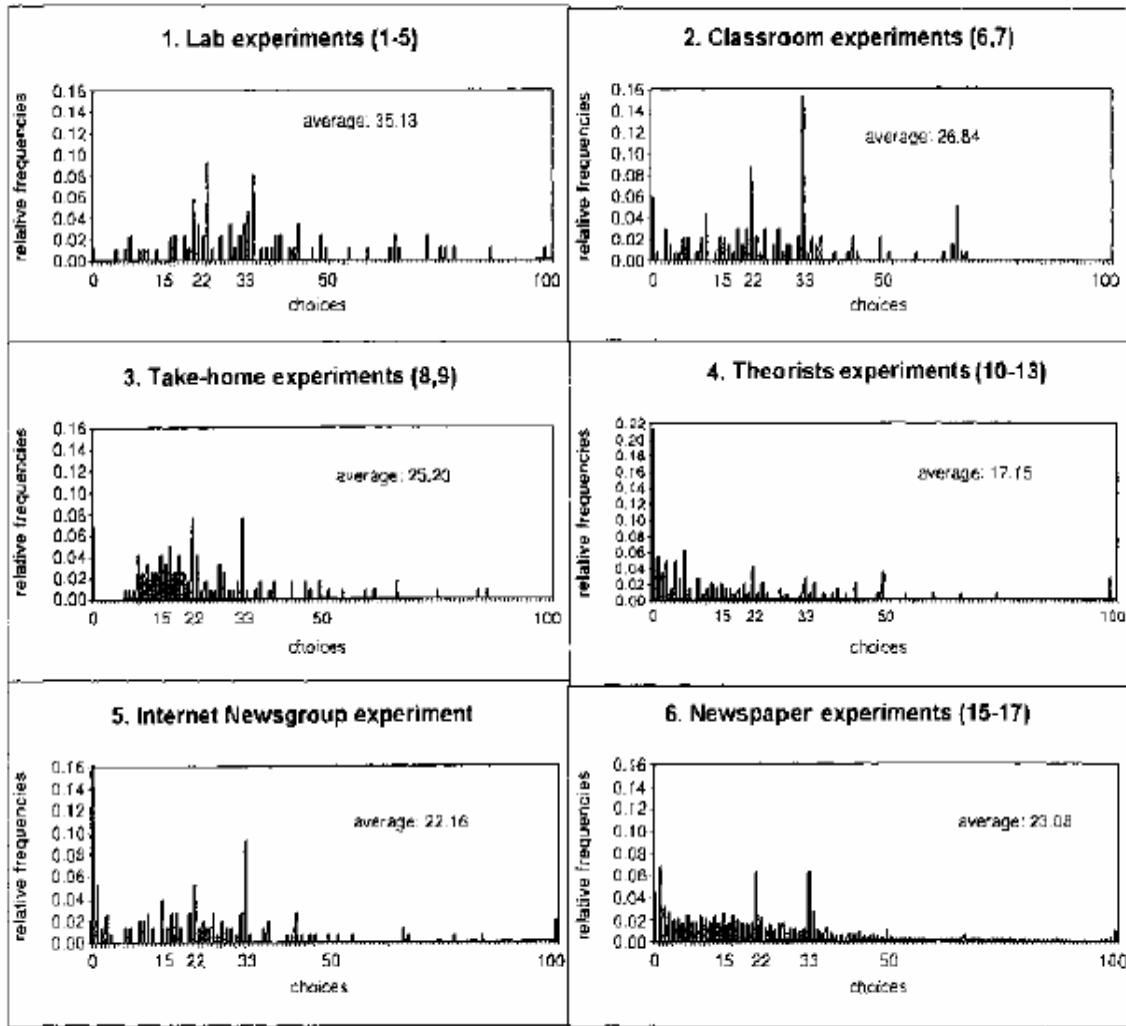


FIGURE 3. RELATIVE FREQUENCIES OF CHOICES IN THE SIX GROUPS OF EXPERIMENTS

### Summary

- Yes, subject pools do matter but ... (qualitatively the same picture)
- People of various walks of life do not engage in many steps of reasoning
- "Experimenters" do rather well (in the sense that they figure out what others think), at least in this experiment

### Johnson et al., Detecting failures of backward induction: Monitoring information search in sequential bargaining (JET 2002)

## Motivation

To tease apart why people do not manage to implement the subgame-perfect solution in alternating offer games [explain specifically how used in the present paper], or in simple bargaining games for that matter:

- Is it due to social preferences?
- Is it due to cognitive limitations?

## Design and implementation

Use MouseLab and undergraduate students to find an answer.

Three experimental studies:

- Bargaining with other players
- Bargaining with robots and instructions
  - turning off “social preferences”
  - also, teaching subjects
- Mixing trained and untrained subjects

Each subject plays 8 (16, 16) three-round alternating offer games, rematched each round in group of 10 with another member of the group.

Payment in cash at the end according to performance (half of dollar earnings and show-up fee)

Experimenter tracks choices but (importantly) also information acquisition (and therefore, via inference, information processing): thus, a non-invasive way to look into the head of subjects! (different from various other techniques such as fMRI etc.)

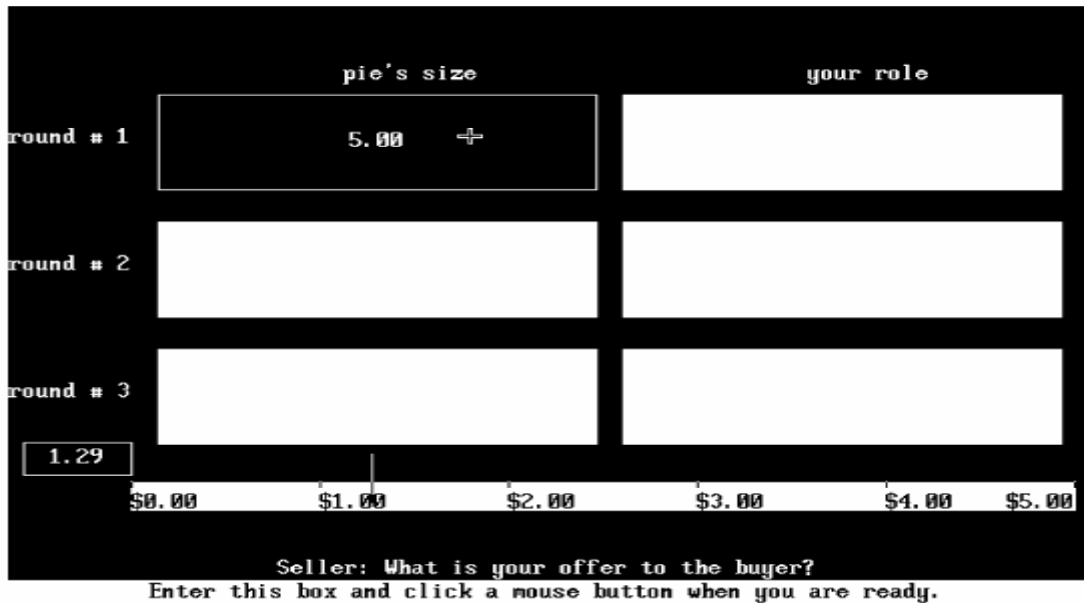


FIG. 1. MOUSELAB computer display.

Important (questionable) assumption – subjects do not use memory

### Summary

- Social preferences (accounting for about one third) and limited cognition (accounting for about two thirds) both play their role. (What role they play depends on the same factors as those in Cherry et al., Andreoni & Miller, and List (JPE 2006))
- Subjects can be taught to think strategically very quickly ...
- Equilibrium choices are highly correlated with equilibrium reasoning
- MouseLab is a really cool (and very underused) tool !

### Rubinstein, Instinctive and Cognitive Reasoning: A study of response times (EJ 2007)

### Motivation

Rather than fMRI [to be discussed later in this course] or similar expensive, small-sample (and hence noisy) studies, Rubinstein wants “to explore the deliberation process of decision makers based on their response times.” (p. 1244)

### Definition

“Response time [RT] is defined here as the number of seconds between the moment that our server receives the request for a problem until the moment that an answer is returned to the server.” (p. 1245, see also p. 1257, 5.(a) )

- no control for differential server speed, or transfer speeds
- no control for differential speed of reading etc.

“The magic of a large sample gives us a clear picture of the relative time responses.” (p. 22; indeed statistical tests highly significant – no surprise given the numbers involved.; see p. 1257, 5.(b))

A standard criticism of survey experiments is that in the absence of monetary rewards behaviour is less realistic. However, in my experience there is no significant difference between survey results and results in experiments with monetary rewards; see also Camerer and Hogarth (1999). In any case, we are not interested here in the absolute distribution of responses in real life problems (and note that even with real payments the experiment is still far from a real life situation), but only in the relative response times of the different choices. Thus, the absence of real rewards should not have any significant impact.

(p. 1257) -> worksheet [recall earlier lecture on financial and social incentives]

## Basic working hypothesis

Action that require lesser response time are more instinctive (i.e., on the basis of an emotional response); those requiring more response time are more cognitive. [We'll return to this issue also in L 13, 14 when we discuss issues in neuroeconomics.]

## Basic methodology

Classify “intuitively” (“I have done so intuitively.” p. 1245, see also p. 1258, 5.(c)):

As mentioned earlier, the classification of choices was done intuitively. An alternative and more formal approach would be to base classification on other sources of information such as the results of a survey in which subjects were asked whether they consider a choice to be instinctive or not. Of course, such an approach would have its own deficiencies. In any case, the distinction between intuitive and cognitive responses was used here only as a suggestive explanation for the huge differences in time response between actions.

See related questions on worksheet.

## Example 4 (The Beauty Contest Game)

Each of the students in your class must choose an integer between 0 and 100 in order to guess '2/3 of the average of the responses given by all students in the class'.

Each student who guesses 2/3 of the average of all responses rounded up to the nearest integer, will receive a prize to be announced by your teacher (or alternatively will have the satisfaction of being right!).

What is your guess?

Nash equilibrium?

Results?

Table 4  
Example 4: Results

	$n = 2,423$	0-1	2-13	14-15	16-21	22	23-32	33-34	35-49	50	51-100
	86 sec	11%	9%	2%	6%	4%	10%	11%	11%	16%	20%
A	15%	269	213	47	137	99	249	262	267	393	487
B	49%	126 sec	91 sec	89 sec	84 sec	82 sec	157 sec	84 sec	113 sec	94 sec	70 sec
C	36%	89 sec	70 sec								70 sec

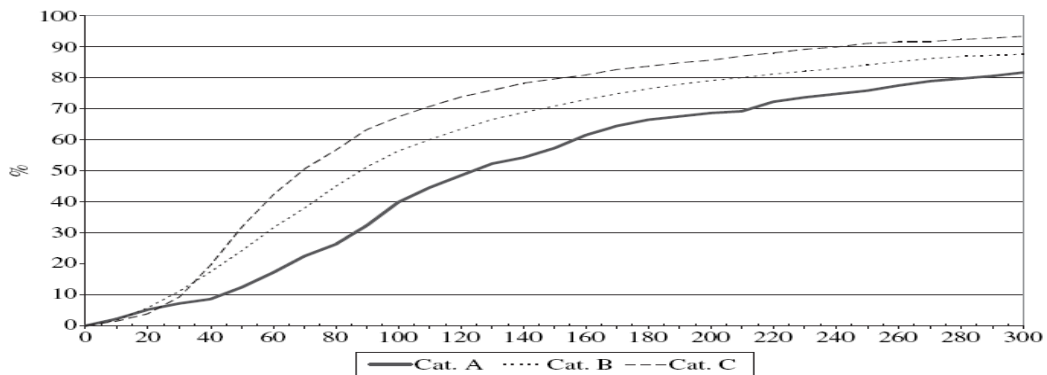


Fig. 4. Example 4: Response Time Frequencies

Where

A = responses of 33 – 34 and 22

C = responses of 50 or more

B = responses of “victims of Game Theory” and “the subjects whose strategy was to give the best response to a wild guess.” (p. 12)

Is Nagel’s classification wanting, as Rubinstein suggests? (p. 13)

The results cast doubt on the classification used by Nagel and others whereby the whole range of 20–25 is classified as one group. In my data, the MRT of the 4% who chose 22 was 157 seconds while the MRT among the 8% who chose 20, 21, 23, 24 or 25 was only 80 seconds. This must mean that there is little in common between the choice of 22 and the rest of the category which Nagel called ‘Step 2’.



Example 5 (The Centipede Game; recall our earlier look at Parco et al. 2002, and your reading of Palacios-Huerta & Volij 2006, tbd later today)

You are playing the following ‘game’ with an anonymous person. Each of the players has ‘an account’ with an initial balance of \$0. At each stage, one of the players (in alternating order – you start) has the right to stop the game.

If it is your turn to stop the game and you choose not to, your account is debited by \$1 and your opponent’s is credited by \$3.

Each time your opponent has the opportunity to stop the game and chooses not to, your account is credited by \$3 and his is debited by \$1.

If both players choose not to stop the game for 100 turns, the game ends and each player receives the balance in his account (which is \$200; check this in order to verify that you understand the game).

At which turn (between 1 and 100) do you plan to stop the game? (If you plan not to stop the game at any point write 101).

What’s the (subgame perfect) Nash equilibrium?

Results?

Says Rubinstein (p. 1252):

The Centipede Game is another prime example of the tension between Nash equilibrium and the way in which games are actually played. Assuming that the players care only about the amount in their own account, the only Nash equilibrium strategy for player 1 is to stop the game at turn 1. However, this is a highly unintuitive action. The response 101 seems to be the instinctive one. The cognitive actions are in the upper range of the responses (98, 99, 100). A choice in the range 2–97 seems to be a reasonless one.

Table 5  
Example 5: Results

<i>n</i> = 1,361	1	2–97	98–100	101
%	12	11	20	57
median	132 sec	80 sec	163 sec	123 sec

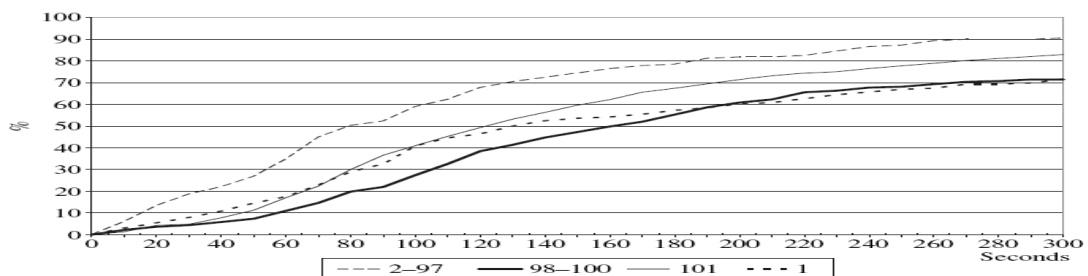


Fig. 5. Example 5: Response Time Frequencies

Again, the question is whether financial incentives make a difference. Recall J.E. Parco, A. Rapoport & W.E. Stein. "Effects of financial incentives on the breakdown of mutual trust," Psychological Science 2002 (13) 292-297 (which is available on the internet; just type into google "parco rapoport psych science").

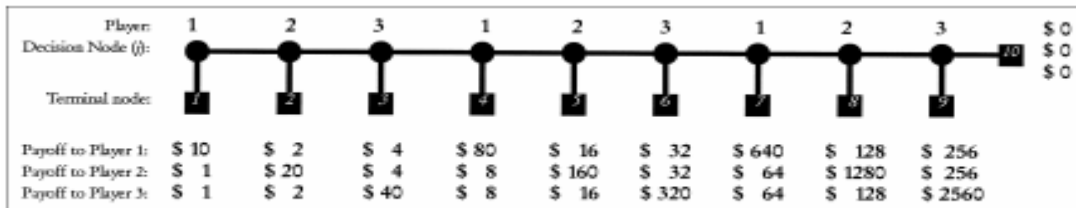


Fig. 2. The three-person, nine-move centipede game used in Rapoport, Stein, Parco, and Nicholas (2000).

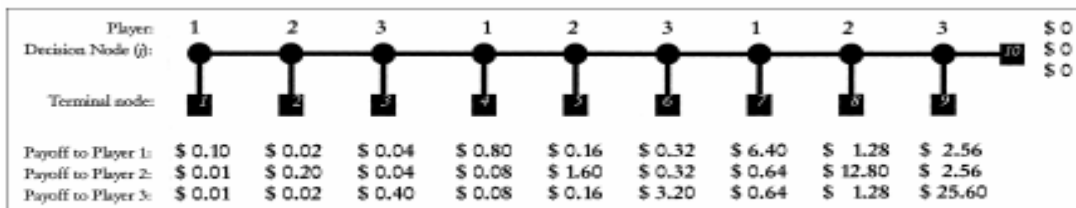


Fig. 3. The three-person, nine-move centipede game used in the present study.

**Table 1. Proportion of games ending at each terminal node**

Session	N <sup>a</sup>	Terminal node								
		1	2	3	4	5	6	7	8	9
<b>Low-pay 9-move game (present study)</b>										
1	300	.027	.043	.093	.240	.263	.227	.073	.017	.013 <sup>b</sup>
2	300	.023	.067	.250	.243	.263	.097	.037	.007	.013
Across sessions	600	.025	.055	.172	.242	.263	.162	.055	.012	.013
<b>High-pay 9-move game (RSPN)</b>										
1	300	.463	.317	.110	.050	.027	.020	.010	.003	.000
2	300	.393	.277	.157	.087	.090	.017	.023	.013	.003
3	300	.303	.280	.187	.093	.053	.037	.010	.003	.003
4	300	.407	.257	.183	.077	.037	.017	.013	.003	.007
Across sessions	1,200	.392	.283	.159	.077	.037	.023	.014	.006	.010

<sup>a</sup>Number of games (five groups of 3 randomly matched players per trial participating in 60 trials).  
<sup>b</sup>A single Player 3 continued at the 9th decision node.  
<sup>c</sup>RSPN = Rapoport, Stein, Parco, and Nicholas (2000).

Recall Parco et al.'s conclusion.

Another concern is the rather curious framing of the task (which probably interacts with the lack of financial incentives.)

### Rydval, Ortman, Ostatnicky (2008), Three Simple Games and How to Solve Them. (Manuscript)

**Figure 1: The guessing games in normal-form representation**

Game 2p2n: 2 players, 2 numbers

		Player 2	
		0	1
Player 1	0	M/2,M/2	M,0
	1	0,M	M/2,M/2

Game 2p3n: 2 players, 3 numbers

		Player 2		
		0	1	2
Player 1	0	M/2,M/2	M,0	M,0
	1	0,M	M/2,M/2	M,0
	2	0,M	0,M	M/2,M/2

Game 3p2n: 3 players, 2 numbers

Player 3's choice = 0

		Player 2	
		0	1
Player 1	0	M/3,M/3,M/3	M/2,0,M/2
	1	0,M/2,M/2	0,0,M

Player 3's choice = 1

		Player 2	
		0	1
Player 1	0	M/2,M/2,0	M,0,0
	1	0,M,0	M/3,M/3,M/3

### Reasoning class A

Wrong reasoning – e.g., due to misrepresenting the strategic nature of the guessing game or making a numerical mistake, or irrelevant belief-based reasoning.

### Reasoning class B

Reasoning based on listing contingencies involving own dominant choice of 0, but without explicitly explaining why 0 is the dominant choice.

### Reasoning class C

Reasoning explicitly recognizing and explaining why 0 is the dominant choice, with or without listing contingencies.

**Table 2: Frequency of subjects sorted by reasoning classes, choices and beliefs**

	Total	Class A	Class A/B	Class A/C	Class B	Class B/C	Class C
All subjects	<b>112</b>	<b>66</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>7</b>	<b>30</b>
Choice=0	<b>62</b>	17	2	3	3	7	30
Choice=1	<b>50</b>	49	1	0	0	0	0
Belief=0	<b>49</b>	12	1	2	3	4	27
Belief=1	<b>63</b>	54	2	1	0	3	3
Choice=0 & Belief=0	<b>46</b>	9	1	2	3	4	27
Choice=0 & Belief=1	<b>16</b>	8	1	1	0	3	3
Choice=1 & Belief=0	<b>3</b>	3	0	0	0	0	0
Choice=1 & Belief=1	<b>47</b>	46	1	0	0	0	0

**Table 3: Percentages (rounded to integers) of reasoning classes and choices for each game**

	Total	Class A	Class A/B	Class A/C	Class B	Class B/C	Class C
Game 2p2n (28 subj.)	<b>100</b>	<b>57</b>	<b>4</b>	<b>0</b>	<b>4</b>	<b>11</b>	<b>25</b>
Choice=0	57	14	4	0	4	11	25
Choice=1	43	43	0	0	0	0	0
Game 2p3n (41 subj.)	<b>100</b>	<b>76</b>	<b>2</b>	<b>5</b>	<b>2</b>	<b>2</b>	<b>12</b>
Choice=0	39	17	0	5	2	2	12
Choice=1	61	59	2	0	0	0	0
Game 3p2n (43 subj.)	<b>100</b>	<b>44</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>7</b>	<b>42</b>
Choice=0	70	14	2	2	2	7	42
Choice=1	30	30	0	0	0	0	0

### Kovalchik, Camerer, Grether, Plott, Allman, Aging and decision making: a comparison between neurologically healthy elderly and young individuals (JEBO 2005)

How many experiments on how many populations? Specifically what are the populations?  
 4, 2, neurologically healthy elderly (ave age 82, N = 50, 70 % female) and young individuals (probably from PCC, N = 51, 51 % female)

What are the tasks used in those experiments?

- Confidence (exploring meta-knowledge, see Hertwig & Ortmann book chapter, lecture 9)
- decisions under uncertainty
- differences between WTP – WTA (as in Plott Zeiler)
- strategic thinking (as in Bosch-Domenech et al)

What exactly was the confidence task? (Make sure you read Appendix A; do you see any problem with the questions?)

- 20 trivia questions (general knowledge questions? These questions seem to reflect the age of the experimenters! They may be trivia questions but I doubt whether they are legitimate general knowledge questions!)
- all questions two possible answers
- subjects had to try to give answer and provide a confidence assessment of their answer

What was the result of the confidence task? (Understand Figure 1)

- older 74.1 correct, younger 66.1 correct
- calibration? (some overconfidence – see also p. 83 lines 2 – 5)

82

S. Kovalchik et al. / J. of Economic Behavior & Org. 58 (2005) 79–94

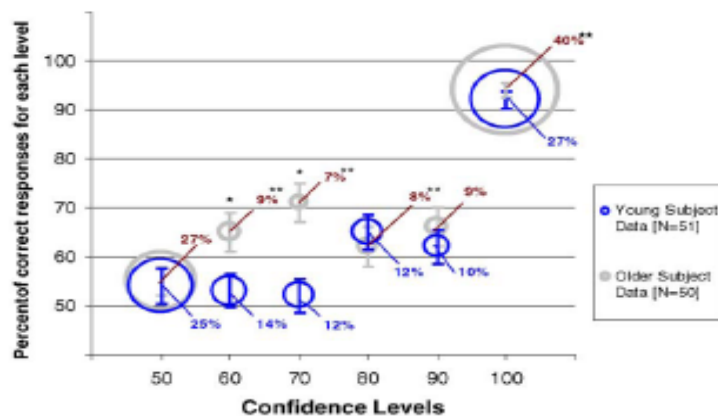


Fig. 1. Proportion of correct responses ( $\pm$ standard error of mean (S.E.M.)) for a given confidence level out of the total responses with that confidence. The width of each bubble reflects the percentage of responses that were given at each confidence level out of the total responses. Exact percentages for response distribution are labeled next to bubbles for each population. \*The hypothesis that the proportions correct are the same is rejected at  $P < 0.05$ . \*\*The hypothesis that the proportions choosing this response are the same is rejected at  $P < 0.05$ .

Do you agree with the authors' interpretation of their results? ("One interpretation of these results is that older subjects have learned through experience to temper their overconfidence and, thus, look more like experts." (p. 82)) Can you think of another explanation?

What exactly was the WTA – WTP task? And how was it implemented?

(How was it different from PZ 2005?)

- subjects interviewed one at a time and performed either as buyer or seller
- each round were told that they own the item in front of them
- then asked to report value of item them
- then asked to state WTA (sellers) and WTP (buyers)
- then BDM procedure in simplified form
- total of three rounds (the first two – pen and frame -- hypothetical, the third – coffee mug -- real)
- anonymous transfer

What was the result of that task?

Table 2  
Statistics for WTA/WTP [mean, median (S.D.)]

	Round		
	Hypothetical 1 (pen)	Hypothetical 2 (frame)	Actual (mug)
Older WTA (N=25)	8.84, 3.5 (17.82)	6.61, 7 (3.74)	2.48, 2.5 (1.7)
Older WTP (N=25)	5.13, 3.95 (6.12)	9.34, 6.5 (6.21)	3.25, 2 (3.04)
Younger WTA (N=26)	2.2, 1.5 (3.2)	7.46, 5 (8.63)	3.88, 2.38 (4.88)
Younger WTP (N=25)	1.62, 1.25 (1.12)	4.98, 4 (3.59)	2.24, 2 (1.75)

The data shows the mean, median (standard deviation) for the WTAs (seller offers) and WTPs (buyer offers). Data for each of the three rounds and the items used are given. Only the actual round used real cash. For the actual round the difference between the WTA and WTP prices within each group is not significant ( $P > 0.25$ ).

What exactly was the strategic thinking task?  
 (How was it different from the task in Bosch-Domenech? Or was it?)  
 What was the result of that task?

**Stem and Leaf Plots for the Beauty Contest Game**

	Younger Subjects	Older Subjects
0	47	
1	5679	4788
2	000023478889	12255677779
3	0223333345555556677788	2355557
4	222235	457
5	0002	0028
6	58	25
7		5
8		6

Fig. 5. The plot shows the total number of subjects and number selections for the *p*-beauty contest (theory of mind task). Players try to choose a number close to  $2/3$  times the average of 10 numbers provided by participating subjects. The values in the left column are the "stems" or the tens digit, and the values on the right list the digits. Each numerical response is therefore 10 times the left column stem plus the middle or right column "leaf" value. For example, the most extreme guesses were 4, given by a younger subject, and an 86 given by an older subject. Values greater than 50 may indicate confused responses. Older subjects were more likely to give a response above 50 ( $P < 0.07$ ).

p. 88: "Our results show that both the old and the young samples behave similarly on this task."

A SHORT LIST OF PRACTICAL POINTS ON HOW TO GO ABOUT DATA GENERATION AND ANALYSIS (see p. 64 of Friedman & Cassar whose chapter 5 I will send you if you send me message):

This is a long chapter, so a short list of practical points might be helpful:

- 1 Before beginning your experiment, think about how you will analyze the data.
- 2 Use good design to avoid collinearity, omitted variables, etc.
- 3 Choose your laboratory protocols to reduce measurement errors.
- 4 Choose your treatments to produce good, representative samples.
- 5 Choose a design that allows you to use efficient statistics (e.g. matched-pairs).
- 6 Search the literature and your imagination to find effective graphical displays and summary statistics.
- 7 Look for irregularities and outliers in your data.
- 8 Use standard parametric and nonparametric hypothesis tests to draw your conclusions.

All true!

Also true (well, at least in my view):

You have a problem if you have to torture the data to make them confess!

Ideally, you should be able to tell a story (your story!) by way of descriptive data (graphs, summary statistics). Tufte (1983) is a good book indeed. But even simpler, **take note of what you like in articles you read.**

And, yes, it is very important to go through the „qualitative phase“ and really understand the data. Such an analysis will help you spot unexpected (irr)regularities (recording mistakes, computer malfunctions, human choice idiosyncracies, etc.) [Because of the ease with which programs like STATA can be used nowadays, some experimenters are tempted to skip this step. Bad idea!]

Importantly, IT IS WORTH BEARING IN MIND THAT STATISTICAL TESTS ARE UNABLE TO DETECT FLAWS IN EXPERIMENTAL DESIGN AND IMPLEMENTATION ... (Ondrej's write-up p. 7)



A little excursion of a more general nature:

The following text has been copied and edited from a wonderful homepage that I recommend to you enthusiastically: <http://www.socialresearchmethods.net/>

It's an excellent resource for social science research methods.

Four interrelated components that influence the conclusions you might reach from a statistical test in a research project [Sadly, it has to be said that most experimentalists do not think much about the interrelatedness of these components, or any of these components other than the alpha level; but the prevalence of a bad practice does not mean, that one ought, or has to, adopt it]:

The four components are:

- **sample size**, or the number of units (e.g., people) accessible to the study
- **effect size**, or the salience of the treatment relative to the noise in measurement
- **alpha level** ( $\alpha$ , or significance level), or the odds that the observed result is due to chance
- **power**, or the odds that you will observe a treatment effect when it occurs

Given values for any three of these components, it is possible to compute the value of the fourth. For instance, you might want to determine what a reasonable sample size would be for a study. If you could make reasonable estimates of the effect size, alpha level and power, it would be simple to compute (or, more likely, look up in a table) the sample size.

Some of these components will be more manipulable than others depending on the circumstances of the project. For example, if the project is an evaluation of an educational program, the sample size is predetermined. Or, if the drug dosage in a program has to be small due to its potential negative side effects, the effect size may consequently be small. The goal is to achieve a balance of the four components that allows the maximum level of power to detect an effect if one exists, given programmatic, logistical or financial constraints on the other components. Of course, financial constraints are always a concern to experimental economists.

Figure 1 shows the basic decision matrix involved in a statistical conclusion. All statistical conclusions involve constructing [two mutually exclusive hypotheses](#), termed the null (labeled  $H_0$ ) and alternative (labeled  $H_1$ ) hypothesis. Together, the hypotheses describe all possible outcomes with respect to the inference. The central decision involves determining which hypothesis to accept and which to reject.

For instance, in the typical case, the null hypothesis might be:

**H<sub>0</sub>: Program Effect = 0**

while the alternative might be

**H<sub>1</sub>: Program Effect  $\neq$  0**

The null hypothesis is so termed because it usually refers to the "no difference" or "no effect" case. (E.g., you might want to test whether asset legitimacy makes a difference, or asset legitimacy and anonymity make a difference but you claim in your null hypothesis they don't.) Usually in social research we expect that our treatments and programs will make a difference. So, typically, our theory is described in the alternative hypothesis.

Figure 1 below is a complex figure that you should take some time studying.

First, look at the header row (the shaded area). This row depicts reality -- whether there really is a program effect, difference, or gain. Of course, the problem is that you never know for sure what is really happening (unless you're God). Nevertheless, because we have set up mutually exclusive hypotheses, one must be right and one must be wrong. Therefore, consider this the view from God's position, knowing which hypothesis is correct. The first column of the 2x2 table shows the case where our program does not have an effect; the second column shows where it does have an effect or make a difference.

The left header column describes the world we mortals live in. Regardless of what's true, we have to make decisions about which of our hypotheses is correct. This header column describes the two decisions we can reach -- that our program had no effect (the first row of the 2x2 table) or that it did have an effect (the second row).

Now, let's examine the cells of the 2x2 table. Each cell shows the Greek symbol for that cell. Notice that the columns sum to 1 (i.e.,  $\alpha + (1-\alpha) = 1$  and  $\beta + (1-\beta) = 1$ ): If one column is true, the other is irrelevant -- if the program has a real effect (the right column) it can't at the same time not have one. Therefore, the odds or probabilities have to sum to 1 for each column because the two rows in each column describe the only possible decisions (accept or reject the null/alternative) for each possible reality.

Below the Greek symbol is a typical value for that cell.

The value of  $\alpha$  is typically set at .05 in the social sciences. A newer, but growing, tradition is to try to achieve a statistical power of at least .80. Below the typical values is the name typically given for that cell (in caps). Note that two of the cells describe errors -- you reach the wrong conclusion -- and in the other two you reach the correct conclusion.

Type I [false positive] is the same as the  $\alpha$  or significance level and labels the odds of finding a difference or effect by chance alone. (Is there a psychological, or reporting, bias here?)

Type II [false negative] suggest that you find that the program was not demonstrably effective. (There may be a psychological bias here too but probably a healthy one.)

Think about what happens if you want to increase your power in a study !

	<p><b>H<sub>0</sub> (null hypothesis) true</b></p> <p><b>H<sub>1</sub> (alternative hypothesis) false</b></p> <p><b>In reality...</b></p> <ul style="list-style-type: none"> <li>There is <i>no</i> relationship</li> <li>There is <i>no</i> difference, no gain</li> <li>Our theory is <i>wrong</i></li> </ul>	<p><b>H<sub>0</sub> (null hypothesis) false</b></p> <p><b>H<sub>1</sub> (alternative hypothesis) true</b></p> <p><b>In reality...</b></p> <ul style="list-style-type: none"> <li><b>There <i>is</i> a relationship</b></li> <li><b>There <i>is</i> a difference or gain</b></li> <li><b>Our theory is <i>correct</i></b></li> </ul>
<p><b>We accept the null hypothesis (H<sub>0</sub>)</b></p> <p><b>We reject the alternative hypothesis (H<sub>1</sub>)</b></p> <p><b>We <u>say</u>...</b></p> <ul style="list-style-type: none"> <li>"There is no relationship"</li> <li>"There is no difference, no gain"</li> <li>"Our theory is wrong"</li> </ul>	<p style="text-align: center;"><b>1-<math>\alpha</math></b></p> <p style="text-align: center;">(e.g., .95)</p> <p style="text-align: center;"><b>THE CONFIDENCE LEVEL</b></p> <p>The odds of saying there is no relationship, difference, gain, when in fact there is none</p> <p>The odds of correctly not confirming our theory</p> <p><i>95 times out of 100 when there is no effect, we'll say there is none</i></p>	<p style="text-align: center;"><b><math>\beta</math></b></p> <p style="text-align: center;">(e.g., .20)</p> <p style="text-align: center;"><b>TYPE II ERROR</b></p> <p>The odds of saying there is no relationship, difference, gain, when in fact there is one</p> <p>The odds of not confirming our theory when it's true</p> <p><i>20 times out of 100, when there is an effect, we'll say there isn't</i></p>




<p><b>We reject the null hypothesis (<math>H_0</math>)</b></p> <p><b>We accept the alternative hypothesis (<math>H_1</math>)</b></p> <p><b>We say...</b></p> <ul style="list-style-type: none"> <li> <b>"There is a relationship"</b></li> <li> <b>"There is a difference or gain"</b></li> <li> <b>"Our theory is correct"</b></li> </ul>	<p style="text-align: center;"><math>\alpha</math></p> <p style="text-align: center;">(e.g., .05)</p> <p style="text-align: center;"><b>TYPE I ERROR</b></p> <p style="text-align: center;"><b>(SIGNIFICANCE LEVEL)</b></p> <p>The odds of saying there is an relationship, difference, gain, when in fact there is not</p> <p>The odds of confirming our theory incorrectly</p> <p><i>5 times out of 100, when there is no effect, we'll say there is on</i></p> <p>We should keep this small when we can't afford/risk wrongly concluding that our program works</p>	<p style="text-align: center;"><math>1-\beta</math></p> <p style="text-align: center;">(e.g., .80)</p> <p style="text-align: center;"><b>POWER</b></p> <p>The odds of saying that there is an relationship, difference, gain, when in fact there is one</p> <p>The odds of confirming our theory correctly</p> <p><i>80 times out of 100, when there is an effect, we'll say there is</i></p> <p>We generally want this to be as large as possible</p>
---	--	--

Figure 1. The Statistical Inference Decision Matrix

We often talk about alpha ( $\alpha$ ) and beta ( $\beta$ ) using the language of "higher" and "lower." For instance, we might talk about the advantages of a higher or lower  $\alpha$ -level in a study. You have to be careful about interpreting the meaning of these terms. When we talk about *higher*  $\alpha$ -levels, we mean that we are *increasing* the chance of a Type I Error. Therefore, a *lower*  $\alpha$ -level actually means that you are conducting a *more rigorous* test. With all of this in mind, let's consider a few common associations evident in the table. You should convince yourself of the following:

- the lower the  $\alpha$ , the lower the power; the higher the  $\alpha$ , the higher the power
- the lower the  $\alpha$ , the less likely it is that you will make a Type I Error (i.e., reject the null when it's true)
- the lower the  $\alpha$ , the more "rigorous" the test
- an  $\alpha$  of .01 (compared with .05 or .10) means the researcher is being relatively careful, s/he is only willing to risk being wrong 1 in a 100 times in rejecting the null when it's true (i.e., saying there's an effect when there really isn't)
- an  $\alpha$  of .01 (compared with .05 or .10) limits one's chances of ending up in the bottom row, of concluding that the program has an effect. This means that both your statistical power and the chances of making a Type I Error are lower.
- an  $\alpha$  of .01 means you have a 99% chance of saying there is no difference when there in fact is no difference (being in the upper left box)
- increasing  $\alpha$  (e.g., from .01 to .05 or .10) increases the chances of making a Type I Error (i.e., saying there is a difference when there is not), decreases the chances of making a Type II Error (i.e., saying there is no difference when there is) and decreases the rigor of the test
- increasing  $\alpha$  (e.g., from .01 to .05 or .10) increases power because one will be rejecting the null more often (i.e., accepting the alternative) and, consequently, when the alternative is true, there is a greater chance of accepting it (i.e., power)

Robert M. Becker at Cornell University illustrates these concepts masterfully, and entertainingly, by way of the OJ Simpson trial (note that this is actually a very nice illustration of the advantages of contextualization although there may be order effects here ☺):

<http://www.socialresearchmethods.net/OJtrial/ojhome.htm>

H<sub>0</sub>: OJ Simpson was innocent

(although our theory is that in fact he was guilty as charged)

H<sub>A</sub>: Guilty as charged (double murder)

Can H<sub>0</sub> be rejected, at a high level of confidence? I.e. ...

Type I error? Returning a guilty verdict when the defendant is innocent.

Type II error? Returning a not guilty verdict when the defendant is guilty.

The tradeoff (The Jury's Dilemma): Do we want to make sure

we put guilty people in jail (that would mean, to choose a higher  $\alpha =$

to have less stringent demands on evidence needed)

or we keep innocent people out of jail (that would mean , to choose a lower  $\alpha =$

to have higher demands on evidence needed)

Says Becker, "The standard of reasonable doubt may vary from jury to jury and case to case, but generally, juries unlike social scientists, may be more likely to make (or feel comfortable with) a Type II Error based on the notion of "innocent until proven guilty" (beyond a reasonable doubt) This, of course, makes the prosecutors' life more difficult who have to increase the amount (sample size) and the persuasiveness (effect size) of their evidence in order to increase the chances that the jury would conclude that their theory is indeed the correct theory (power). (By the same token, the defense will try to reduce the effect size through various strategies ... .)

## Drawing a Conclusion

IF we believe there is a correlation (beyond a reasonable doubt) between the evidence and OJ, THEN

Return a Guilty verdict...

ELSE

Return a Not Guilty verdict (there was not enough evidence, power was too low, perhaps the theory is only based on chance, circumstantial)



JURY

Click on Boxes		Reality Truth	
		$H_0$ True	$H_0$ False
What the jury concludes	Accept $H_0$	<b>CONFIDENCE LEVEL</b> OJ is Innocent Not Guilty Verdict	<b>TYPE II Error</b> OJ is Guilty Not Guilty Verdict
	Reject $H_0$	<b>TYPE I Error</b> OJ is Innocent Guilty Verdict	<b>POWER</b> OJ is Guilty Guilty Verdict

**OJ Receives Justice...OJ is a Free Man!**



**CONFIDENCE LEVEL ( $1 - \alpha$ )**

OJ is Innocent ( $H_0$  is true)  
 Not Guilty Verdict ( Jury accepts  $H_0$ )

The jury found reasonable doubt...they correctly concluded that the evidence did not correlate with OJ...they correctly concluded that Marcia and Chris' theory was wrong

**OJ Wrongly Sent to Prison...Free OJ!**

**TYPE I Error ( $\alpha$ )**

OJ is Innocent ( $H_0$  is true)  
 Guilty Verdict (Jury Rejects  $H_0$ )

The jury did not find reasonable doubt...they incorrectly correlated the evidence with OJ...they incorrectly accepted Marcia and Chris' theory

## OJ Gets Off....America Stunned!



**TYPE II Error ( $\beta$ )**

OJ is Guilty ( $H_0$  is False)  
Not Guilty Verdict (Jury Accepts  $H_0$ )

The jury found reasonable doubt...they did not correlate the evidence with OJ...they failed to accept Marcia and Chris' accurate theory (perhaps due to a lack of statistical power)

## OJ Found Guilty...OJ Fans are Stunned...He's Staying in Jail!

**POWER ( $1 - \beta$ )**

OJ is Guilty ( $H_0$  is False)  
Guilty Verdict (Jury Rejects  $H_0$ )

The jury did not find reasonable doubt...they correctly correlated the evidence with OJ...they correctly accepted Marcia and Chris' theory (statistical Power was high)

### 1. Introduction

#### 1.1 Descriptive statistics

**Descriptive statistics** - tools for presenting various characteristics of subjects' behavior as well as their personal characteristics in the form of tables and graphs, and with methods of summarizing the characteristics by measures of central tendency, variability, and so on.

One normally observes variation in characteristics between (or across) subjects, but sometimes also within subjects – for example, if subjects' performance varies from round to round of an experiment.



**Inferential statistics** - formal statistical methods of making inferences (i.e., conclusions) or predictions regarding subjects' behavior.

Types of variables (Stevens 1946)

- *categorical* variables (e.g., gender, or field of study)
- *ordinal* variables (e.g., performance rank)
- *interval* variables (e.g., wealth or income bracket)
- *ratio* variables (e.g., performance score, or the number of subjects choosing option A rather than option B).

Different statistical approaches may be required by different types of variables.

### 1.1.1 Measures of central tendency and variability

#### **Measures of central tendency**

- the *arithmetic mean* (the average of a variable's values)
- the *mode* (the most frequently occurring value(s) of a variable)
- the *median* (the middle-ranked value of a variable)
  - useful when the variable's distribution is asymmetric or contains outliers

#### **Measures of variability**

- the *variance* (the average of the squared deviations of a variable's values from the variable's arithmetic mean)
- an unbiased estimate of the population variance,  $\hat{s}^2 = ns^2/(n-1)$ , where  $s^2$  is the sample variance as defined in words directly above, and  $n$  is the number of observations on the variable under study)
- the *standard deviation* (the square root of the variance)
- the *range* (the difference between a variable's highest and lowest value)
- the *interquartile range* (the difference between a variable's values at the first quartile (i.e., the 25<sup>th</sup> percentile) and the third quartile (i.e., the 75<sup>th</sup> percentile))
- Furthermore, ... measures assessing the shape of a variable's distribution – such as the degree of symmetry (*skewness*) and peakedness (*kurtosis*) of the distribution – useful when comparing the variable's distribution to a theoretical probability distribution (such as the normal distribution, which is symmetric and moderately peaked).

### 1.1.2 Tabular and graphical representation of data

**ALWAYS inspect the data by visual means before conducting formal statistical tests!  
And do it on as disaggregated level as possible!**

## **1.2 Inferential statistics**

We use a *sample statistic* such as the sample mean to make inferences about a (unknown) *population parameter* such as the population mean.<sup>1</sup>

---

<sup>1</sup> As further discussed below, random sampling is important for making a sample representative of the population we have in mind, and consequently for drawing valid conclusions about population parameters based on sample statistics. Recall the problematic recruiting procedure in Hoelzl Rustichini (2005) and Harrison's et al (2005) critique of the unbalanced subject pools in Holt & Laury (2002).

Difference between the two is the **sampling error**, it decreases with larger sample size. Sample statistics draw on measures of central tendency and variability, so the fields of descriptive and inferential statistics are closely related: A sample statistic can be used for summarizing sample behavior as well as for making inferences about a corresponding population parameter.

### 1.2.1 Hypothesis testing (as opposed to estimation of population parameters – see 2.1.1.)

*classical hypothesis testing model*

$H_0$ , of *no effect* (or *no difference*) versus

$H_1$ , of the *presence of an effect* (or *presence of a difference*)

where  $H_1$  is stated as either nondirectional (two-tailed) if no prediction about the direction of the effect or difference, or directional (one-tailed) if prediction (researchers sometimes speak of two-tailed and one-tailed statistical tests, respectively).

A more conservative approach is to use a nondirectional (two-tailed)  $H_1$ .

Can we reject  $H_0$  in favor of  $H_1$ ?

Example: Two groups of subjects facing different experimental conditions:

Does difference in experimental conditions affects subjects' average performance?

$H_0: \mu_1 = \mu_2$  and  $H_1: \mu_1 \neq \mu_2$ , or  $H_1: \mu_1 > \mu_2$  or  $H_1: \mu_1 < \mu_2$ , if we have theoretical or practical reasons for entertaining a directional research hypothesis,

where  $\mu_i$  denotes the mean performance of subjects in Population  $i$  from which Sample  $i$  was drawn. How confident are we about our conclusion?

### 1.2.2 The basics of inferential statistical tests

- compute a *test statistic* based on sample data
- compare to the theoretical probability distribution of the test statistic constructed assuming that  $H_0$  is true
- If the computed value of the test statistic falls in the extreme tail(s) of the theoretical probability distribution – the tail(s) being delimited from the rest of the distribution by the so called *critical value(s)* – conclude that  $H_0$  is rejected in favor of  $H_1$ ; otherwise conclude that  $H_0$  of no effect (or no difference) cannot be rejected. By rejecting  $H_0$ , we declare that the effect on (or difference in) behavior observed in our subject sample is *statistically significant*, meaning that the effect (or difference) is highly unlikely due to chance (i.e., random variation) but rather due to some systematic factors.

By **convention**, *level of statistical significance* (or significance level),  $\alpha$ , often set at 5% ( $\alpha=.05$ ), sometimes at 1% ( $\alpha=.01$ ) or 10% ( $\alpha=.10$ ).

Alternatively, **one may instead (or additionally) wish to report the exact probability value (or *p*-value),  $p$ , at which statistical significance would be declared.**

**The significance level at which  $H_0$  is evaluated and the type of  $H_1$  (one-tailed or two-tailed) ought to be chosen (i.e., predetermined) by the researcher prior to conducting the statistical test or even prior to data collection.**

The critical values of common theoretical probability distributions of test statistics, for various significance levels and both types of  $H_1$ , are usually listed in special tables in appendices of statistics (text) books and in Appendix X of Ondrej's chapter.

### 1.2.3 Type I and Type II errors, power of a statistical test, and effect size

Lowering  $\alpha$  (for a given  $H_1$ )

- increases the probability of a *Type II error*,  $\beta$ , which is committed when a false  $H_0$  is erroneously accepted despite  $H_1$  being true.

- decreases the *power* of a statistical test,  $1 - \beta$ , the probability of rejecting a false  $H_0$ .

Thus, in choosing a significance level at which to evaluate  $H_0$ , one faces a tradeoff between the probabilities of committing the above statistical errors.

Other things equal, the larger the sample size and the smaller the sampling error, the higher the likelihood of rejecting  $H_0$  and hence the higher the power of a statistical test.

The probability of committing a Type II error as well as the power of a statistical test can only be determined after specifying the value of the relevant population parameter(s) under  $H_1$ .

Other things equal, the test's power increases the larger the difference between the values of the relevant population parameter(s) under  $H_0$  and  $H_1$ .

This difference, when expressed in standard deviation units of the variable under study, is sometimes called the *effect size* (or Cohen's  $d$  index).

Especially in the context of parametric statistical tests, some scientists prefer to do a power-planning exercise prior to conducting an experiment: After specifying a minimum effect size they wish to detect in the experiment, they determine such a sample size that yields what they deem to be sufficient power of the statistical test to be used.

Note, however, that one may not know a priori which statistical test is most appropriate and thus how to perform the calculation. In addition, existing criteria for identifying what constitutes a large or small effect size are rather arbitrary (Cohen (1977) proposes that  $d$  greater than 0.8 (0.5, 0.2.) standard deviation units represents a large (medium, small) effect size).

Other things equal, however, the smaller the (expected) effect size, the larger the sample size required to yield a sufficiently powerful test capable of detecting the effect. See, e.g., [S] pp. 164-173 and pp. 408-412 for more details.

Criticisms of the classical hypothesis testing model:

Namely, with a large enough sample size, one can almost always obtain a statistically significant effect, even for a negligible effect size (by similar token, of course, a relatively large effect size may turn out statistically insignificant in small samples).

Yet if one statistically rejects  $H_0$  in a situation where the observed effect size is practically or theoretically negligible, one is in a practical sense committing a Type I error. For this reason, one should strive to assess whether or not the observed effect size – i.e., the observed magnitude of the effect on (or difference in) behavior – is of any practical or theoretical significance. To do so, some researchers prefer to report what is usually referred to as the magnitude of *treatment effect*, which is also a measure of effect size (and is in fact related to Cohen's  $d$  index). We discuss the notion of treatment effect in Sections 2.2.1 and 2.3.1, and see also [S] pp.1037-1061 for more details.

Another criticism: improper use, particularly in relation to the true likelihood of committing a Type I and Type II error. Within the context of a given research hypothesis, statistical comparisons and their significance level should be specified *prior* to conducting the tests. If additional *unplanned* tests are conducted, the overall likelihood of committing a Type I error in such an analysis is inevitably inflated well beyond the  $\alpha$  significance level prespecified for the additional tests. For explanation, and possible remedies, see Ondrej's text,

Alternatives:

the *minimum-effect hypothesis testing model*

the *Bayesian hypothesis testing model*

See Cohen (1994), Gigerenzer (1993) and [S] pp. 303-350 for more details, also text.

### **1.3 The experimental method and experimental design**

How experimental economists and other scientists design experiments to evaluate research hypotheses.

Proper design and execution of your experiment ensure reliability of your data and hence also the reliability of your subsequent statistical inference. (Statistical tests are unable to detect flaws in experimental design and implementation.)

A typical research hypothesis involves a prediction about a *causal* relationship between an *independent* and a *dependent* variable (e.g.. effect of financial incentives on risk aversion, or on effort, etc.)

A common experimental approach to studying the relationship is to compare the behavior of two groups of subjects: the *treatment* (or experimental) group and the *control* (or comparison) group.

The independent variable is the experimental conditions – manipulated by the experimenter – that distinguish the treatment and control groups (one can have more than one treatment group and hence more than two levels of the independent variable).

The dependent variable is the characteristic of subjects' behavior predicted by the research hypothesis to depend on the level of the independent variable (one can also have more than one dependent variable).

In turn, one uses an appropriate inferential statistical test to evaluate whether there indeed is a statistically significant difference in the dependent variable between the treatment and control groups.

What we describe above is commonly referred to as *true experimental designs*, characterized by a random assignment of subjects into the treatment and control groups (i.e., there exists at least one adequate control group) and by the independent variable being exogenously manipulated by the experimenter in line with the research hypothesis.

These characteristics jointly limit the influence of confounding factors and thereby maximize the likelihood of the experiment having *internal validity*, which is achieved to the extent that observed differences in the dependent variable can be unambiguously attributed to a manipulated independent variable.

*Confounding factors* are variables systematically varying with the independent variable (e.g., yesterday's seminar?), which may produce a difference in the dependent variable that only appears like a causal effect. Unlike true experiments, other types of experiments conducted outside the laboratory – such as what is commonly referred to as natural experiments exercise less or no control over random assignment of subjects and exogenous manipulation of the independent variable, and hence are more prone to the potential effect of confounding variables and have lower internal validity.

Random assignment of subjects conveniently maximizes the probability of obtaining control and treatment groups equivalent with respect to potentially relevant individual differences such as demographic characteristics. As a result, any difference in the dependent variable between the treatment and control groups is most likely attributable to the manipulation of the independent variable and hence to the hypothesized causal relationship. Nevertheless, the equivalence of the control and treatment groups is rarely achieved in practice, and one should control for any differences between the control and treatment groups if deemed necessary (e.g., as illustrated in Harrison et al 2005, in their critique of Holt & Laury 2002).

Similarly, one should not simply assume that a subject sample is drawn randomly and hence is representative of the population under study. Consciously or otherwise, we often deal with nonrandom samples. Volunteer subjects, or subjects selected based on their availability at the time of the experimental sessions, are unlikely to constitute true random samples but rather *convenience samples*. As a consequence, the *external validity* of our results – i.e., the extent to which our conclusions generalize beyond the subject sample(s) used in the experiment – may suffer.<sup>2</sup>

Choosing an appropriate experimental design often involves tradeoffs. One must pay attention to the costs of the design in terms of the number of subjects and the amount of money and time required, to whether the design will yield reliable results in terms of internal and external validity, and to the practicality of implementing the design.

In other words, you may encounter practical, financial or ethical limitations preventing you from employing the theoretically best design in terms of the internal and external validity.

#### **1.4 Selecting an appropriate inferential statistical test**

- Determine whether the hypothesis (and hence your data set) involves one or more samples.
- Single sample: use a *single-sample* statistical test to test for the absence or presence of an effect on behavior, along the lines described in the first example in Section 1.2.1.
- Two samples: use a *two-sample* statistical test for the absence or presence of a difference in behavior, along the lines described in the second example in Section 1.2.1.
- Most common single- and two-sample statistical tests in Sections 2 to 6; other statistical tests and procedures intended for two or more samples are not discussed in this book but can be reviewed, for example, in [S] pp. 683-898.

---

<sup>2</sup> In the case of convenience samples, one usually does not know the probability of a subject being selected. Consequently one cannot employ methods of survey research that use *known* probabilities of subjects' selection to correct for the nonrandom selection and thereby to make the sample representative of the population. One should rather employ methods of correcting for how subjects select into participating in the experiment (see, e.g., Harrison et al., UCF WP 2005, forthcoming in JEBO?).

When making a decision on the appropriate two-sample test, one first needs to determine whether the samples – usually the treatment and control groups/conditions in the context of the true experimental design described in Section 1.3 – are independent or dependent.

*Independent samples design (or between-subjects design, or randomized-groups design)* – where subjects are randomly assigned to two or more experimental and control groups – one employs a *test for two (or more) independent samples*.

*Dependent samples design (or within-subjects design, or randomized-blocks design)* – where each subject serves in each of the  $k$  experimental conditions, or, in the matched-subjects design, each subject is matched with one subject from each of the other  $(k-1)$  experimental conditions based on some observable characteristic(s) believed to be correlated with the dependent variable – one employs a *test for dependent samples*.

One needs to **ensure internal validity of the dependent samples design by controlling for order effects** (so that differences between experimental conditions do not arise solely from the order of their presentation to subjects), and, in the matched-subjects design, by ensuring that matched subjects are closely similar with respect to the matching characteristic(s) and (within each pair) are assigned randomly to the experimental conditions.

Finally, in *factorial designs*, one simultaneously evaluates the effect of several independent variables (factors) and conveniently also their interactions, which usually requires using a *test for factorial analysis of variance* or other techniques (which are not discussed in this book but can be reviewed, e.g., in [S] pp.900-955).

Sections 2 to 6: we discuss the most common single- and two-sample *parametric* and *nonparametric* inferential statistical tests.

The parametric label is usually used for tests that make stronger assumptions about the population parameter(s) of the underlying distribution(s) for which the tests are employed, as compared to non-parametric tests that make weaker assumptions (for this reason, the non-parametric label may be slightly misleading since nonparametric tests are rarely free of distributional and other assumptions).

Some researchers instead prefer to make the parametric-nonparametric distinction based on the type of variables analyzed by the tests, with nonparametric tests analyzing primarily categorical and ordinal variables with lower informational content (see Section 1.1).

**Behind the alternative classification is the widespread (but not universal) belief that a parametric test is generally more powerful than its nonparametric counterpart *provided* the assumption(s) underlying the former test are satisfied, but that a violation of the assumption(s) calls for transforming the data into a format of (usually) lower informational content and analyzing the transformed data by a nonparametric test.**

Alternatively, ... use parametric tests even if some of their underlying assumptions are violated, but make adjustments to the test statistic to improve its reliability.

While the reliability and validity of statistical conclusions depends on using appropriate statistical tests, one often cannot fully validate the assumptions underlying specific tests and hence faces the risk of making wrong inferences. For this reason, **one is generally advised to conduct both parametric and nonparametric tests to evaluate a given statistical hypothesis, and – especially if results of alternative tests disagree – to conduct multiple experiments evaluating the research hypothesis under study and jointly analyze their results by using meta-analytic procedures.** See, e.g., [S] pp.1037-1061 for further details. Of course, that's only possible if you have enough resources.

## **2. $t$ tests for evaluating a hypothesis about population mean(s)**

### **3. Nonparametric alternatives to the single-sample $t$ test**

### **4. Nonparametric alternatives to the $t$ test for two independent samples**

### **5. Nonparametric alternatives to the $t$ test for two dependent samples**

## **6. A brief discussion of other statistical tests**

### **6.1 Tests for evaluating population skewness and kurtosis**

### **6.2 Tests for evaluating population variability**



## “Linear Public Goods Experiments: A Meta-Analysis”

Author: Jennifer Zelmer  
*Experimental Economics*, 6: 299-310 (2003)

### Introduction

- **1<sup>st</sup> Fact:** EconLit tracks over 600 journals, as well as a wide range of books and dissertations (AEA, 1999).
- ✓ **1<sup>st</sup> Problem:** single experiments or studies in the social sciences rarely provide definitive answers to a research question (Wolf, 1986).
- **2<sup>nd</sup> Fact:** the literature about how different *factors* affect individuals’ willingness to contribute to public goods has accumulated many data and results (sometimes conflicting) for more than two decades.
- ✓ **2<sup>nd</sup> Problem:** survey articles and qualitative reviews do not provide *estimates* of the *effect size* of factors based on the totality of evidence.

...these problems can be approached with the meta-analysis... 3

### Introduction

#### ■ Meta-Analysis

**definition:** “the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings into a single, generalizable finding” (Plath 1992, as cited in Zelmer p. 300);

**main steps:**

1. Identifies a *sample* (as complete as possible) of studies pertaining to the same issue;
2. Describes features and results in a *consistent quantitative way*;
3. Applies statistical techniques to *aggregate* the findings across studies and objectively *examine the relationship* between study characteristics and outcomes;
4. Concludes with a systematic and detailed description of the method used to integrate study results, *ensuring replicability* of the findings.

**consideration:**

applications of meta-analysis in economics continue to be relatively rare (van de Bergh et al., 1997; Croson and Marks, EE 2000, for provision point public good experiments; see various references of Zelmer to latter paper). 4

## Study objective

- “The objective of this meta analysis is to **synthesize** the results of existing experimental evidence on the impact of a variety of [...] **factors** on the **extent of cooperation observed** in standard linear public goods experiments using the voluntary contribution mechanism” (p. 301)

## Theoretical framework

### Voluntary contributions to public goods

- What are **public goods**?  
Commodities for which use of the good by one agent does not preclude its use by others (Pigou, 1932). (**non-rival**)
- Under what conditions are individuals more likely to voluntarily contribute to the provision of public goods?
- The game- theoretical framework in which the primary sources face this problem is the **voluntary contribution mechanism** in linear public goods environment (VCM).

6

## Theoretical framework

### Voluntary contributions to public goods

- Subjects are divided into groups and play the same game for a finite number of periods.
- Endowment for each period:  $w_i$ .
- $w_i$  has to be divided between:
  - $x_i$  = contribution to private account (constant return to him/herself only)
  - $g_i$  = contribution to public good (benefits to all group)
- at the end of each period subjects (usually) learn:
  - the aggregate contribution to the public good;
  - their earnings for the period.

7

## Theoretical framework

### Voluntary contributions to public goods

- Individual  $i$ 's utility function:  $u_i = \alpha x_i + \beta g_i \rightarrow \text{maximize}$
- Budget constraint:  $w_i = x_i + g_i$
- Public good identity:  $G = \sum_i g_i$
- Non negativity constraint:  $g_i \geq 0$
- $\alpha$  and  $\beta$  (*m.p.c. return*) are constants;
- Subjects in finitely repeated games (i.e. one shot games) have a dominant strategy to contribute nothing to the public good;
- **Nash-equilibrium** is full free riding;
- **Pareto-efficient** outcome is for all subjects to contribute their entire endowment to the public good

8

# Theoretical framework

## Voluntary contributions to public goods

### An example...<sup>1</sup>

- 2x2 matrix game
- Number of subjects = 2
- $w_i = 10$ ;  $\alpha = 1$ ;  $\beta = 0.7$ ;
- *only two actions are possible* for the subjects:
  - to contribute for all the endowment (invest all)
  - to contribute for nothing (invest nothing)

(Prisoner Dilemma)

		Agent 2	
		Invest nothing (\$0)*	Invest all (\$10)
Agent 1	Invest nothing (\$0)*	10, 10	7, 7
	Invest all (\$10)	7, 17	14, 14

<sup>1</sup> This example was taken from Saijo and Nakamura (1995), one of the articles used in the meta-analysis.

9

## Methods

### 1. Searching

A search of the economics literature of *standard single-stage linear public goods in VCM*, was conducted using three sources: **EconLit**, Internet documents in Economics Access Service (**IDEAS**), and references cited in John Ledyard's (1995) survey of experimental research related to public goods.

- **Four** keyword searches: "public goods" and "experiment\*", "voluntary contribution" and "experiment", "variable contribution" and "experiment" and "cooperation" and "experiment".
- **Two** subject heading searches were conducted using the **JEL classification system**:
  - i. area H410 and keyword "experiment"
  - ii. areas C900,C910,C920 and C990, and the keyword "public goods"

• **Detected 349 potential primary sources.** ←

10

## Methods

### 2. Selection criteria

Titles and abstracts were screened using these **inclusion criteria**:

1. unique **reports** of a laboratory experiment, observations gathered in a **controlled environment**;
2. standard V.C.M. in a single-stage linear public goods environment, where  $\beta < 1$ ;
3. reported **group-level results** for at least one of the outcome of interest;
4. Could be obtained through electronic access or libraries (at Toronto and York Universities) or *www*.

(where it was possible, potential relevant studies were retrieved for a more detailed evaluation = full review)

• **27 studies** were included in the meta-analysis. ←

11

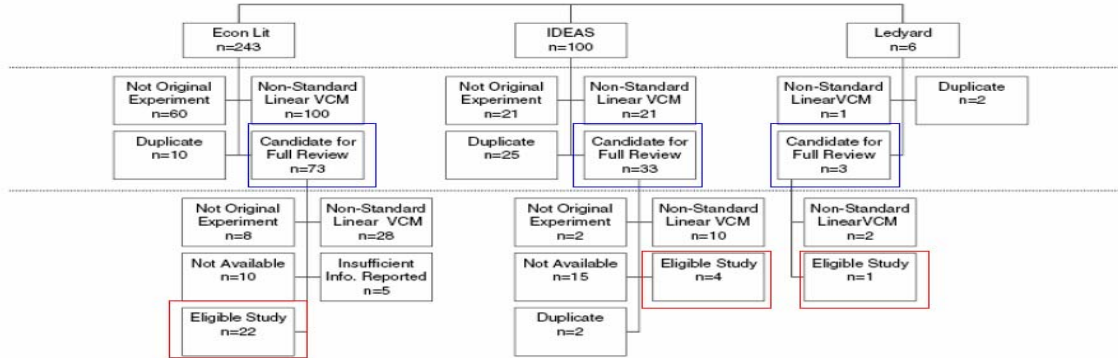
27/349 =

Is less than the 10% of the *primary sources*, obtained through the search phase!

# Methods

## 2. Selection criteria

Flow-chart of the search and review process.



12

# Methods

## 2. Selection criteria

Final coverage of the meta-analysis:

- 711 distinct groups (in experimental session with the same conditions)
- 7240 experimental periods (together)
- 13 groups were excluded due to missing data for the dependent variable

Table 1. Selected parameters for experiments included in the meta-analysis.

Author(s)	Year of publication	# Periods	Group size	MPCR
Andreoni	1988	3-10	5	0.5
Andreoni	1995a	10	5	0.5
Andreoni	1995b	10	5	0.5
Asch, Gligioni, and Pelfico	1993	5	8-12	0.3
Cason and Khan	1999	6-24	4	0.3
Crosan	1996	10	4	0.5
Crosan	2000	10	4	0.125
Dorsey	1992	10	4	0.3
Falkinger, Fehr, Gächter, and Wimmer-Ebmer	1999/2000 <sup>a</sup>	10	4-16	0.1-0.4
Fehr and Gächter	1999/2000 <sup>b</sup>	6-10	4	0.4
Fisher, Isaac, Sholtzberg, and Walker	1995	10	4	0.3-0.75
Gächter, Fehr, and Knörr	1996	10	4	0.4
Gosser, Holt, and Laury	1999/2002 <sup>c</sup>	17	2-4	0.2-0.8
Harbaugh and Krause	2000	10	6	0.33-0.67
Isaac and Walker	1988a	10	4	0.003
Isaac and Walker	1988b	10	4-10	0.3-0.75
Isaac and Walker	1998	10	4	0.3
Isaac, Walker, and Thomas	1984	10	4-10	0.3-0.75
Isaac, Walker, and Williams	1994	10-60	4-100	0.03-0.75
Keser and van Winden	1996/2000 <sup>d</sup>	25	4	0.5
Laury, Walker, and Williams	1995	15-30	4	0.55
McCorkle and Watts	1996	1	49	0.04
Nowell and Tränkle	1994	13	4	0.3
Ockenfels and Weimann	1999	10	5	0.33
Sajjo and Nakamura	1995	10	7	0.7
Weimann	1994	10	5	0.5
Wilson and Sell	1997	18	6	0.003

13

# Methods

## 3. Validity assessment and data abstraction

• For the selected studies, author abstracted **bibliographic details**, **contextual information** (for each experiment) and **data** on each session in the experiments, and put them in a Microsoft Access Database, in order to improve **data reliability** and to facilitate potential **replication** of data extraction and coding.

• For this procedure a pilot test was conducted using 7 experimental reports. Two experts (A. Muller and S. Mastelman) reviewed the **study selection criteria** and **abstraction protocols** prior to implementation.

14

# Model adopted

**Method:** Meta-regression method, weighted least squares<sup>2</sup> of group level results.

**Dependent variable:** average efficiency of the group's contributions over the session.

**Explanatory variables:**

1. *literature's relevant characteristics* of the public goods environment:
  - m.p.c. return
  - group size
  - gender of subjects
  - extent of subject experience
  - extent of communication allowed to subjects
2. *variable describing study design and experimental design:*
  - cash or benefits?
  - fully computerized environment or not?
3. *dummy variable to track quality problems* reported by the primary searches
4. *dummy variables indicating publication info* (published journal or w.paper?)
5. *dummy variables for each experiments*, where possible

<sup>2</sup> in the weighted least squares (WLS) regression, each data point (set of value) potentially receives a different weight. The appropriate weight to assign is one which is proportional to how well the dependent variable is known or inversely proportional to the variability of the dependent variable.

16

# Methods

## 3. Validity assessment and data abstraction

### • Comments on data abstraction:

✓ **type** and **extent** of data reported in primary studies **varied significantly**

✓ more reported information: **m.p.c.**, **group size**, **fully computerize or not**;

✓ less reported information: **nature of the subjects pool**, **date and place** of the experiment, **average payments** to subjects

✓ three variables were taken off the quantitative analysis because of inconsistent reporting: **average payments to subjects** as a proxy for salience rewards (!!!), year of the experiment, methods use to randomize subjects to different treatment groups.

✓ for selected variables, standardized imputations were conducted according to the study protocol

15



# Results

## Quantitative data synthesis – Model adopted

### Not reported data:

- results of analyses about the ***decay in efficiency over a session***
- proportion of complete free riders
- parameter estimates and other information on the dummy variables of each experiment

} Few primary studies included these data

} To facilitate interpretation of results

---

# Results

Quantitative data synthesis  
Weighted least squares  
regression results  
(adjusted  $r^2=0.6115$ )

$\Delta_+$

Factors which  
(significantly) affect  
mean contributions  
to the public good in  
a **positive** way.

Table 3. Weighted least squares results—Average contributions as a percent of the total endowment.

Variable	Estimate	Std error	p-value	sig
Intercept	-14.87	18.44	0.4207	
# Periods	-0.44	0.29	0.1376	
Friendship among subjects	1.50	7.05	0.8320	
Group size	0.15	0.09	0.0948	
Cash rewards	15.36	10.92	0.1605	
Fully computerized environment	1.10	5.72	0.8479	
Marginal per capita return	39.53	6.12	<.0001	**
Male subjects only	1.00	13.16	0.9395	
Female subjects only	8.00	12.60	0.5260	
Child subjects	44.85	22.49	0.0472	*
Heterogeneous MPCR	-0.54	12.65	0.9657	
Heterogeneous endowments	-14.51	7.10	0.0421	*
Experienced subjects	-6.15	2.55	0.0167	*
Communication allowed	40.46	4.16	<.0001	**
Punishment of subjects allowed	1.86	6.16	0.7637	
Economics training	6.05	5.87	0.3039	
Positive framing	19.30	7.90	0.0151	*
Optimum announced	-0.46	12.99	0.9716	
End of session announced	6.48	9.98	0.5168	
Quality problems identified	-5.55	6.95	0.4255	
Imperfect monitoring of group contributions	2.25	7.19	0.7550	
Beliefs re: others' behaviour solicited	-20.00	8.49	0.0193	*
Constant groups for session ("partners")	15.67	3.54	<.0001	**
Subjects from western europe	-0.55	10.14	0.9568	
Subjects from eastern europe	-10.78	11.20	0.3368	
Japanese subjects	-10.60	13.22	0.4232	
Published in journal	-6.28	15.24	0.6807	

19

# Results

Quantitative data synthesis  
Weighted least squares  
regression results  
(adjusted  $r^2=0.6115$ )

$\Delta_-$

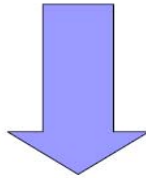
Factors which  
(significantly) affect  
mean contributions  
to the public good in  
a **negative** way.

Table 3. Weighted least squares results—Average contributions as a percent of the total endowment.

Variable	Estimate	Std error	p-value	sig
Intercept	-14.87	18.44	0.4207	
# Periods	-0.44	0.29	0.1376	
Friendship among subjects	1.50	7.05	0.8320	
Group size	0.15	0.09	0.0948	
Cash rewards	15.36	10.92	0.1605	
Fully computerized environment	1.10	5.72	0.8479	
Marginal per capita return	39.53	6.12	<.0001	**
Male subjects only	1.00	13.16	0.9395	
Female subjects only	8.00	12.60	0.5260	
Child subjects	44.85	22.49	0.0472	*
Heterogeneous MPCR	-0.54	12.65	0.9657	
Heterogeneous endowments	-14.51	7.10	0.0421	*
Experienced subjects	-6.15	2.55	0.0167	*
Communication allowed	40.46	4.16	<.0001	**
Punishment of subjects allowed	1.86	6.16	0.7637	
Economics training	6.05	5.87	0.3039	
Positive framing	19.30	7.90	0.0151	*
Optimum announced	-0.46	12.99	0.9716	
End of session announced	6.48	9.98	0.5168	
Quality problems identified	-5.55	6.95	0.4255	
Imperfect monitoring of group contributions	2.25	7.19	0.7550	
Beliefs re: others' behaviour solicited	-20.00	8.49	0.0193	*
Constant groups for session ("partners")	15.67	3.54	<.0001	**
Subjects from western europe	-0.55	10.14	0.9568	
Subjects from eastern europe	-10.78	11.20	0.3368	
Japanese subjects	-10.60	13.22	0.4232	
Published in journal	-6.28	15.24	0.6807	

20

## Other results



Variables included in the meta-regression which have **relatively little overall influence** on the meta-analysis result (either positively or negatively).

Table 3. Weighted least squares results—Average contributions as a percent of the total endowment.

Variable	Estimate	Std error	p-value	sig
Intercept	-14.87	18.44	0.4207	
# Periods	-0.44	0.20	0.1376	
Friendship among subjects	1.50	7.05	0.8320	
Group size	0.15	0.09	0.0948	
Cash rewards	15.36	10.92	0.1605	
Fully computerized environment	1.10	5.72	0.8479	
Marginal per capita return	39.53	6.12	<.0001	**
Male subjects only	0.00	13.05	0.9395	
Female subjects only	8.00	12.60	0.5260	
Child subjects	44.85	22.49	0.0472	*
Heterogeneous MPCR	-0.54	12.65	0.9657	
Heterogeneous endowments	-14.51	7.10	0.0421	*
Experienced subjects	-6.15	2.55	0.0167	*
Communication allowed	40.46	4.16	<.0001	**
Punishment of subjects allowed	1.86	6.16	0.7637	
Economics training	0.05	5.87	0.3039	
Positive framing	19.30	7.90	0.0151	*
Optimum announced	-0.46	12.99	0.9716	
End of session announced	6.48	9.98	0.5168	
Quality problems identified	-5.55	6.95	0.4255	
Imperfect monitoring of group contributions	2.25	7.19	0.7550	
Beliefs re: others' behaviour solicited	-20.00	8.49	0.0193	*
Constant groups for session ("partners")	15.67	3.54	<.0001	**
Subjects from western europe	-0.55	10.14	0.9568	
Subjects from eastern europe	-10.78	11.20	0.3368	
Japanese subjects	-10.60	13.22	0.4232	
Published in journal	-6.28	15.24	0.6807	

21

## Discussion of the results

### Results which support the Ledyard (1995) view:

- higher *m.p.c* returns →  $\Delta_+$ ;
- allowing **communication** →  $\Delta_+$ ;
- heterogeneous endowments →  $\Delta_-$ ;
- **group size** and gender are → not significant;
- **experienced** subject →  $\Delta_-$ .

See Bosch-Domenech et al. (2002); possible explanations of the discrepancy:

- How is the term "trained" used?
- Different contexts?  
→ to be discussed!

### Results which do not support the Ledyard (1995) view:

- **repetition** → has not a significant effect<sup>3</sup> on decreasing contribution;
- **"economic training"** → not significant in decreasing contribution<sup>4</sup>;
- relationship **friendship/group**: maintaining the same group →  $\Delta_+$ ;
- **framing effect**: not yet studied by Ledyard (see Andreoni, 1995b) →  $\Delta_+$ .

<sup>3</sup> a separate analysis showed that at least in those studies where data were reported, contributions declined sharply between the first and the last periods (non-linear relationship due to the end of the game).

<sup>4</sup> maybe due to the low levels of training among most subjects categorized as "economics-trained"

22





## Discussion of the results

Meta-analysis also provided:

- **parameter estimates** for each variable → useful in developing hypotheses regarding the combined effect of different factors;
- **a priori power calculations**: as the likelihood on detecting a difference of a specified size if (it exists) and **focusing experimenters on areas where experimental evidence is less strong**;

For the author, it is likely that meta-analysis will become more popular in economics; but this kind of analysis strongly depends on **consistent** and **complete reporting** of the methods and results of primary sources.

Many of the articles reviewed for this research **do not include key information** about experimental design and results. → Solution: to use existing guidelines (Palfrey and Porter, 1991).

Also increasing **experimental data-sharing** would be useful.

23

On with the show ☺

## 10 Can public goods experiments inform policy?

### Interpreting results in the presence of confused subjects

*Stephen J. Cotten, Paul J. Ferraro, and Christian A. Vossler*

This is a chapter in Cherry et al (2008), Environmental Economics, Experimental Methods, Routledge.

- VCM (= voluntary contributions mechanism) is the cornerstone of experimental investigations on the private provision of public goods
- Standard experimental investigation places individuals in a context-free setting where the public good, which is non-rival and non-excludable in consumption, simply money
- Specifically, “tokens” have to be divided between a private and a public account
- Typically, parameterized/designed so that each player has a dominant strategy of not contributing (to the public account)
- In one-shot (single-round) VCM experiments, subjects contribute – contrary to the theoretical prediction – about 40% - 60 %
- In finitely-repeated VCM experiments, subjects contribute about the same initially but contributions then decline towards zero (but rarely ever zero)

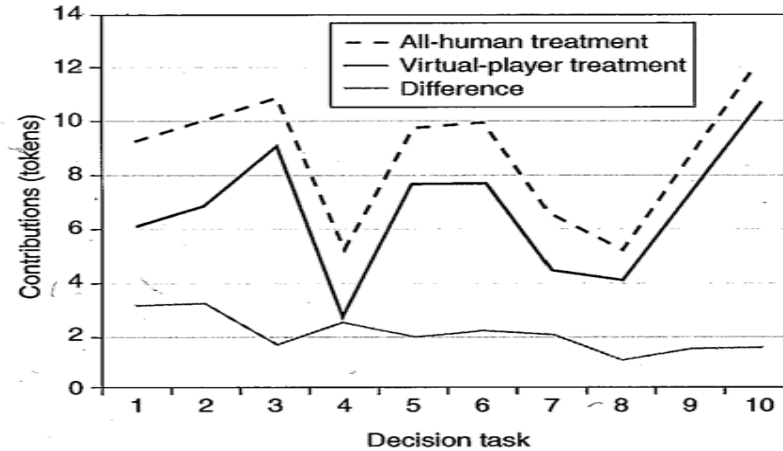
- “Thus, there seem to be motives for contributing that outweigh the incentive to free ride” (CFV 194)
- Possible “motives”: “pure altruism”, “warm-glow” (also called, “impure altruism”), “conditional cooperation”, “confusion”
- “Confusion” describes individuals’ failure to identify (in the laboratory set-up) the dominant strategy of no contribution (a realistic concern, see Rydval, Ortmann, Ostatnicky, Three Simple Games and How to Solve Them, now forthcoming in Journal of Economic Behavior and Organization: <http://www.cerge-ei.cz/pdf/wp/Wp347.pdf>)
- Findings:
  - o Palfrey & Prisbey (AER 1997) - find warm-glow but no evidence of pure altruism
  - o Goeree et al. (JPubE 2002) - find pure altruism but no warm-glow
  - o Fischbacher et al. (EL 2001) – find conditional cooperation but no pure / impure altruism, as do Fischbacher & Gaechter (manuscript 2004)
  - o etc. (contradictory gender effects, but see Ortmann & Tichy JEBO 1999)
  - o apparent lack of correspondence between contributions behavior in experimental and naturally occurring settings (e.g., Laury & Taylor JEBO 2008)
- Could it be that these findings are the result of confusion that “confounds” the interpretation of behavior in public good experiments? (p. 195)
  - o One new experiment, two old ones
  - o Using the “virtual-player” method to sort out pro-social motives such as altruism ...
- Finding:
  - o “The level of confusion in all experiments is both substantial and troubling.” (p. 196)
  - o “The experiments provide evidence that confusion is a confounding factor in investigations that discriminate among motives for public contributions, ... “ (p. 196)
- Solutions:
  - o Increase monetary rewards in VCM experiments ! (inadequate monetary rewards having been identified as potential cause of contributions provided out of confusion)
  - o Make sure instructions are understandable ! (poorly prepared instructions having been identified as possible source of confusion)
  - o Make sure, more generally, that subjects manage to identify the dominant strategy ! (the inability of subjects to decipher the dominant strategy having been identified as a possible source of confusion)
  - o “Our results call into question the standard, “context-free” instructions used in public good games.” (p. 208)

In more detail:

- Andreoni (AER 1995) first to argue that (parts of) what looks like kindness in VCM experiments is really confusion. Andreoni finds that other-regarding behavior (kindness, altruism) and confusion are “equally important”
- Houser & Kurzban (AER 2002) did the same thing but they used a different set-up:
  - a “human condition” (the standard VCM game)
  - a “computer condition” (the standard VCM game, played by one human player and three non-human (or, “virtual”) players.
  - Each round, the aggregate computer contribution to the public good is three-quarters of the average contribution observed for that round in the human condition.
  - Basic idea: confusion and other-regarding behavior present in the human condition but not in the computer condition
  - Basic result: Confusion accounts for about 54 percent of contributions to all public good contributions.
- Ferraro et al. (JEBO 2003) and Ferraro & Vossler (manuscript 2005), with designs similar to Houser & Kurzban find that 54 and 52 percent contributions come from confused subjects.
- Palfrey & Prisbey (1997) find a similar result in their own experiment (not using virtual players) and estimate with their model that “well over half” of the contributions in the classic VCM experiments by Isaac et al. (Public Choice 1984) are attributable to error.
- Goeree et al. (JPubE 2002) find in their own experiment (not using virtual players) both a positive and significant effect on coefficients that correspond to (pure) altruism and decision error (confusion); no point estimate is given,
- Fischbacher & Gaechter (manuscript 2004) find in their own experiment (not using virtual players) that “at most 17.5% “ are contributed by confused subjects; they also argue that none of their subjects exhibits altruism or warm-glow (no subject stated they would contribute if other group members would not). In Fischbacher & Gaechter’s view, all non-confused subjects are “conditional cooperators”
- Summary: every study that looks for confusion finds that it plays a significant role in observed contributions.

- The virtual-player method has three (four, five) important features:
  - o Virtual players (that are preprogrammed to execute decisions that are made by human players in otherwise identical treatments)
  - o Split-sample design (where each participant is randomly assigned to play with humans or (human condition) with virtual players (computer condition))
  - o A procedure that ensures that human participants understand how the non-human, virtual players behave.
  - o Random assignment of subjects to the human condition or the computer condition – important assumption here that subjects are drawn from the same population.
  - o “Twins” in multiple-round public goods games where the group contributions are announced after each round, so that history starts to play a role ...

- Some graphs:



*Figure 10.1* GHL application, comparison of all-human and virtual-player contributions.

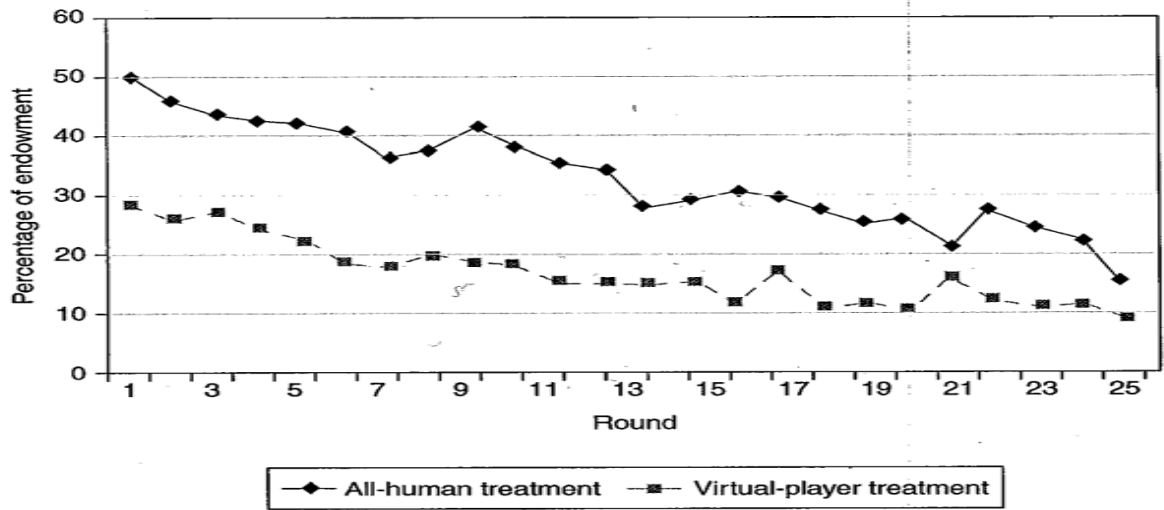


Figure 10.2 Ferraro and Vossler (2005) experiment, mean contributions.