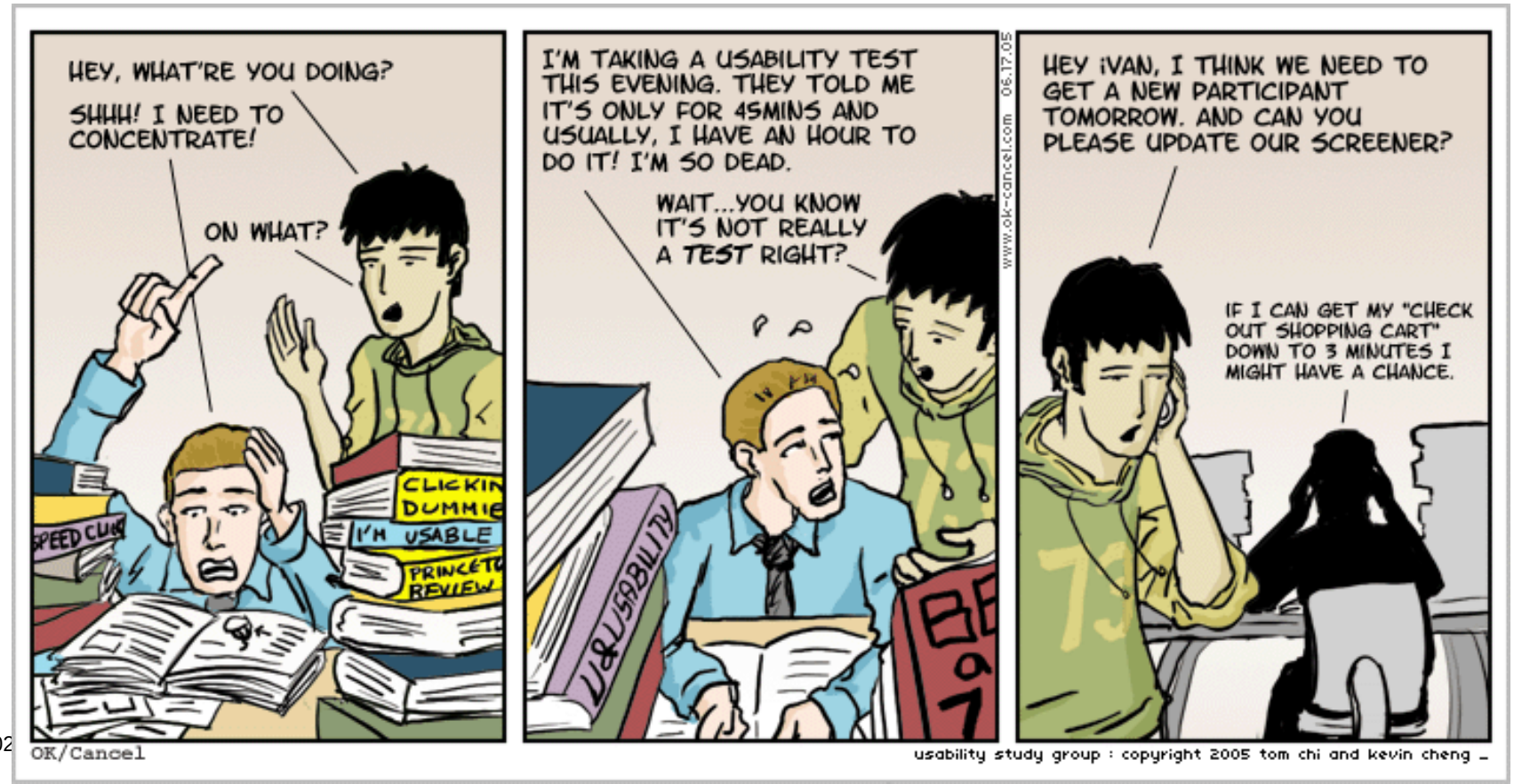
Research Methods
User Evaluation Experiments

Edwin Blake
edwin@cs.uct.ac.za

Purpose

Show you when you should use user experiments

Show you how to design experiments that involve users



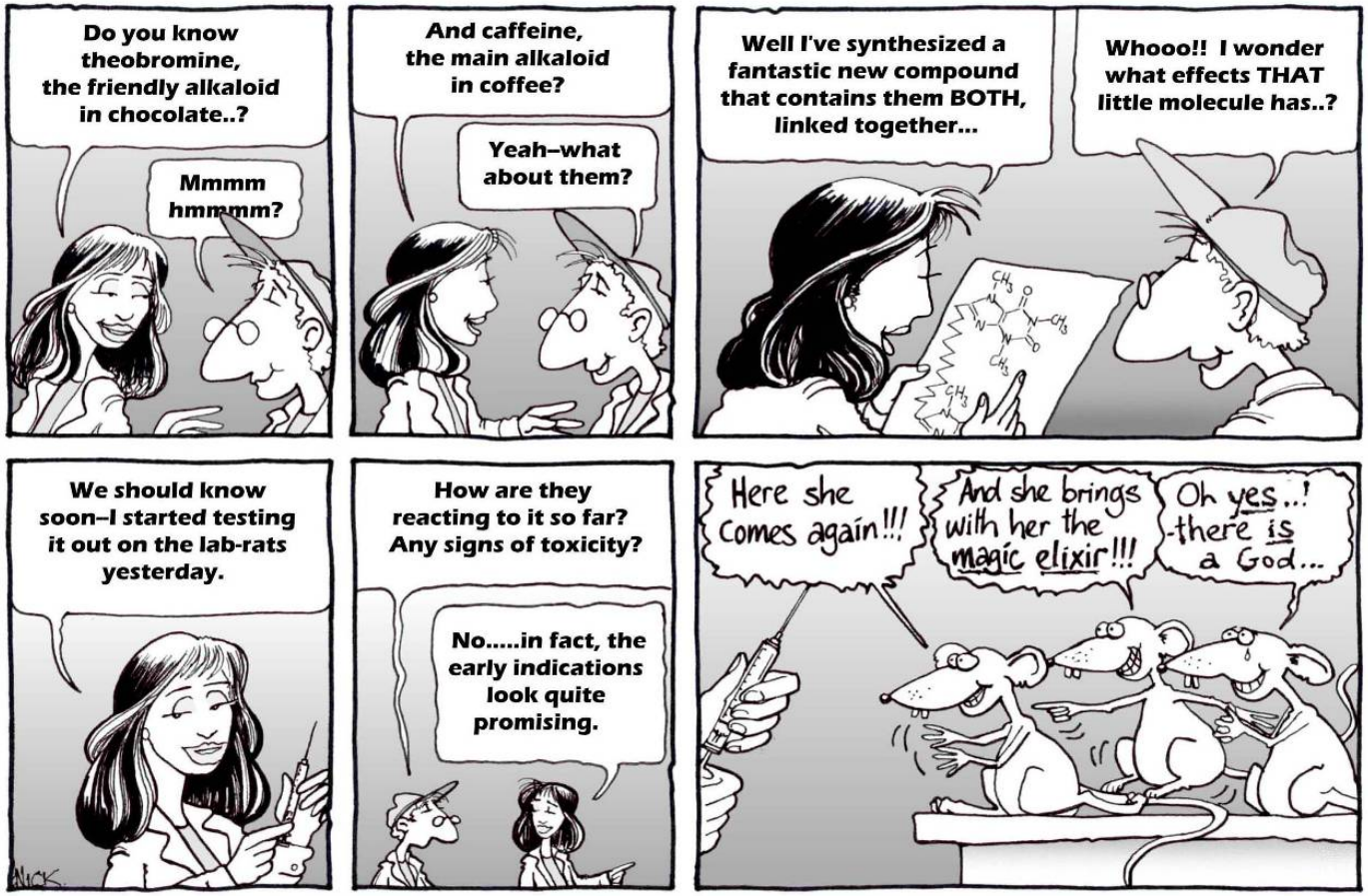
Outline

Overview

Designing a Study

Running a Study

Measurement



Why do User Evaluation Experiments?

To show how good your system or technique is

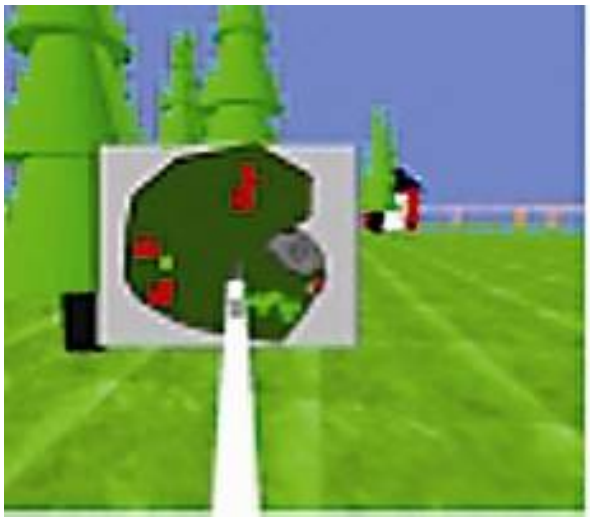
- Example: “A methodology for the evaluation of travel techniques for immersive virtual environments” (Bowman et al, 2001)
- Do experiments in a testbed

To test a theory

- Example: San VE: the addition of environmental sound to a VE increases Presence

To improve applications

- In terms of usability
- If it is supposed to be better for users: prove it!



Steps in Experimentation

Precise and unambiguous *hypothesis* to confirm or refute

Specified *experimental system* which is modified systematically (depending on the type of study)

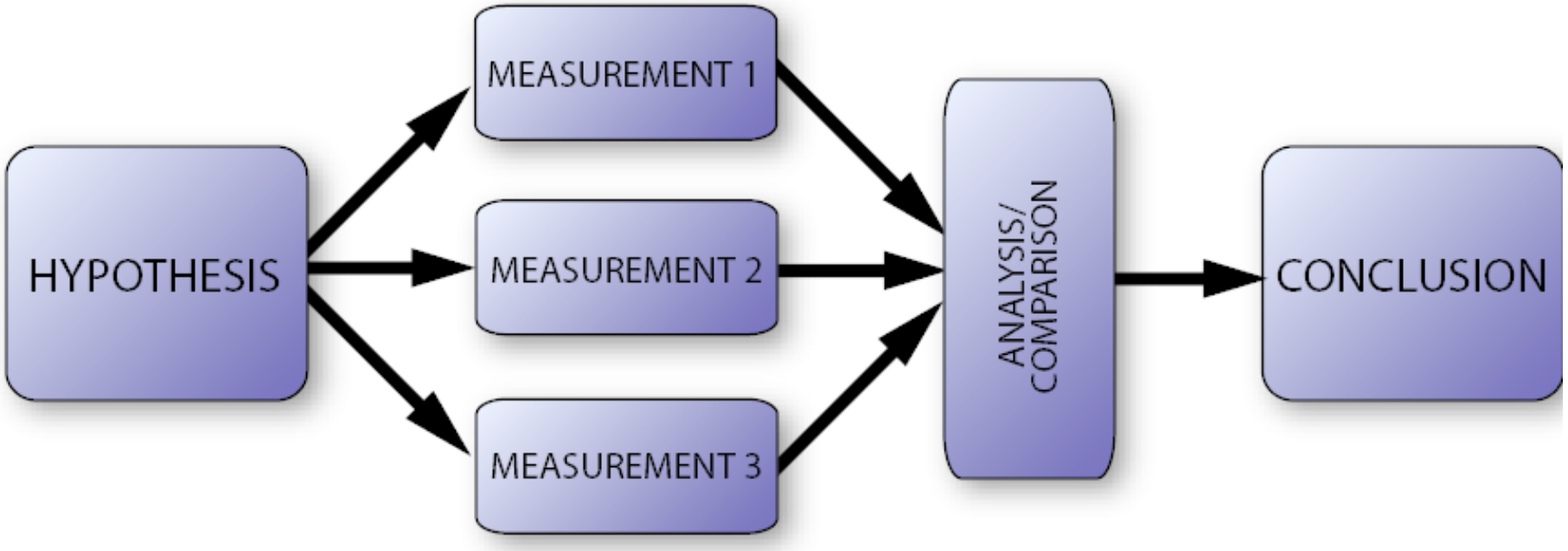
Use of *controls* to ensure experiment tests hypothesis

Measurement of results

Analysis of measured data

Report of procedures and results so experiment is repeatable

How a study runs through time…



The golden rules

You must know two things before you start:

- What am I trying to show?
- How am I going to analyze this data?
- “Front, back, then middle” strategy

Your constant ‘sanity check’

- Think about what you want to conclude — can you support it with the evidence you plan to collect?

Outline

Overview

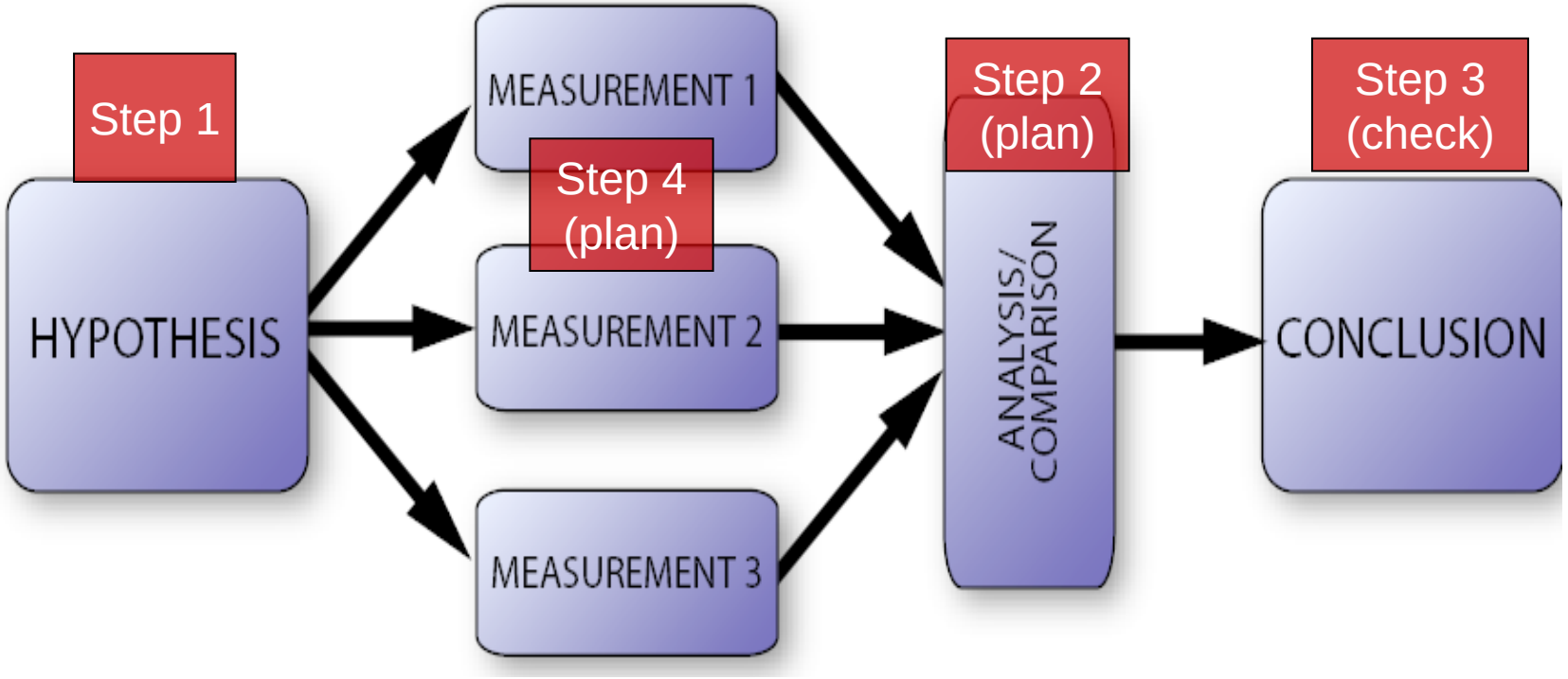
Designing a Study

- Reliability and Validity of Design
- Types of Study
- Causality
- Sampling

Running a Study

Measurement

How you design it…



Iterate through these steps a number of times…

A perfect design allows no rival explanations for the effect you have found!

Research Design

- Practical plan of how to perform research
  - What tests will be performed
  - Which measurements/scales will be used
  - The design is a scaffold on which the entire study hangs.
- Faulty design = false conclusions
  - Performance of travel techniques in VR means speed and accuracy — No (not only)
  - Other factors, such as spatial awareness, ease of learning, ease of use, presence and user comfort, might be important — YES
  - Once data has been collected, design cannot be changed!
- Carefully evaluate design before performing the study

Terminology

Variable: a thing that is measured

- Independent variable (IV): manipulated by the researcher
- Dependent variable (DV): changes as a result of the manipulation
- Extraneous variable: any variable other than the IV that influences the DV

Hypothesis: falsifiable statement of the relationship between 2 or more variables

Operationalization: to turn concept into variable(s)

Intervention: the “thing” done to the subjects to check for changes in the DV

- In a study to see how nutrition affects IQ, the feeding program is the intervention

Reliability and Validity

Reliability:

- Consistency across time or with equivalent tests / designs

Validity:

- Does the design / test measure what it says it does, even when applied to different people, settings or times
- A design has validity when it has *the capacity to study what it aims to study*

Ensuring Design Validity

Validity is evaluated before the design is run

Validity of designs is improved by eliminating rival hypotheses

Identify as many possible rival hypotheses as you can, and change the design to reduce their impact

- E.g. If you suspect time spent video gaming might be a rival hypothesis, record hours for each subject

Parts of Design Validity

Validity of designs come in 2 parts:

Internal validity

- can the design sustain the conclusions?

External validity

- can the conclusions be generalized to the population effectively?
- Campbell (1979) came up with a list of threats to validity

Internal Validity

Each design is only capable of supporting certain types of conclusions

Says nothing about if the results can be applied to the real world

Generally, more control = higher internal validity

External Validity

Can the findings of the study be generalized?

- Do they speak only of our sample, or of a wider group?

Says nothing about the truth of the result being generalized

Generally, bigger samples with valid measures = better external validity

Threats to Internal Validity I

1. Co-varying events
   - Another, unseen variable might be causing the effect we are seeing
2. Maturation
   - Changes over time can be caused by a natural learning process
3. Reactivity (testing effect)
   - People realize that they are being studied, and respond they way they think is appropriate

Threats to Internal Validity II

4. Instrument decay
   - Instruments with low reliability lead to inaccurate findings/missing phenomena
5. Regression to the mean
   - Studying extreme scores can lead to inflated differences, which would not occur in moderate scorers
6. Subject mortality
   - If subjects drop out, it creates a bias to those who didn't

Threats to External Validity

1. Subject selection
   - Selecting a sample which does not represent the population well will prevent generalization
2. Operationalization of the variables
   - We take a concept (wide scope) and make it a variable(s) (narrow scope) – will we find the same results with a different operationalization of the same concept?
   - Bad measures of variables

Some design decisions

Unit of analysis: what are you talking about?

- Groups? Individuals? Ideologies?

Time:

- is the study *longitudinal* (follow people over a long time)
- or *cross-sectional* (a snapshot in time)?

These decisions affect the conclusions that can be drawn!

- Must be carefully chosen

Types of Study

Descriptive: paints a picture of how things are right now.

Relational: investigates the relationship between two or more variables without manipulation.

Experimental: investigates how one or more variables *cause* another variable.

Descriptive Designs

Descriptive research aims to simply describe the population by looking at a sample

Very simple design

- Everyone is measured on the same variables.
- All variables are independent
- Often a large number of variables
- No hypothesis, just questions

Descriptive Design Example

Manninen and Pirkola (1997) Comparative Classification of Multi-User Virtual Worlds

- Surveyed 15 worlds:
  - Text (eg IRC)
  - 2D (eg Ultima On-Line)
  - 3D (eg DIVE, Quake)
- Displayed and compared critical features

Table 1. Classification matrix of surveyed multi-user virtual worlds.

Technical Category / Application Area	TEXT	2D	3D
Creative/Constructivist			AlphaWorld
Educational/Training	APECKS		
Game	Dark Chambers BatMUD	Ultima On-line The Realm	Meridian59 Quake WarBirds
Social/Chat	IRC	Worlds Away	Worlds Chat Blaxxun Onlive Traveler
Work/Research			DIVE

Table 3. Overall Technology Comparison.

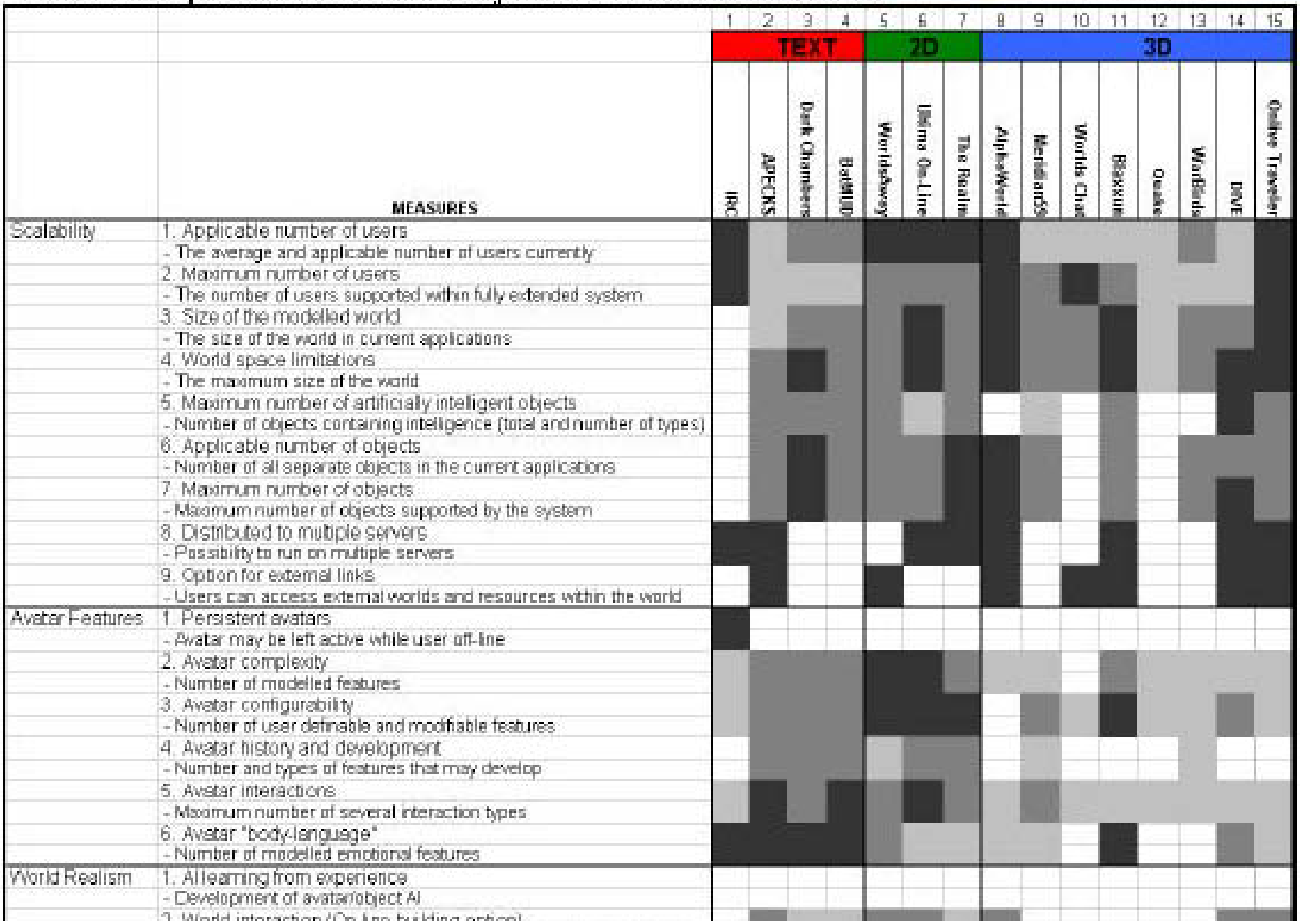
	TEXT	2D	3D
Scalability			
Avatar Features			
World Realism			
User Interface			
Communication			

Criteria for Comparison

Criteria:

- Scalability
  - E.g. Maximum number of users, maximum size of world
- Avatar Features
  - E.g. Number of configurable options, modelled emotional features
- World Realism
  - E.g. Number of objects that can be interacted with, physical laws modelled
- User Interface
  - E.g. Navigation and control, sound support
- Communication
  - E.g. Audio and visual types and availability

Table 2. Competence matrix of surveyed multi-user virtual worlds.



Relational Designs

- Relational research aims to find relationship between variables in the population
  - By looking at relationships in a sample
- Simple design
  - No manipulations, observe relationships as they occur naturally
  - IV/DV decided on basis of hypotheses

Relational Example: Walking in Place

Hypothesis that the correlation between proprioception and sensory data is an important factor in maintaining presence

Navigating a VE by using a ‘mouse’ breaks this match

Compared a ‘walking in place’ method with a point-and-click (‘flying’) method

- Slater M, Usoh M, Steed A (1995) Taking steps: the influence of a walking technique on presence in virtual reality. ACM Trans Comput-Hum Interact 2:201-219.

Relational Example: Experimental Environments

16 subjects
- 8 in a walking-in-place group
- 8 in a point-and-click group

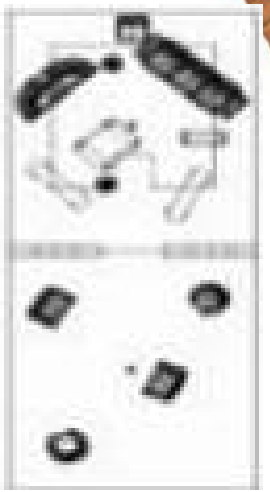
Task was to take an object to the chair

A between groups design

Each subject had a virtual body that they could see when looking down

HMD display with Virtual Research Flight Helmet 360 × 240 with 75 deg. horizontal FOV

Polhemus sensors for head tracking



Relational Example: Some Measured Variables

Presence Questionnaire

- Three questions
  - The sense of ‘being there’
  - Whether the environment the virtual room was remembered as somewhere visited rather than only images seen
  - The extent to which the pit room became the dominant reality, and that the real lab was forgotten
- Each measured on a 7-point scale
  - (1) Not at all
  - (7) Very much so
- Final score is x = the number of ‘6’ or ‘7’ answers given (x = 0,1,2,3)

Path taken to the chair

Extent to which they ‘associated’ with their Virtual Body

Relational Example: Results

For the ‘walkers’ – greater their association with their VB the greater their presence score

For the ‘pointers’ — no correlation

If they associated with their VB then the ‘walkers’ reported higher presence than the ‘pointers’

Path across the precipice associated with lower reported presence

Relational Experiment Extension

Walking > Walking-in-Place > Flying, in VEs

- Repeat of Slater experiment showing subjects in immersive VE have higher presence when walk-in-place instead of push-button-fly

Added real walking

- (RW) Really walking (UNC wide-area tracker)
- (WIP) Walking in place
- (PAC) Point and click (‘flying’)

Examined in terms of ease of locomotion (simplicity, straightforwardness, naturalness) and presence

Between groups experiment (11 subjects per group)

Results showed (RW,WIP) >> PAC

No significant difference between RW and WIP

- Usoh M, Arthur K, Whitton MC, Bastos R, Steed A, Slater M, Brooks Jnr F.P (1999) Walking > walking-in-place > flying, in virtual environments. (SIGGRAPH ’99), 359-364

Further Studies

Meehan/Insko (2002) exposed 10, 52 and 33 subjects in 3 different studies.

Heart rate increase when in the pit room.

Static haptics further significantly increased heart rate.

Heart rate correlated with subjective self-report of presence level.

Meehan M, et al (2002) Physiological measures of presence in stressful virtual environments. ACM Transactions on Graphics 21:645-652

Causality

Limits of relational research

- Relational research only tells us that there is a relationship
- Cannot tell if the relationship is causal

A cause always co-occurs with the effect

- (Drinking a lot) and (feeling light headed) co-occur because drinking causes drunkenness

But: 2 non causally related events can also co-occur because they are both effects of the “third variable”

- People who have sore throats usually also get fevers — sore throats cause fever ? (3rd variable is the flu virus)

Causality II

Knowing if a relationship is causal is important

- Interventions only guaranteed to work if the relationship is causal

Causality:

A *causes* B if and only if:

A exists then B exists and

A does not exist, then B does not exist

Relational studies test only “if A exists then B exists”!!

Experimental designs

At least 2 groups:

- Experimental “if A exists, B exists”
- Control “if A does not exist, B does not exist”

Same IVs (cause) and DVs (effect) are measured on all groups

- Manipulate the existence of the IV
- All other things are kept constant
  - The only difference between experimental and control groups is the IV

Experimental Design Example

Brown et al. (2003) The Effects of Mediation in a Storytelling Virtual Environment

- Explored effects of visual and audio mediation in a storytelling VE
- 4 versions of the VE, varying in audio and visual mediations
- Took presence, enjoyment and story involvement

Sampling

Population*:* All possible observations of a particular variable

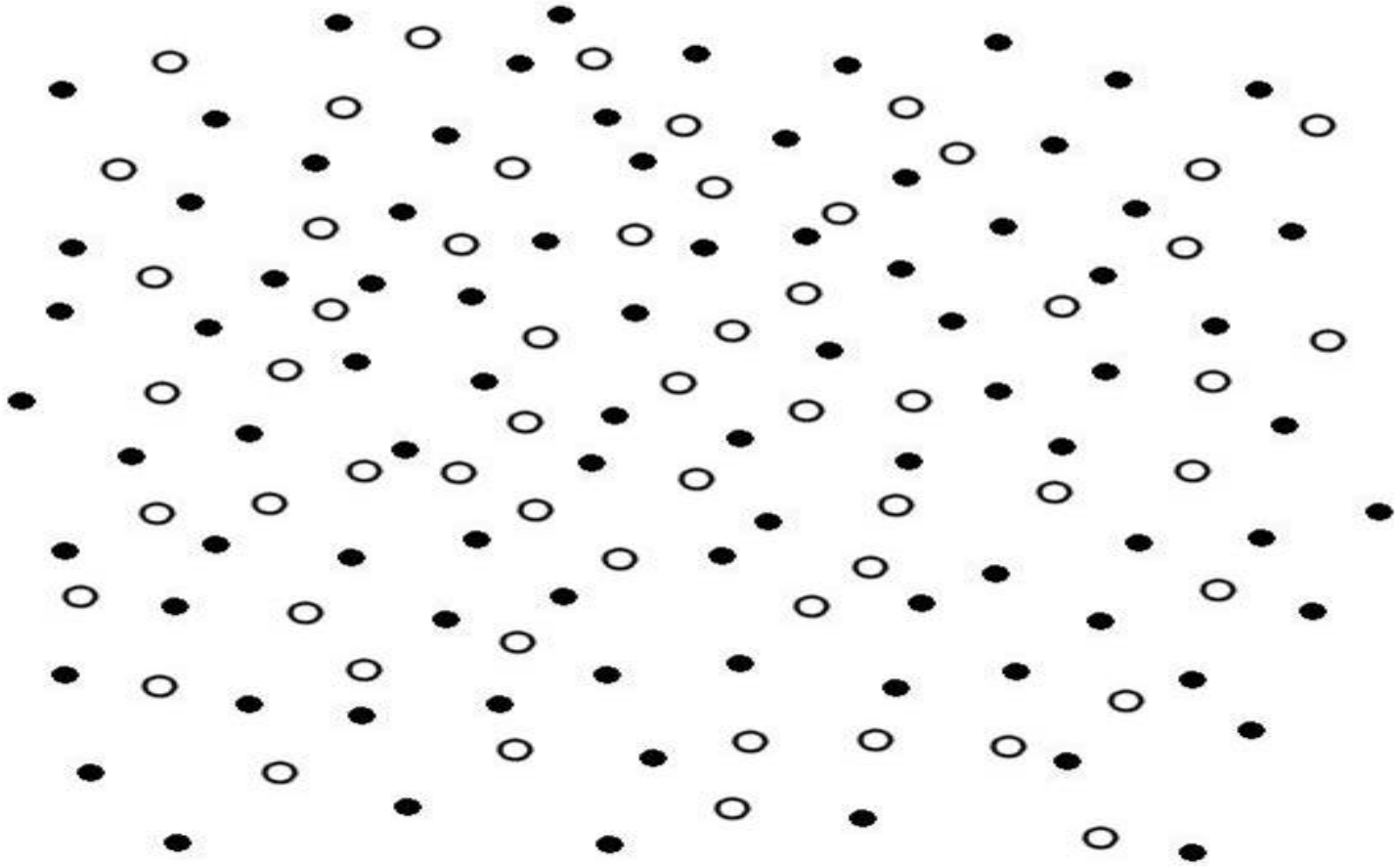
- Cannot experiment on populations!

Sample: subset of a population *selected to estimate the behaviour of the population*

- We must know what the *actual* parent population is, otherwise we draw false conclusions

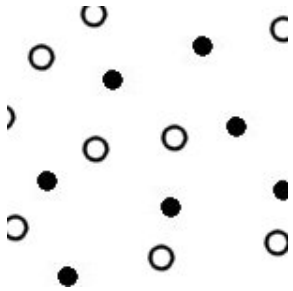
Careful sampling will ensure that results can be generalized

Sample Example



- What percentage of these dots are black and which are white?

Subsample example



In this small sub section (the sample), there are: 53% black, 47% white

In the big picture, there are in fact: 59% black, 41% white

- Even though we used a sample, we approximated the big picture well!!

Who and how many to use — your sample

Requirement 1: Represent your population

- Representation of your population of interest
- If using a control group, use an ecologically valid one
- Different populations will give different results, so beware of spurious differences

Requirement 2: Satisfy the needs of your statistics

- More observations are better (stats will tell you how many)
- Go for as much variance as you can
- Avoid outliers (super-extremes)

Deciding Who to Choose

Two basic sampling methods:

- **Probability sampling**: Each member of the population has a certain probability of being selected
- **Non-probability sampling**: Members selected by other means than mathematical rules (e.g. convenience, access)
  - Problematic for most statistical analyses
  - Suited for qualitative research

Outline

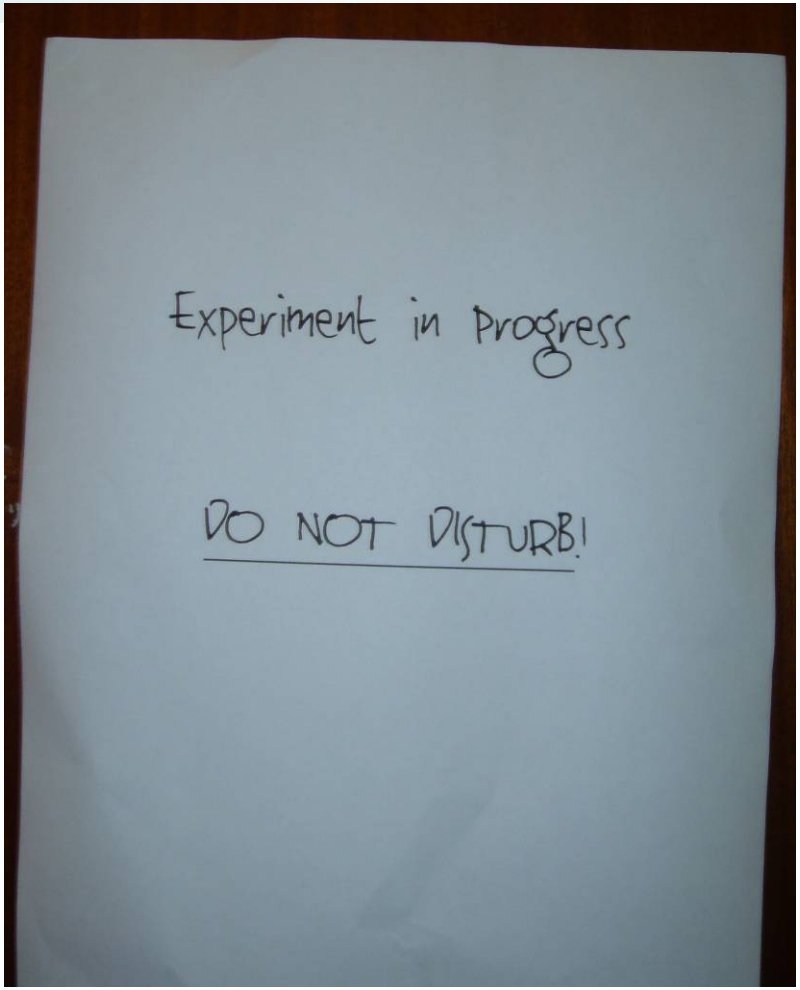
Overview

Designing a Study

Running a Study

- Recruiting Subjects
- The Experimental Space
- Procedure

Measurement



Recruiting Subjects

You wouldn’t throw a party without inviting guests — it doesn’t make sense!

A study can’t happen unless you have subjects

Sometimes getting the sample you want requires imagination, ingenuity and contacts

Recruiting Subjects
Children & Young Adults

University students (easiest)

- Visit lecture
- Aim for a good campus cross-section

School children (more work)

- Obtain ethics approval from your faculty
- Contact school
- Consent forms
- Negotiate whether the study will take place at school, during school time or as an after school activity
- Ensure that a teacher will be present

Recruiting Subjects
Specific Samples

Specialists

- Approach companies, NGO’,s etc.

Certain groups, for example

- HIV positive individuals
- Rural health-care givers
- Disabled
- Elderly

First steps:

- find out where they hang out and who their gatekeepers are?

Recruiting Subjects Online

If your study can be automated it can be distributed online

This can draw huge samples

Must use huge samples to protect against lack of subject control

Recruiting Subjects
The Lure

Intangible

- An interesting activity
- Making a contribution to science and technology

Tangible Rewards

- Cash (sometimes required by ethics board)
- A prize
  - e.g. All participants will be entered into a draw to win an iPod

    Cost is R2500, for a sample of 980

    versus

    980 subjects @R20 each = R19 600

The Experimental Space Maximise Subject Throughput

Run multiple sessions in the experimental space per day

For single-user applications

- Allow multiple subjects to experience interactive application simultaneously
- While ensuring that each subject is not distracted by the other subjects
- Control equipment

The Experimental Space
Maximise Subject Throughput

Pipelining:

- If study takes the form of
  <user experiences application> then
  <user completes questionnaire>,
  use different venues so that these activities can run simultaneously

Get Help:

- Make use of more than one experimenter

Control in experimentation

Keep even apparently insignificant aspects of procedure constant between observations

- Room conditions, length of experiment, your interaction with the subject
- Record those that you cannot keep constant
- The best way to ensure that there are no differences between groups is to assign subjects carefully
  - It is impossible to manually create even groups
  - Use Random assignment

Procedure

You want everyone in the same experimental condition to be treated identically

You don't want to think about 'how' on the day of the experiment – you want to go into 'execute' mode

So plan the experimental procedure to the last detail

- Wording
- Sequence
- Logistics

Procedure
Introduction to the Study

Tell the subjects what they need to know and make them want to give their full attention

BUT

Don’t give away the game AND

Don’t bias them or try to get them on your side

- You want to dodge the experimenter effect

- The thin line between informed consent and experimenter bias

Procedure
Assigning Groups

Random

Try to eliminate subjects' awareness of different experimental conditions

- Can influence subjects response
- e.g. high school study:
  - 1 class, some children are required to stay in class and read a story text (control group) while others are taken out of class to experience an interactive storytelling application (experimental group).
  - How might this have affected the two groups' response to the study?

Procedure Training

Provide users a chance to become familiar with the application they will be required to use

Provide a version in which they can train

- e.g. A study involving VR application should provide a virtual environment in which the user can practice the use of the controls and navigation

Procedure
The Interactive Experience

If processing subjects simultaneously be aware of timing

Helpful to have an experimenter present to control the room and possibly catch observations that may otherwise go unnoticed

Fly-on-the-wall and silent footed

If a subject interacts with the experimenter — note it, you may want to exclude their data

What about video recording?

Procedure
Collecting Data

- Retrospective questionnaire
  - Existing questionnaires which have been proven statistically sound
  - New questionnaires created for the study
- Observation during users' interaction with application
  - Naturalistic observations: can't record everything and can't record ad-hoc

Procedure Debriefing

Say your thankyou’s

Handle any questions, mention possible side-effects

Give reward (cash/prize)

Non-disclosure agreement

Further contact

How to stuff up your own work

Research is a specific social situation — you *can* cause subjects to behave in a particular way

- **Experimenter effect**: Your race, gender etc. suggests to people how they "should" behave
- **Demand Characteristics**: The research setting or measures can "suggest" to the subjects how to behave

- You can eliminate this by careful control or observation

Outline

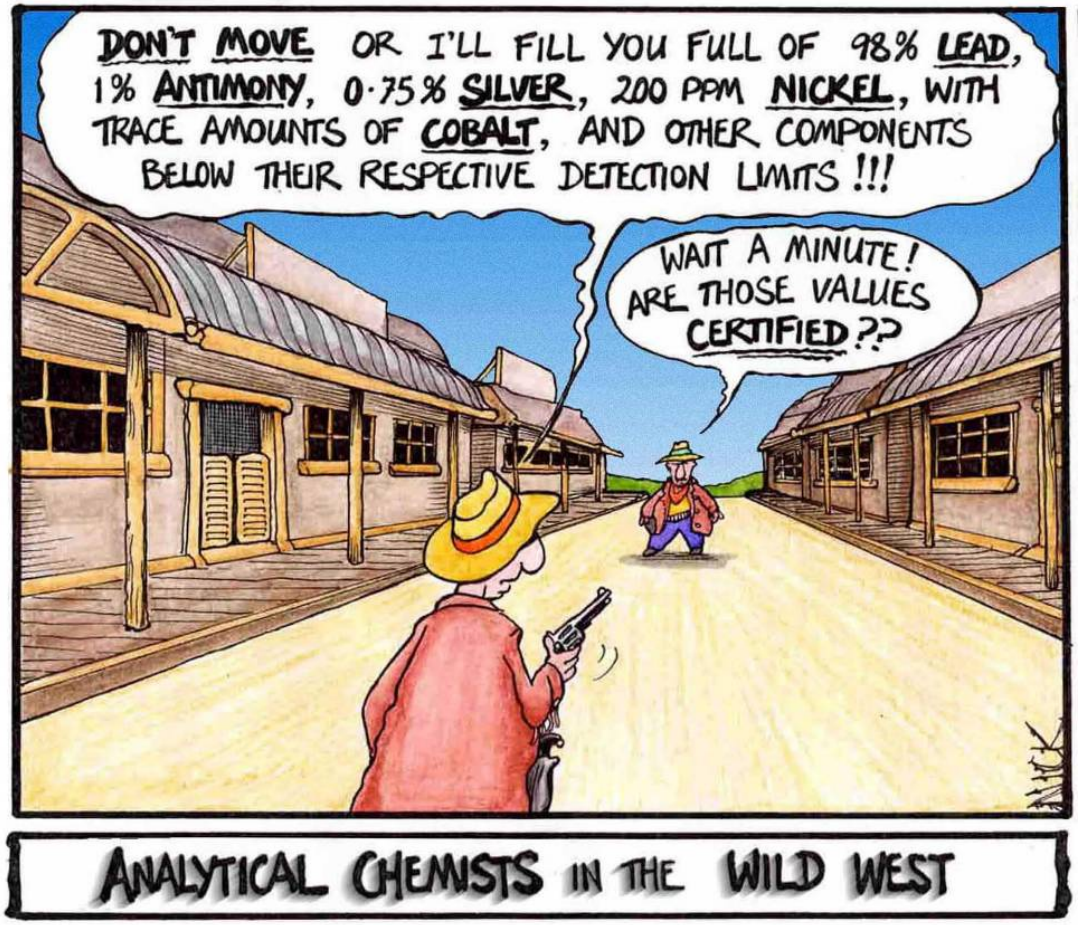
Overview

Designing a Study

Running a Study

Measurement

- Valid and Reliable
- Using Questionnaires



Measuring Results

Various ways of measurement:

- Most common – questionnaire or scale
- Many others
  - Physical measures, e.g. heart rate for anxiety
  - System measures, e.g. logging of user key strokes
  - Interview, shows subjective responses
  - Observation and talk aloud

Scales

Scales must be **valid** and **reliable**

- The more of each of these properties, the better the scale
- If at all possible, use an established questionnaire when asking questions of users

Reliability: stability of a measure over time

- Low reliability implies that other variables (“noise variables”) are being measured also

Validity: the degree to which a scale measures what it is supposed to

Validity in scales

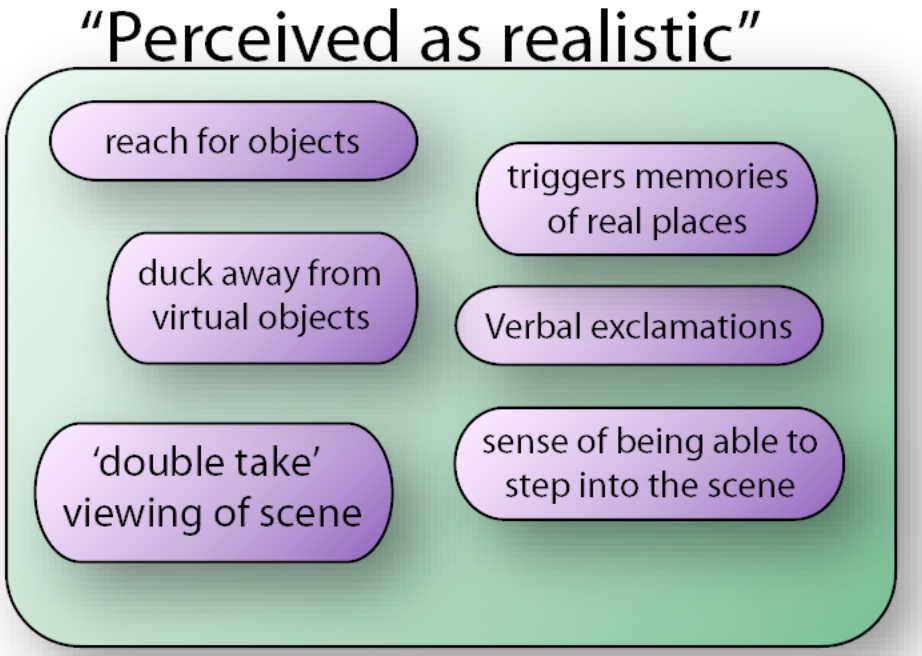
Validity is subdivided into many types — most important:

- **Criterion Related Validity**: Degree to which new scale matches established scale
  - By comparing to a scale known to be valid, you can be sure yours is valid
  - Run a sub-study in which you give the subjects your scale and the established one
- **Construct Validity**: Does the scale actually measure the construct?
  - Eg: measuring cranial circumference to measure intelligence
  - Closely tied into the theory of the construct
  - Most difficult to achieve, most important
  - Measures lacking in construct validity are almost useless

Behaviour Domains

All constructs you want to measure have a ***behaviour domain***

To measure it, all we have to do is sample it correctly!



Questionnaires

A questionnaire is valid to the extent that inferences made from it are appropriate, meaningful and useful

Construct Validity:

- A scale measures what it claims to (all the items belong to the same behaviour domain)

Concurrent/Content Validity:

- A scale's items measure the same factor/phenomenon

Using psychometric techniques we can:

- Evaluate the overall soundness of a questionnaire
- Identify specific problematic items

Some Questions from a Presence Questionnaire

Likert type scale

1. To what extent did you have a sense of being in place X?
   Not at all 1 2 3 4 5 6 7 Very much so

2. To what extent were there times during the experience when X became the ‘reality’ for you, and you almost forgot about the ‘real world’ of the laboratory in which the whole experience was really taking place?
   Never 1 2 3 4 5 6 7 Almost all the time

3. When you think back about your experience, do you think of the virtual X more as *images that you saw*, or more as *somewhere that you visited ?*
   Only as images that I saw 1 2 3 4 5 6 7 Somewhere that I visited

Reminder — Design Validity

Check Internal validity:

- Does the design allow me to make the conclusion I want to?
- ‘scientifically ok’

Check External validity:

- Does the design allow me to say something about my chosen population?
- ‘real world relevance ok’

Beware: These two give a constant sum!

- more control = less real world relevance

Conclusion

Project Management

- Research is uncertain, so plan carefully using network analysis, gantt charts and risk management
- Be sure to update your planning during the project

Experimental Computer Science

- CS deals with Information Artefacts, unlike most Sciences
- All the more reason to apply experimental methods

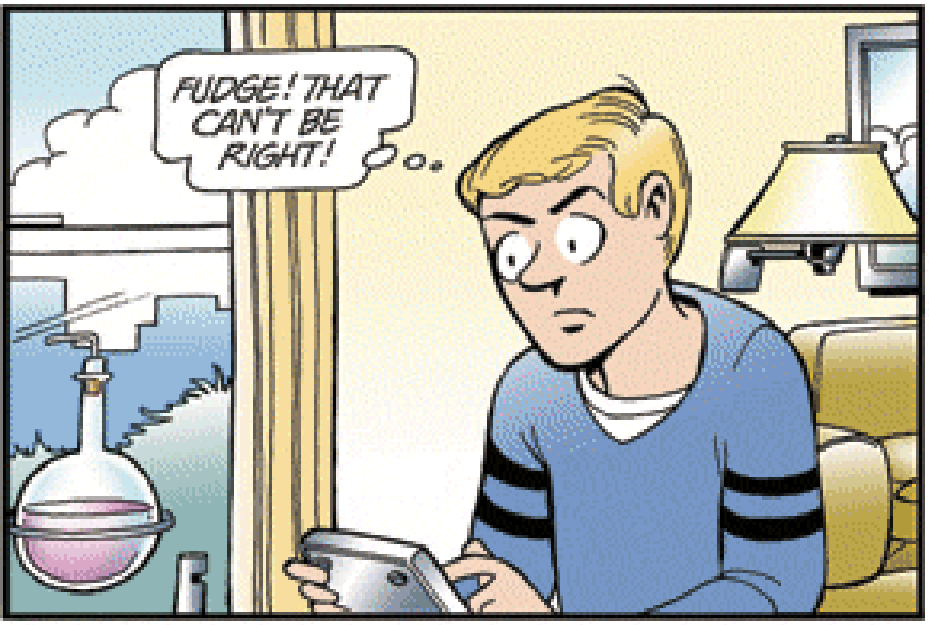
User Evaluation

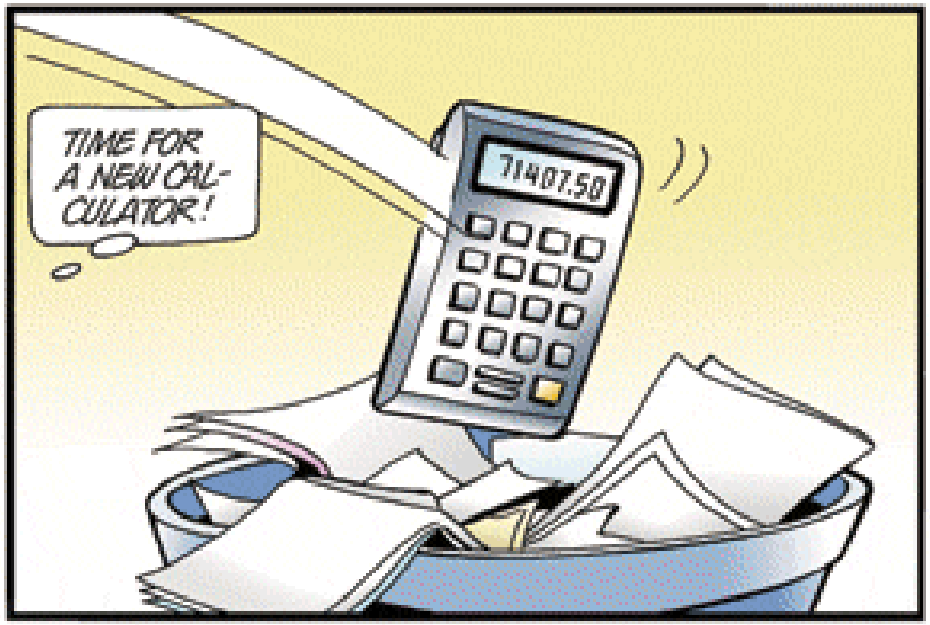
- Pay careful attention to the design of the experiment and the validity of measures

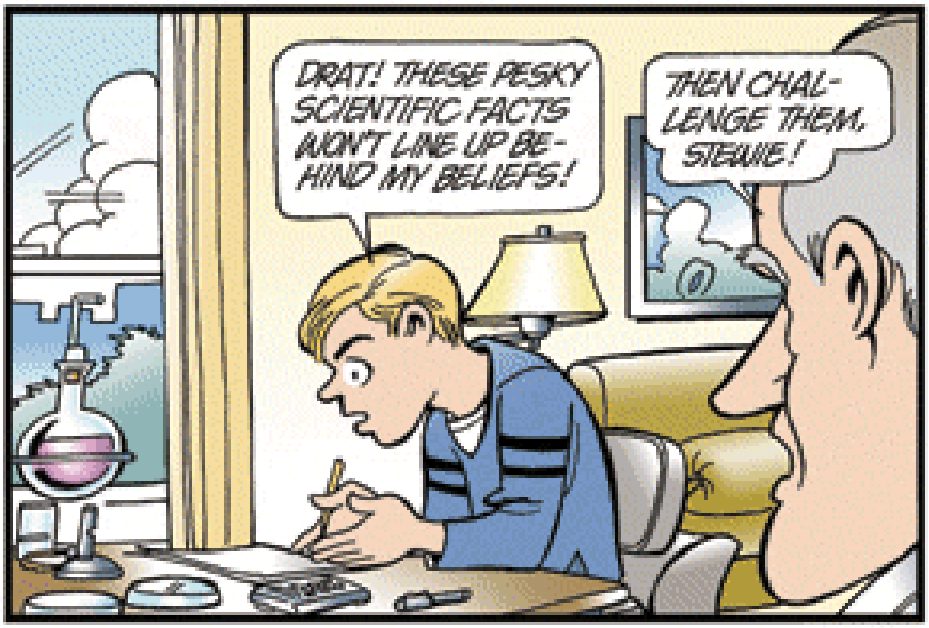
Next

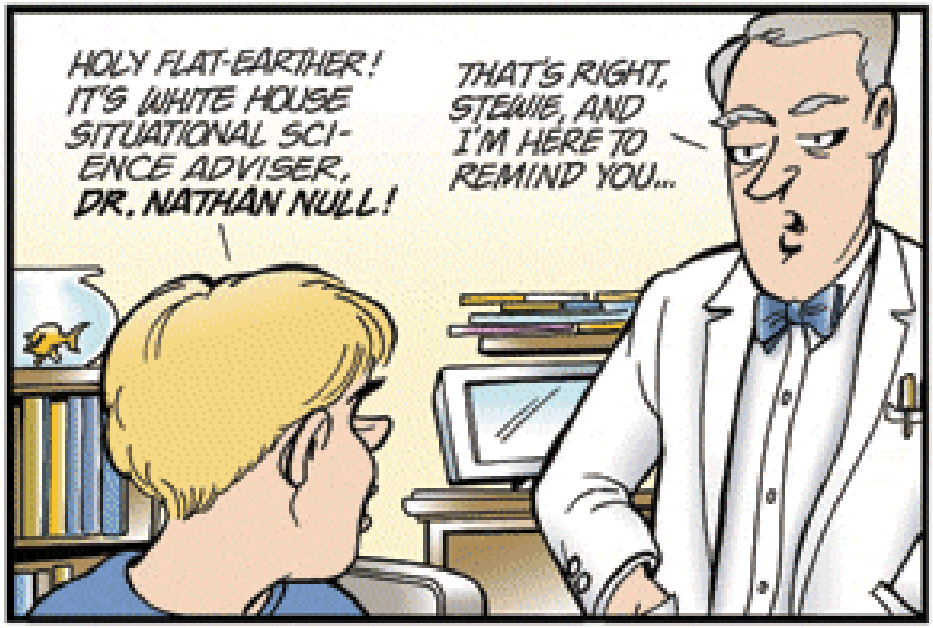
Technical Writing

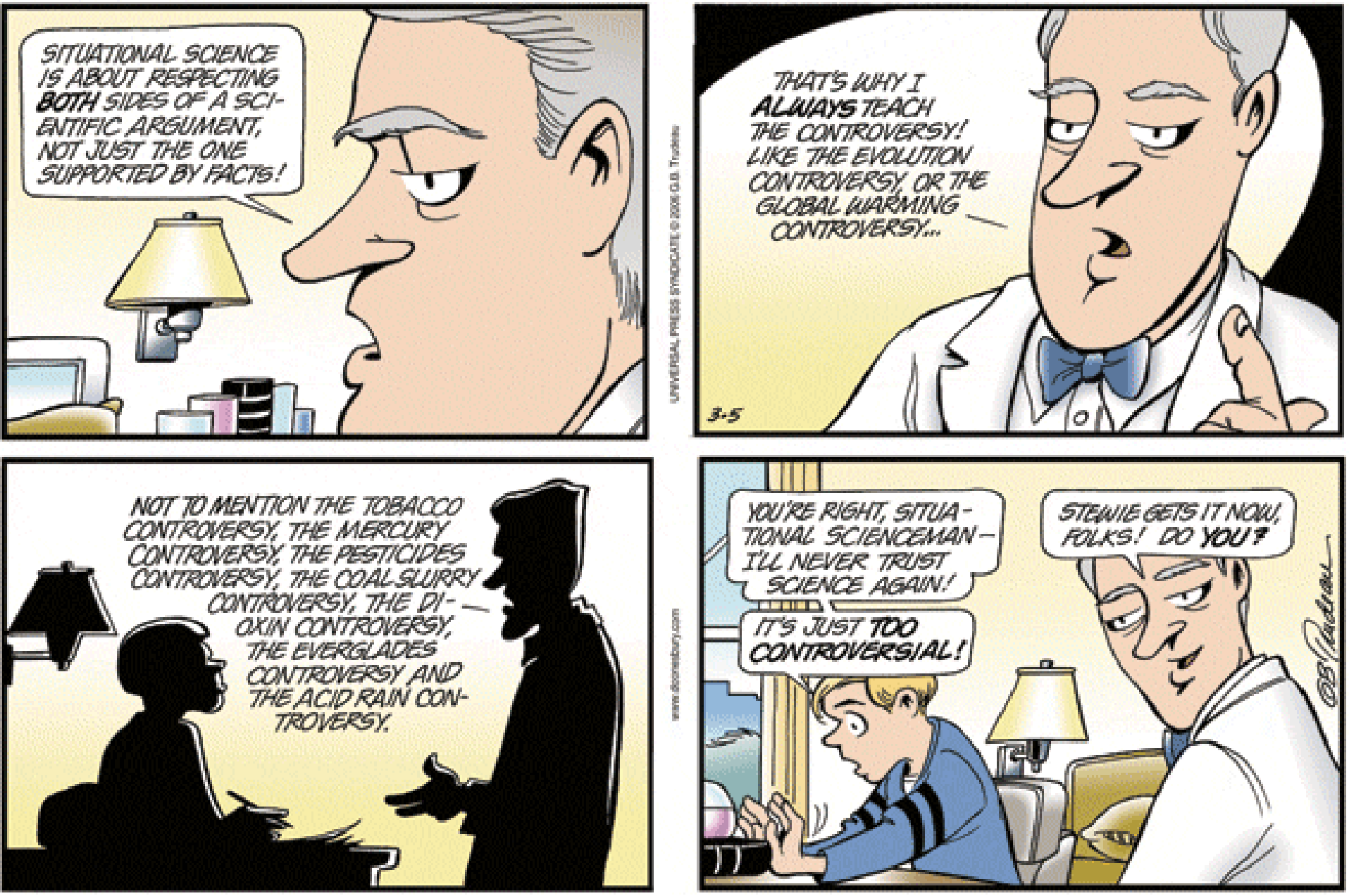
- Writing is a skill that requires practice. Don’t leave all your writing to the end of a research project


FUDGE! THAT CAN'T BE RIGHT!


TIME FOR A NEW CALCULATOR!
71407.58


DRAT! THESE PESKY SCIENTIFIC FACTS WON'T LINE UP BEHIND MY BELIEFS!
THEN CHALLENGE THEM, STEWIE!


HOLY FLAT-EARTHER! IT'S WHITE HOUSE SITUATIONAL SCIENCE ADVISER, DR. NATHAN NULL!
THAT'S RIGHT, STEWIE, AND I'M HERE TO REMIND YOU...


SITUATIONAL SCIENCE IS ABOUT RESPECTING BOTH SIDES OF A SCIENTIFIC ARGUMENT, NOT JUST THE ONE SUPPORTED BY FACTS!
THAT'S WHY I ALWAYS TEACH THE CONTROVERSY! LIKE THE EVOLUTION CONTROVERSY, OR THE GLOBAL WARMING CONTROVERSY...
UNIVERSAL PRESS SYNDICATE © 2009 G.B. Trudeau
3-5
NOT TO MENTION THE TOBACCO CONTROVERSY, THE MERCURY CONTROVERSY, THE PESTICIDES CONTROVERSY, THE COAL SLURRY CONTROVERSY, THE DIOXIN CONTROVERSY, THE EVERGLADES CONTROVERSY AND THE ACID RAIN CONTROVERSY.
www.doonesbury.com
YOU'RE RIGHT, SITUATIONAL SCIENCEMAN— I'LL NEVER TRUST SCIENCE AGAIN!
IT'S JUST TOO CONTROVERSIAL!
STEWIE GETS IT NOW, FOLKS! DO YOU?