Experimental Design Strategies

Rob MacCoun

CSLS Miniseries in Empirical Research Methods, 5 Nov 2010

Roadmap

- If experiments are the answer, what is the question?
- Counterfactuals
- Internal validity
- External validity; mundane vs. exp realism
- Construct validity
- Statistical conclusion validity

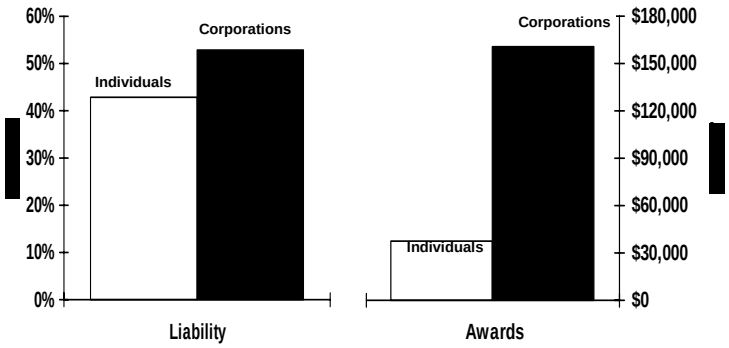
There are various technical appendix slides at the end of the handout.

What experiments offer

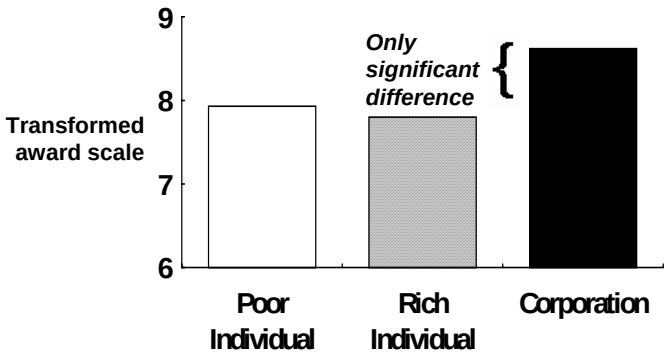
- Great for:
  - *Causal inference (why are A and B correlated?)*
  - *Theory testing*
  - *Low-risk test of interventions that haven't been adopted in the real world (e.g., change of law or new procedure)*
- Bad when:
  - *Goal is point estimation (forecasting, etc.)*
  - *External validity is more important than internal validity*
  - *Ethical, political, legal barriers*

Juries appear to treat corporations differently

(Chin & Peterson, 1985 archival analysis)



Jurors do treat corporations differently, but not because of wealth (MacCoun, 1996, mock jury experiment)



Flouting the Law

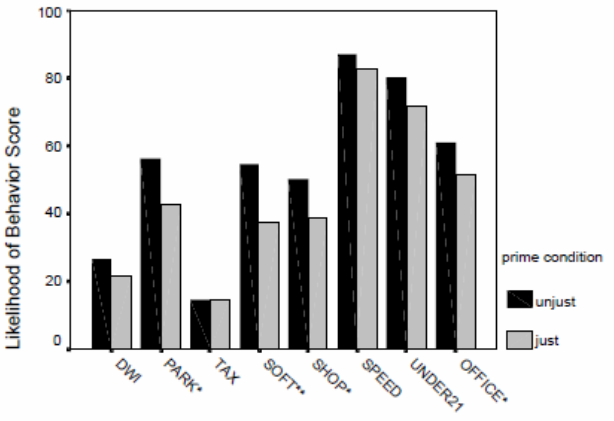
Janice Nadler

"What happens when a person's common-sense view of justice diverges from the sense of justice he or she sees enshrined in particular laws? In particular, does the perception of one particular law as unjust make an individual less likely to comply with unrelated laws?"

TABLE 1. CONTENT OF NEWSPAPER STORIES CONTAINING PRIMES

News Story	*General Emphasis (both versions)*	*Just Prime Version*	*Unjust Prime Version*
Civil Forfeiture	Purpose and application of (actual) laws permitting the government to seize property under certain circumstances	Emphasized the law enforcement benefits of civil forfeiture laws	Emphasized the civil liberties concerns surrounding civil forfeiture laws
Income Tax	Proposed legislation ostensibly pending before Congress that would affect the amount of income tax paid by middle class taxpayers	Emphasized positive effects of income tax paid by middle class people	Emphasized negative effects of income tax paid by middle class people
Landlord/ Tenant	Proposed legislation ostensibly pending before the state legislature that would permit landlords to conduct warrantless searches of tenants' apartments under certain circumstances	Emphasized importance of empowering landlords to evict drug-dealing tenants	Emphasized the civil liberties and privacy concerns in permitting searches of tenants' apartments

FIGURE 1. WILLINGNESS TO FLOUT AS A FUNCTION OF EXPOSURE TO UNJUST LAWS IN EXPERIMENT 1



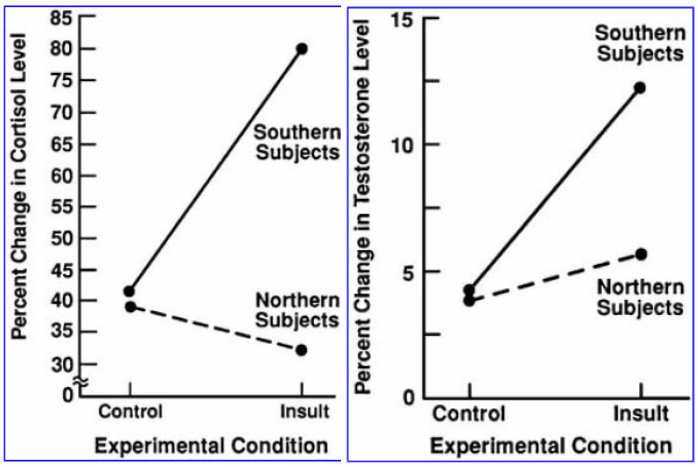
* p < .10

** p < .05

Evaluation of simulation experiments

- Realism
  - experimental vs. mundane realism
  - mundane realism is never complete
  - ultimately, a marketing issue
- Theory testing vs. describing the world
  - can't use simulations for descriptive stats
  - use the theory, not the data, to make predictions about the world
  - *a priori* theories about boundary conditions can be incorporated into theory and tested

Cohen, Nisbett, Bowdle, & Schwarz (1996): “Participants were University of Michigan students who grew up in the North or South. In 3 experiments, they were insulted by a confederate who bumped into the participant and called him an “asshole.”



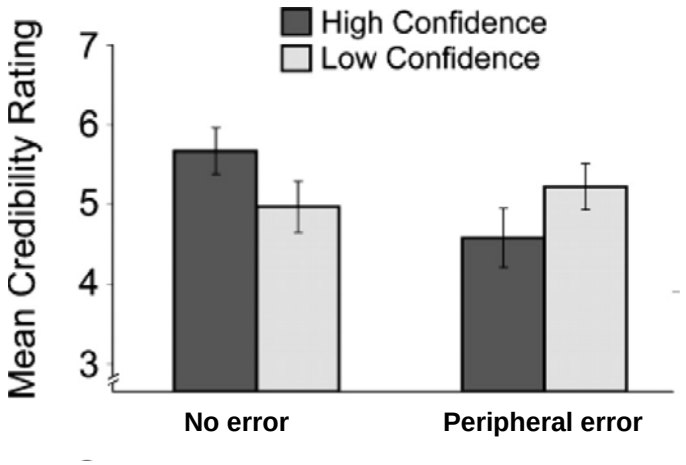
Two diverging theories

- **The ‘confidence heuristic’**
  - Highly confident advisors are presumed to be more accurate, knowledgeable, and credible, even when given feedback that demonstrates otherwise (Price & Stone, 2004).
- **The ‘calibration hypothesis’**
  - Advisors are perceived more credible if they express confidence only when warranted - highly confident but inaccurate advisors lose credibility (Tenney, MacCoun, Spellman, & Hastie, 2007).

Tenney, MacCoun, Spellman, & Hastie (2007, Psych Science)

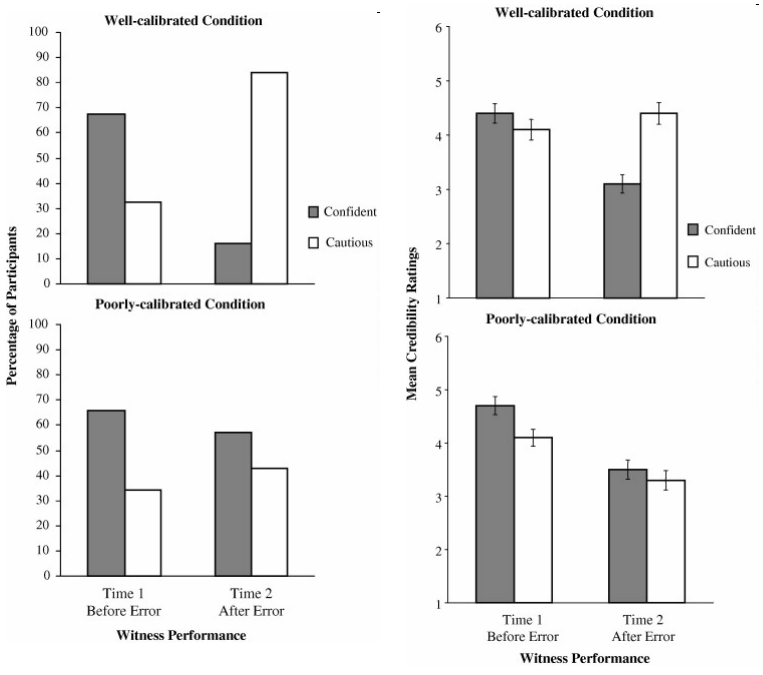
- Hypothesis: People judge source's *calibration*, not (just) their confidence
- Experiment 1: Mock juror study, 48 undergrads, confidence and accuracy manipulated in between-subject design
- Eyewitness to burglary:
  - "Yes, sir, absolutely, I'm certain" vs. "No, sir, I'm not certain"
  - "about 7:00" (contradicted by victim) vs. "about 8:15" (corroborated by victim)

Tenney E R et al. Psychological Science 2007;18:46-50



Tenney, Spellman, & MacCoun (2008, JESP): Exp. 1

- Cautiousness, not calibration?
  - *Maybe in the presence of errors people prefer informants who are more modest, or cautious, in their claims overall.*
- Well-calibrated: Cautious witness is correct about high-confidence assertion and wrong about the low-confidence assertion (as in Tenney et al., 2007)
- Poorly-calibrated: Cautious witness is correct about the low-confidence assertion and wrong about the high-confidence assertion



Exp. 2

- What if there is a good reason for a high-confidence error?
  - Justifiable error should not affect perceived credibility
- Time 1: Two witnesses identify suspect as passenger in vehicle -- one with confidence, the other cautiously
  - CONFIDENT > CAUTIOUS
- Time 2: Both shown to be in error (Time 2) about the identification
  - BOTH WITNESSES LOSE CREDIBILITY
- Time 3: A *justification* for the error is given: the passenger had an identical twin!
  - CAUTIOUS WITNESS REGAINS CREDIBILITY

Ethical and political problems with randomization

- Withholding possible benefit from controls?
  - cancelling study midstream raises threats to statistical conclusion validity (discussed later)
- Exposing treatment group to extra hardships, risks?
  - informed consent creates selection bias, expectancy effects
- "Equipoise" criterion in medical research
- Lotteries as a fair allocation rule when there is scarcity

The NEW ENGLAND JOURNAL of MEDICINE

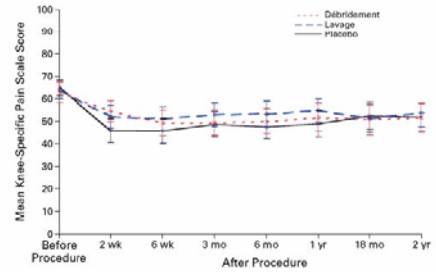
HOME | ARTICLES | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | Keyword, Ti

ORIGINAL ARTICLE

A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee

J. Bruce Moseley, M.D., Kimberly O'Malley, Ph.D., Nancy J. Petersen, Ph.D., Terri J. Menke, Ph.D., Baruch A. Brody, Ph.D., David H. Kuykendall, Ph.D., John C. Hollingsworth, Dr.P.H., Carol M. Ashton, M.D., M.P.H., and Nelda P. Wray, M.D., M.P.H.

N Engl J Med 2002; 347:81-88 | July 11, 2002



Some patients were randomly assigned to a placebo surgery condition in which "three 1-cm incisions were made in the skin." "…Incisional erythema developed in one patient, who was given antibiotics. In a second patient, calf swelling developed in the leg that had undergone surgery; venography was negative for thrombosis."

THE SPECIFIC DETERRENT EFFECTS OF ARREST FOR DOMESTIC ASSAULT*

LAWRENCE W. SHERMAN
University of Maryland, College Park and Police Foundation

RICHARD A. BERK
University of California, Santa Barbara

with

42 Patrol Officers of the Minneapolis Police Department,
Nancy Wester, Donileen Loseke, David Rauma, Debra Morrow, Amy Curtis, Kay Gamble, Roy Roberts, Phyllis Newton, and Gayle Gubman

The specific deterrence doctrine and labeling theory predict opposite effects of punishment on individual rates of deviance. The limited cross-sectional evidence available on the question is inconsistent, and experimental evidence has been lacking. The Police Foundation and the Minneapolis Police Department tested these hypotheses in a field experiment on domestic violence. Three police responses to simple assault were randomly assigned to legally eligible suspects: an arrest; "advice" (including, in some cases, informal mediation); and an order to the suspect to leave for eight hours. The behavior of the suspect was tracked for six months after the police intervention, with both official data and victim reports. The official recidivism measures show that the arrested suspects manifested significantly less subsequent violence than those who were ordered to leave. The victim report data show that the arrested subjects manifested significantly less subsequent violence than those who were advised. The findings falsify a deviance amplification model of labeling theory beyond initial labeling, and fail to falsify the specific deterrence prediction for a group of offenders with a high percentage of prior histories of both domestic violence and other kinds of crime.

Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence?

Jack Glaser , Jay Dixit , Donald P. Green

Journal of Social Issues
Volume 58, Issue 1, pages 177–193, Spring 2002

Our goal was to compare factors that are likely to inspire hate crime, specifically those discussed above: economic threat (i.e., job competition), territorial threat (i.e., minority in-migration to neighborhoods), and genetic threat (interracial marriage). In order to accomplish this, we visited various IRC chat rooms sponsored by White supremacist groups and conducted randomized interviews. Posing as a new visitor to the chat rooms, our interviewer presented scenarios of different kinds of threats and recorded the responses. These responses were then coded for their advocacy of violence so that we could compare the extent to which different types of threat differentially inspire advocacy of hate crime.

Table 1. Scenarios Comprising 3 × 3 Design of the Quasi-Experimental Survey

	Marriage	In-migration	Job competition
Personal	My sister is talking about getting married to this Black man.	I found out this Black couple is moving in next door to me.	I found out I'm competing with a Black man for my promotion at work.
Local	Lots of White women in my neighborhood are getting married to Black men.	Lots of Blacks are moving into my neighborhood.	At my work, White people have to compete with Blacks for promotions.
National	All over the country, Black men are getting married to White women.	All over the country, Blacks are moving into White neighborhoods.	All over the country, Blacks are taking White people's jobs.

Table 2. Advocacy of Violence as a Function of Threat Type and Level

	Threat level		
Threat type	Personal	Local	National
Interracial marriage	2.46 (2.21) $n = 14$	0.18 (0.67) $n = 14$	1.43 (2.44) $n = 7$
In-migration	1.5 (2.23) $n = 11$	0.0 (0.0) $n = 16$	0.0 (0.0) $n = 9$
Job competition	0.54 (1.47) $n = 12$	0.0 (0.0) $n = 7$	0.29 (1.21) $n = 17$

Two common but flawed designs

One-group pretest-posttest design

$$O \quad X \quad O$$

The static group comparison

$$\begin{array}{c} X \quad O \\ \hline O \end{array}$$

Donald Campbell's taxonomy of threats to validity

(Campbell & Stanley, 1963; Cook & Campbell, 1979)

- Internal validity
- Construct validity
- External validity (generalizability)
- Statistical conclusion validity

Internal validity

- "Did in fact the experimental treatments make a difference in this specific experimental instance?" (C&S, 1963)
  - "...internal validity is the *sine qua non..."* (C&S, 1963)
- Donald Rubin's "potential outcomes" (or "counterfactual analysis") framework is a complementary way of thinking about internal validity

Rubin's Potential Outcomes Framework

- Each individual has two scores
  - Outcome under treatment condition
  - Outcome under comparison condition
  - Sometimes notated as y_1 and y_0
    - another notation is Y^t and Y^c
- Of course, we only observe one of these scores
- The other is "counterfactual" and has to be estimated

Threats to internal validity

1. History
2. Maturation
3. Testing
4. Instrumentation
5. Statistical regression (to the mean)
6. Selection
7. Mortality (differential attrition)

1) History

- Specific events occuring between the first and second measurement *in addition to* the treatment variable
- Examples:
  - highly publicized events
  - exposure to other (non-study) treatments

2) Maturation

- "Processes within the respondents operating as a function of the passage of time *per se*."
- Examples:
  - aging (if long-term study)
  - healing/recovery/remission

3) Testing

- "The effects of taking a test upon the scores of a second testing."
- More generally, any effects of measurement on subsequent outcomes
- Examples:
  - practice effects, public commitment effects, priming effects (enhanced salience)
  - 'contamination' of jury pools
  - ICJ accidental injury survey & claiming?

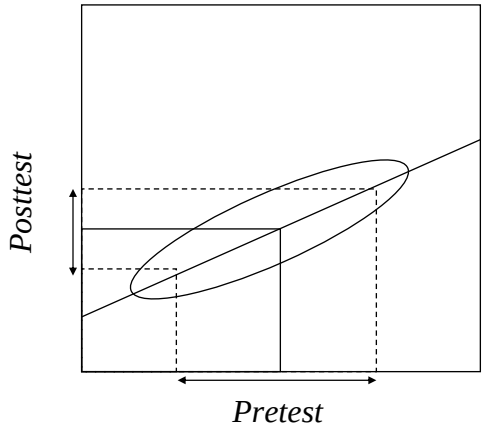
4) Instrumentation

- Changes in the measuring instrument (or the observer) that produce changes in the obtained measurements
- Examples:
  - personnel changes in interview staff
  - changes in coders' standards over time
  - mid-stream revisions in survey questions or procedures
  - addition of video or audio recording

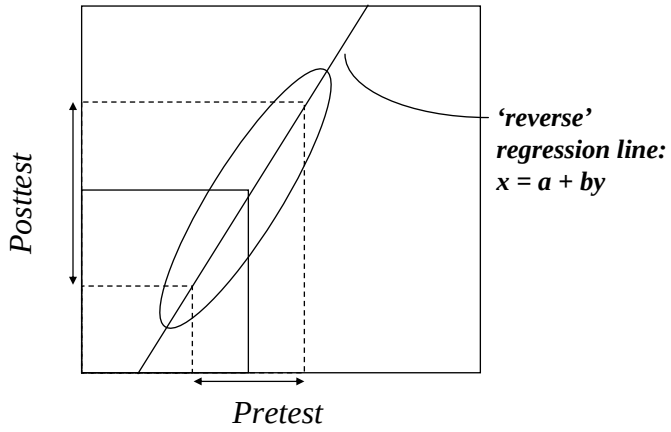
5) Regression to the mean

- Occurs when groups are selected based on extreme (high &/or low) pretest scores
- If less than perfect pretest-posttest correlation, *posttest scores will be closer to mean, regardless of treatment*
- Thus 'the best' will get worse, 'the worst' will get better

mean(z_{yi}) = $r_{xy}z_{xi}$, so if r < 1.00, then z_{yi} closer than z_{xi} to the mean



Strictly artifactual; occurs even if you use *posttest* scores to predict *pretest*...



6) Selection

- Occurs when different processes of recruitment to comparison groups
  - can be artifact of research protocol
  - can be due to respondent *self-selection*
- Examples:
  - students in Catholic vs. public schools
  - addicts in treatment vs. not in treatment
  - effects of pregancy on employment, etc.
- Econometric solutions (Heckman)

7) 'Mortality' (differential attrition)

- Differential attrition from study conditions prior to posttest data
- Involves same concerns raised by nonresponse in surveys
- In essence, "selection out" rather than "selection in"

Strategy 1: *Simple matching*

- Create a comparison group by selecting other cases matched on demographics, etc.
  - Often misnamed a "control group"
- Better than no comparison, but still flawed
  - can never establish that you've matched on every relevant variable
  - Modern matching via "propensity scores" is stronger, but no panacea

Strategy 2: *Random assignment*

- R. A. Fisher (1926): agricultural experiments
- Doesn't require any explicit matching
- Law of large numbers implies that given sufficiently large samples, *no reason to expect any pretreatment differences except by chance*
- (By chance, may have pretest diff's)

Works via 'law of large numbers'

- As cell sizes increase, the experimental groups will become increasingly similar on <u>all</u> dimensions (known and unknown)
  - Random low and high values cancel out
- Doesn't help if small cell sizes
  - randomly assigning 4 classrooms to 2 conditions means cell size is only 2 per condition…even if there are 100 students in each class.

I simulated 100 people, each with a 75% chance of having Trait A, and also a 75% chance of having Trait B. I then randomly assigned them to condition…

	Has Trait A?		Has Trait B?	
Cell size	**Control**	**Treatment**	**Control**	**Treatment**
5	**80%**	**80%**	**20%**	**60%**
10	**80%**	**80%**	**50%**	**80%**
20	**80%**	**80%**	**55%**	**80%**
50	**78%**	**70%**	**64%**	**80%**
"100"*	**77%**	**71%**	**69%**	**69%**

* actually, 94 and 106. Why? Random assignment doesn't guarantee equal cell sizes, so sometimes researchers force cells to be equal. (Like quota sampling)

Random assignment worked right away for Trait A, but by chance, *treatment was confounded with Trait B* until cell sizes got large.

"Natural Experiments"

- Sometimes interventions get allocated via random or quasi-random processes
  - *Exogenous shocks*
  - *effect of Afghan invasion on street price of heroin*
- Rarely truly random, so need to carefully test for treatment confounds
  - *Vietnam draft lottery*

Pretest-posttest control group design

R	***O***	***X***	***O***
R	***O***		***O***

- Very strong for internal validity, though pretesting raise *testing* concerns regarding external validity

Posttest-only control group design

R	***X***	***O***
R		***O***

- Preferable to *pretest-posttest control group* design -- not vulnerable to testing-treatment interaction
- But might want to include the pretests if you expect differential attrition (so you can compare the dropouts and non-dropouts)

Dealing with differential attrition

- Even if groups are equated by randomization at the outset, they may not be comparable after some have dropped out
- Loss of statistical power is bad, but potential bias is worse
- See technical appendix for slides on *Intention to Treat* analysis

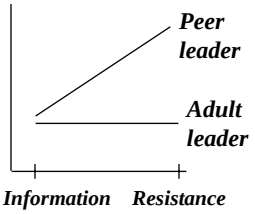
Factorial designs

- Each version of IV_a *crossed* with each version of IV_b
- Allows test for *interaction effects...*

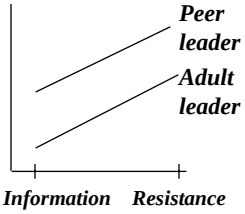
	Information about risks	*Resistance training*
Peer leader	60 students	60 students
Adult leader	60 students	60 students

A "2 x 2" factorial design for drug prevention

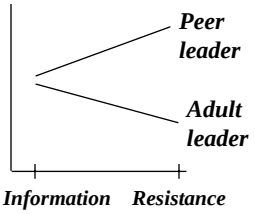
Conditional effect



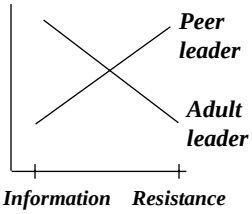
Additive effects



Fan



Crossover



Additional design variants

- *Between*-subjects vs. *within*-subjects ('repeated measures') designs
  - Between: each person exposed to single condition (single level of IV)
  - Within: each person exposed to multiple levels of IV (essential to counterbalance order)
- *Nested* designs
  - e.g., randomly assign class to condition; students nested within class
- *Intentionally confounded* designs (for economy):
  - Latin-squares, hyper-graeco-latin squares, fractional factorials, etc.

Parametric designs

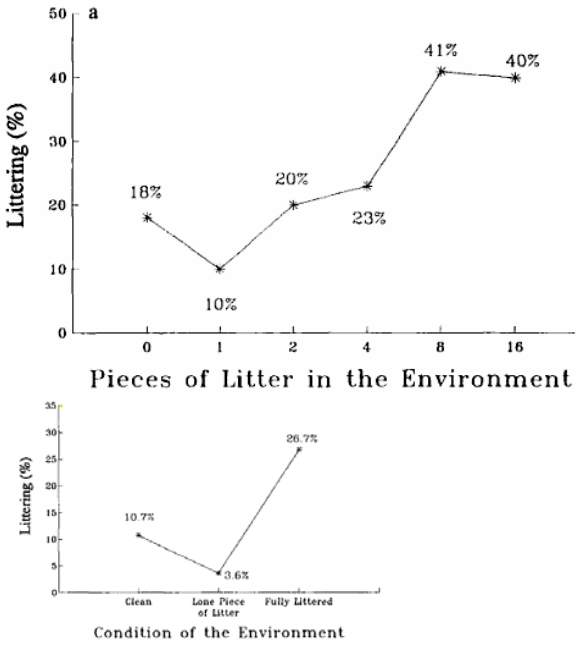
- Simply comparing two levels of a variable will not tell you about its functional form
  - E.g, diminishing marginal utility, U-shaped relationships, S-shaped dose-response curves
    - E.g., Prospect theory vs. alternative theories
  - If you choose two locations on the "wrong" part of the curve, you might reach misleading inferences

Cialdini, Reno, & Kallgren (1990):

Theory predicted that single act of deviance increases compliance, but multiple acts will increase deviance.

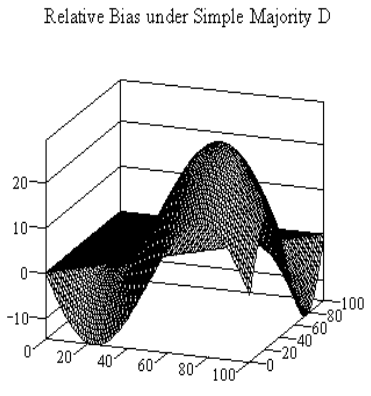
Tested first in parking garage by manipulating number of items of visible litter (paper).

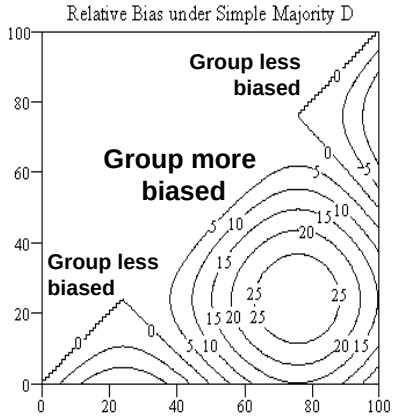
Replicated "checkmark" pattern in another study, using watermelon rind in dormitory mailroom.



Effects of Jury Deliberation on Biases

- Kaplan and Miller (1978) argued that deliberation corrects juror biases
  - *Deliberation emphasizes evidence*
- Mock jury experiment
  - *Varied obnoxiousness of trial actors*
  - *Strong prosecution vs. strong defense case*
  - *Groups shifted in direction of the evidence*
  - *Bias was attenuated by deliberation*
- But Kerr, MacCoun, & Kramer (1996) showed that this was misleading…





Construct validity

- In *measurement* (Cronbach & Meehl, 1954):
  - do these items actually measure the intended latent construct? (e.g., intelligence)
- In *causal inference* (Cook & Campbell, 1979):
  - does the treatment implementation accurately represent the hypothesized treatment?
  - construct validity of outcome measure?

Treatment confounds

- Not explicitly listed in C&S, but extremely common problem in experiments
- In essence, if 'treatment' involved more than one 'thing', which was the cause?
- Examples:
  - different sites or different administrators
  - treatment involves multiple program elements
  - treatment group asked extra questions
  - 'Hawthorne' effect
- May require special control groups

Lab studies show that sequential lineups are fairer than simultaneous lineups.

But controversial Illinois State Police pilot program experiment claimed to find the opposite…

	Simultaneous presentation (n=319)	Sequential presentation (n=229)
Suspect ID	60%	45%
Filler ID	3%	9%
No ID	38%	47%

Gary Wells' critique

- "My main reaction to this report is disappointment and concern that the design of the study does not permit any clear conclusions. The reason is…because the simultaneous lineups never used the double-blind procedure whereas the sequential lineups always used the double-blind procedure."

WHEN AND WHY INDIVIDUALS OBEY FORM-ADHESIVE CONTRACTS: EXPERIMENTAL EVIDENCE OF CONSENT, COMPLIANCE, PROMISE AND PERFORMANCE[†]

ZEV J. EIGEN[††]

Web survey with 480 questions spread out over 480 separate pop-up web pages! How many will participants actually complete?

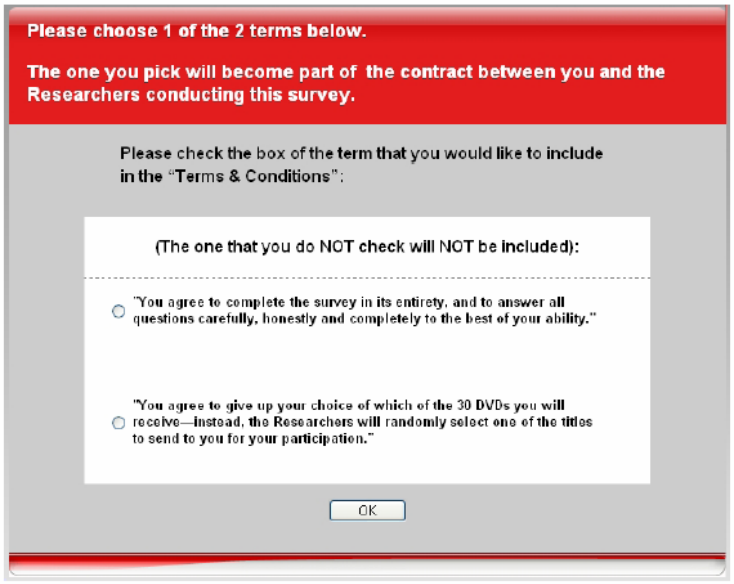
Conventional boilerplate version: 7-paragraph standard end-user contract with "consent to participate" check box.

Substantive choice version: Given two choices (NEXT SLIDE)…

"Results of an online experiment reveal that marginal participation in contract drafting increases drafters' performance of an undesirable contract term."

Problem: Confounded substantive choice with salience of requirements.

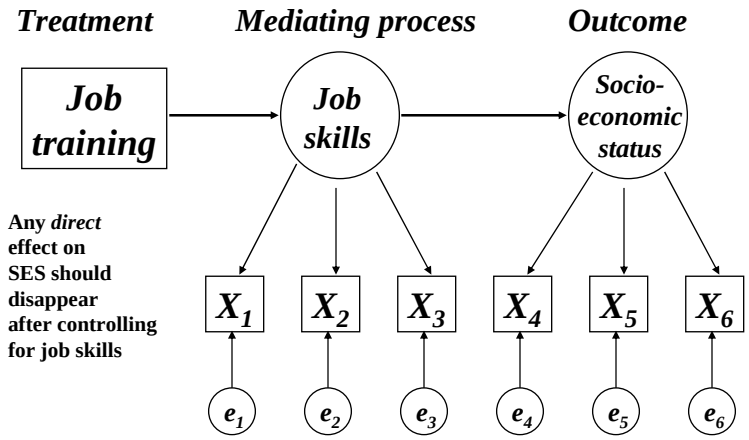
FIGURE 6
CONDITION 3: SUBSTANTIVE CHOICE



Manipulation checks

- Measures to determine that intended treatment was actually experienced ('assessing the take of the IV'):
  - was it administered? properly?
  - did respondent perceive and understand it?
- "Internal analysis"
  - test program effects using manip check rather than assignment as the IV; sacrifices benefits of random assignment

Mediation model



Ideally, use multiple indicators (Xs) for each construct

Expectancy effects

- Hypothesis guessing, 'demand characteristics'
  - respondent modifies responses to try to help or hinder researcher
  - Orne (1962): S's worked for 5 hours summing random #s
  - can require 'cover stories', single blind designs, *placebos*
- Experimenter expectancy effects
  - Rosenthal: gave E's hypotheses, biased results even when E only read instructions
  - requires double blind designs

Problem of failing to reject the null

- Karl Popper: Can only falsify a theory; can't 'confirm' it
- Fisher: Can only reject the null, can't confirm it
  - problem: the null is rarely your hypothesis
  - can we ever know for sure there's "no effect"?
- Failure to reject null could be due to:
  - small sample size
  - weak instantiation of an effective treatment
  - noisy measurement

Risk of using non-experimental approaches? Glazerman, Levy, & Myers (2003)

- 12 case studies on social welfare with:
  - An experimental evaluation
  - 1+ non-experimental (NX) evaluations
- Size of bias (in 1996$ of annual earnings):
  - Regression: $1,101 (about 10% of annual earnings)
  - Matching: $1,143
  - Selection or instrumental variables: $2,791
- **"potential for very large bias"**

Wilde & Hollister (2007)

- Project STAR – Tennessee class size experiment
  - Experimental data from 12 schools
  - Compared to use of propensity score methods for each site
- Estimates were >10 percentile points apart for 8 of 12 schools
- Based on cost-effectiveness criteria, “the nonexperimental estimate would have led to the wrong conclusion in 4 of the 11 cases”

Technical appendices

Significance test controversy

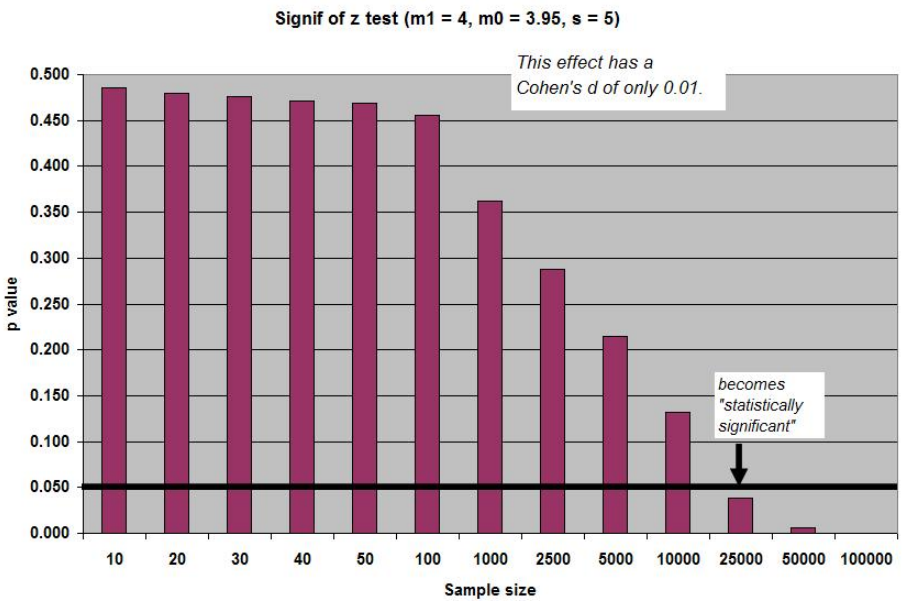
- Arbitrary nature of 'p<.05' criterion
  - extreme aversion to Type I errors
  - neglect of risk of Type II errors
  - 'cliff effect': arbitrary threshold creates binary decisions
- Overreliance on *p-values* (statistical significance) rather than *effect sizes* (substantive significance)
- Complaints about fishing expeditions vs. calls for exploratory data analysis

Confusion about significance

- *p*-value $\neq$ *p*(H_o is true|data)
  - i.e., p-value doesn't tell you "less than 5% probability that there's no effect" -- what we'd really like to know!
  - can only know using Bayes Theorem, but we'd need to know the prior probability, *p*(H_o is true)
- *p*-value = *p*(data|H_o is true)
- *p*-value $\neq$ *p*(Type I error) -- see next slide
- *p*-value = *p*(Type I error|H_o is true)

H_o= 0 ('nil hypothesis') is always false

- "It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?" (Cohen, 1990)
- All 105 possible 2-way crosstabs among 15 attributes of 57,000 Minnesota HS students significant; 96% at $p < .000001$ (Meehl, 1990)



Alternatives to sig. testing

- Confidence intervals
  - avoids dichotomous thinking, highlights uncertainty
  - even better if combined with robust statistics, "bootstrap" standard errors
- Bayesian statistical analysis
- The p_{rep} statistic

The p_{rep} statistic

- Killeen (*Psy Science,* 2005)
- Want to know $p(d_2 > 0 | d_1)$, where d_1 is observed effect size in earlier study and d_2 is effect size in next study
- P_{rep} = area under curve of normal probability table up to $z = d1/\sqrt{2\sigma^2_d}$

In praise of p_{rep}

- Valid? Calculated as .71, .75, and .79 for 3 meta-analyses where effect was replicated 70%, 74%, and 82% of time
- Requires no assumptions about null hypothesis -- compare p_{sig} = p(data|Null is true)
- Easy to communicate: "this effect will replicate 100(p_{rep})% of time"
- But see Geoffrey Iverson et al. (2009a, 2009b) who show that p_{rep} is sometimes misinterpreted, and sometimes too optimistic

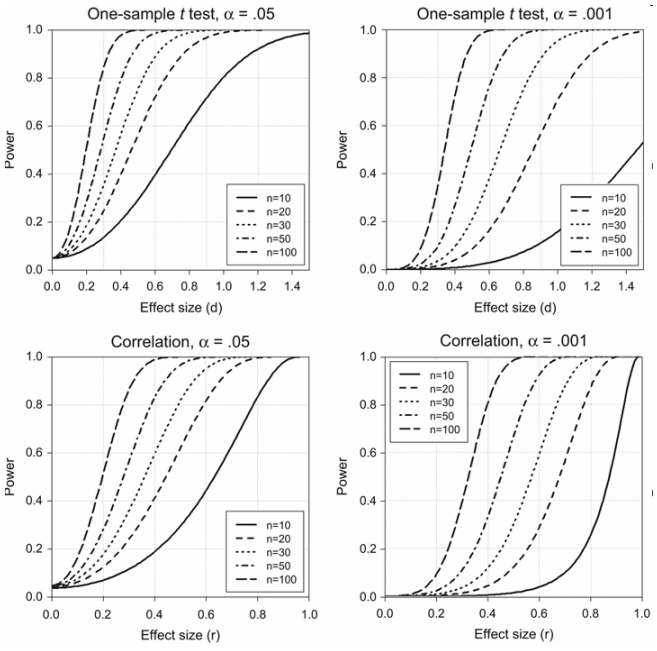
Power analysis

- Power (1 - β) = *p*(Accept H_1|H_1 true)
  - power is a function of significance level (α), sample size *(N),* and population effect size (ES)
- Cohen suggests conventional level of .80
  - i.e., .80 : .05 = 4:1 ratio of Type II:Type I errors
- Average power for medium ES, all articles in *Journal of Abnormal Psychology*
  - 1960: .46 (Cohen, 1962)
  - 1984: .37 (Sedlmeier & Gigerenzer, 1989)

Rossi (1990)

- Power was calculated for 6,155 statistical tests in 221 journal articles published in the 1982 volumes of the *Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology*, and *Journal of Personality and Social Psychology*.
- Power to detect small, medium, and large effects was .17, .57, and .83, respectively.

Yarkoni (2009)



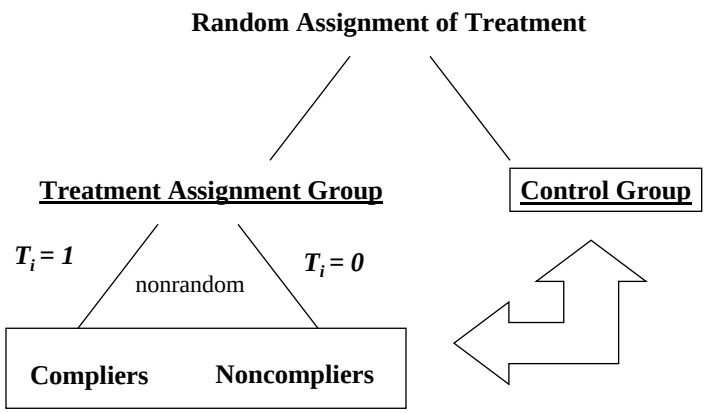
"Minimum detectable difference" (MDD) approach

- When trying to determine sample size, "MDD" refers to the smallest effect size you *want* to detect
  - E.g., smallest effect that would be still worth pursuing based on clinical significance or cost-effectiveness
- When N is fixed by real-world constraints, "MDD" refers to the smallest effect size you *can* detect
  - for a given level of power and alpha—usually .8 and .05

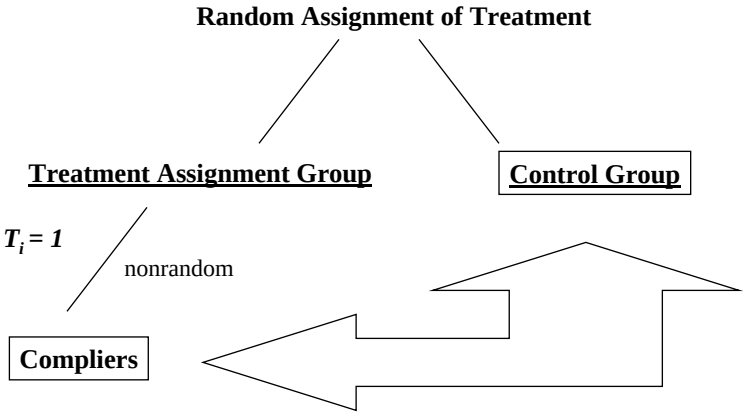
Power for 2x2 interaction?

- Recall that in a 2x2 factorial experiment, you can have significant main effects for each variable, and/or a significant interaction effect involving both IVs.
- The power needed to detect a 2x2 interaction effect in a factorial experiment may be the same as the power needed to detect the main effects of the 2 variables. Or you may need more power. It will depend on the nature of the interaction and the degrees of freedom of the test.
  - For details, see Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin, 110,* 587-595.

***Intention to Treat (ITT)*: includes noncompliers in the treatment effect, biasing it downward. But random assignment is preserved.**

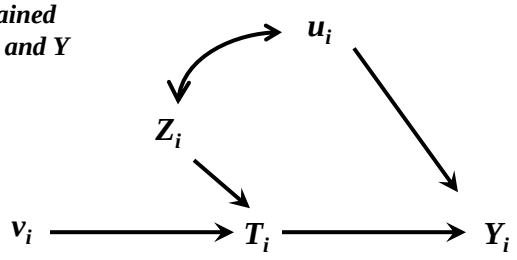


***Average Treatment Effect on the Treated* excludes the non-compliers, so no longer true random assignment. (Threat of selection and attrition biases)**



T = Treatment
Y = Outcome
Z = propensity score
v, u = unexplained variance in T and Y

Propensity Scores (Rosenbaum & Rubin, 1983)



<u>Propensity Score (Z)</u>:
- ***want Z to correlate with u***
- ***controls for "selection on the observables"***
- ***assumes r(u,v) = 0; i.e., controlling for Z, T is uncorrelated with u***

Controlling for propensity scores, can use observed no-treatment outcomes to infer treatment group's no-treatment counterfactual