

P.E.Meehl: Philosophical Psychology Seminar
Tape 1 of 12 Lecture 1 1/5/89

Aims of course: Intrinsic interest; appraise theories; defend against dumb arguments

Metatheory as the empirical theory of theories
But includes rational reconstruction

Object language and metalanguage

Historical overview:

Logical positivism, Vienna Circle.

Persons; doctrine; two aims: (1) Liquidate metaphysics, "pseudo-problems"

(2) Rational reconstruction of science

Berlin group

Verifiability criterion of meaning

Operationism (Bridgman) Psychology courses 50 years behind philosophy

Behaviorism (Watson, Tolman, Hull, Skinner)

PEM did first 10 sessions (Winter Quarter, Jan–Mar 1989). Spring Quarter, several other department members did sessions on various topics. PEM did last two sessions (5/25/89 and 6/1/89).

PEM objected to trailing wire for microphone and requested a wireless one. First session there had been a communication problem, so he put up with a wire (notice he doesn't move about as freely as in later sessions). Second session there was a wireless microphone. But for the third session there was only the hard-wired one again. He is obviously angry throughout the session, and that is why. Subsequently, he informed the people in charge that if they couldn't provide a wireless microphone, there would be no videotaping.

P.E.Meehl: Philosophical Psychology Seminar
Tape 2 of 12 Lecture 2 1/12/89

Popper and related

Popper 1919 eclipse, Adler, Marx, Freud, Astrology, Einstein

Confirmation vs Falsification seeking

Demarcation of science vs non-science

Basic logic: Implicative syllogism, modus tollens, 3rd figure invalid (empirical science)

Risky tests → Stronger corroboration
Spielraum, narrow prediction

Salmon's damn strange coincidence
Bayes's Theorem

Refuse protocol. Theories can control facts.
Ether drift. Mendeleev.

Feyerabend against method.

Lakatos diachronic; Degenerating programs

P.E.Meehl: Philosophical Psychology Seminar
Tape 3 of 12 Lecture 3 1/19/89

Minnesota Center for Philosophy of Science

Reichenbach's contexts discovery/justification: basic, retain
Some appropriate overlap. Example: Latent learning bias in labs
Why false protocol is worst scientific sin
Social influences (e.g. Bethesda fads)
Indirect costs, phony, Universities depend on
Cost of apparatus vs questionnaires, large vs small samples
"Clinical experience" vs hard data

Truth as meta-predicate
Truth-phobia in social science students
Instrumentalism/realism

Skinner on theoretical constructs

Linkage maps and chromosomes (Morgan & Co.); holes in giant chromosome

Theories of truth (correspondence, coherence, pragmatic);
Tarski: Semantic conception.
Quine corners.

Carnap vs Kaufman: 'True' a legitimate predicate

Verisimilitude (brief introduction)

Truth phobia (cont from Lecture 3)

Verisimilitude (cont from Lecture 3):

Ontological, not epistemological;

Not “probable,” that’s an *evidentiary* concept;

Matter of degree. Everything we consider is “complexes.”

Some concepts useable and unavoidable, even if unclear (e.g., probability).

Ex: Kinetic theory of heat [*PM makes error in lecture: Van der Waals
force is not gravitational*]

“Falsified” (literally) ≠ “abandoned”.

Constants of Van der Waals equation not theory derived. Curve fit.

Falsifier becomes corroborator of modified theory.

Ex: Neptune as perturber of Uranus orbit

Two kinds of adjustments: (1) theory itself modified; (2) Alter or add particulars, no theory change.

Formalization (hypothetico-deductive postulates) can be done by logician even if scientist (e.g., Skinner) doesn’t bother.

Core vs periphery of a theory (Lakatos hard core vs protective belt)

Latent learning: Skinner, Hull, Tolman; Freud, neo-Freud; Hull, sH_R core

PEM attempt to explicate “core” by class of all experimental *wffs*

Theory as “complete,” all true *wffs* derivable

PEM Verisimilitude explication-sketch

Kinds of entities (substance, structure, event, state, disposition, field)

Three kinds of theories: functional-dynamic, compositional, developmental

Fisher effects=partial derivatives for continuous case.

Interaction=mixed partial derivatives for continuous case.

Nomological net. Nodes=entities, strands= relations between entities.
Proper subset “operationally” linked.

Levels of verisimilitude (I–X): Examples. Kinds of entities? Connections, Derivatives?
Function forms? Parameters?

Verisimilitude=Similitude to Omniscient Jones’s Theory

Lakatosian defense of core. Basic Formula $T \cdot A_T \cdot C_P \cdot A_1 \cdot C_n \rightarrow (O_1 \supset O_2)$

Defend core despite falsifiers, if theory already has money in the bank,
has a good track record.

Gets credit by risky predictions, “damn strange coincidences”

So $(O_1 \supset O_2)$ had low prior, absent theory

Avagadro Number & molecular reality

X-ray diffraction: Which is theory, which auxiliary?

Two mistakes on significance testing

Weak use, think theory strongly supported by refuting H_0

Strong use, think falsified theory must be abandoned pronto.

Lakatos’ “instant rationality”

Auxiliaries example: Personality theory experiment about anxiety, introversion,
TAT affiliation

Fisher on “controlling all variables.” Lady tea taster

Lakatosian retreat (cont.)

Hard core could be true or have high verisimilitude. Same strategy.

Monkeying with auxiliaries is hydraulic: Repair this falsifier, changes another derivation

Definition of 'hard core': Concepts? Statements?

More peripherals in social and biological sciences

Lakatosian defense justified because of track record

Bulletin style narrative summary of research hard to interpret

"Soft" fields (clinical, counseling, personality, school, community, social, developmental)

Correlation vs experimental manipulation with randomization of subjects

Correlational (non-experimental) data, relying on refuting H_0 , is domain considered here

But include experimental study of interaction with IDs (demographics, traits, status)

having crucial role as theory test

10 obfuscating factors: (a) sizeable, (b) variable, (c) opposed (countervailing),
(d) not accurately estimated

Meta-analysis is not for appraising theories; useful for interventions, technology

10 obfuscators:

1. Loose derivation chain

2. Auxiliaries, when stated, problematic

3. Ceteris paribus clause doubtful

4. Particulars C_n imperfectly realized

[Excursus: Referee dogmatism, quick dismissal. Science as skeptical, avoid superstition.
One failure to replicate, don't assume second try has to be the correct one.]

5. Inadequate power (Cohen)

Estimate degrees of freedom needed from pilot study

Power function

Obfuscators 1–5 tend to give black eye to good theories, of high verisimilitude.

Obfuscators working in other direction, tend to make poor theories look good:

6. Crud factor

In social sciences everything is correlated with everything

Lykken-Meehl study on crud factor (high school students, big N)

MMPI pool. 96% of 550 items discriminate gender

Two pots: Theories in one, facts in other; pair them randomly

Example: Crud factor \rightarrow t-test significant

Some statisticians become dogmatic

6. Crud factor (cont.)

Numerical example, true and false theory track records hardly different
“Theory pot” randomly paired with “variables pot,” zero verisimilitude,
yet appears corroborated

Always a trade-off between Type I and Type II errors so not helpful that
 α can be set low

Distinguish statistical hypothesis from substantive theory. Books rarely do. Terrible!
Probability \times utility in technology. But no analogous rule when “gambling with truth”

Size of crud factor unknown but not negligible in life sciences. *Examples*: *g* factor,
MMPI, SVIB, CPI, teacher rating scales, Thurstone primary abilities

7. Pilot studies

2 questions: Effect exists? *N* needed for α ?

Some *N* easier to raise (e.g., questionnaires) leads to domain and theory bias

8. Bias favoring significant result in MS submissions

9. Bias favoring significant result by referees and editors

10. Detached validation claim for psychometric instruments

Example: Valid variance easily less than 1/2 of reliable. So how know which region is
causing correlation?

These 10 obfuscators work oppositely so net result unpredictable

“Box score” in soft psychology usually runs 2/3 to 4/5 favorable.

Not impressive but people think it is. Asymmetry of falsification
and support—modus tollens.

Can do binomial on box score

[Diagrams of strong, medium, feeble tests, distributions of theory + error tolerances]

Student delusion of testing theory with correlations on MMPI. Due to simplistic
operationism + verificationism + null hypothesis testing habit

Significance tests (cont.)

For technology significance test with overlap is ok; we're discussing testing theories

Ways to improve theory matters:

A. Investigators

1. Expected size of trend, on theory?
2. Power (Cohen), despite its dangers
3. Campbell-Fiske matrix. Put in main suspected nuisance variables.
4. Pilot study (despite dangers)

B. Editors, referees, journal policy

1. Require replication in one paper
2. Tables: *Separations*, standard deviations, sigmas, overlap, not mere significance levels. Give p value, not "n.s.," ".05," ".01". Confidence intervals always preferable (translatable to significance but not reverse). Several overlap cuts (e.g., 10, 25, 50, 75, 90%) passed by experimental group; so reader can choose which to emphasize for his purpose. Percent variance accounted for, in some designs. β -weight \neq causal influence, usually.
3. Journal section for short reports of negative pilot studies.

C. Reviewers of literature

Stress power (usually never mention it)
Emphasize that Box Score doesn't tell much about verisimilitude. Delusion that box score (+) over (-) is "pretty good." If negatives accepted, it's terrible (modus tollens)

Psychologists must first grasp Popper point, then water it down like Lakatos. Most in "soft" psychology haven't reached the Popper stage.

D. Theoreticians

Optimism and pessimism both: 1) We're doing fine with present methods;
2) Soft fields cannot aspire to stronger tests
Numerical point predictions only one kind of strong Popperian test. Ex: Wien's Law, no parameters predicted; "Some function of λ^5 ," look at graph, beautiful fit.
Intermediate theory strength

E. PhD educational practices

Require math for research psychologists: More *math* rather than more *statistics* (much of Fisher statistics of little use to us)
Read experiments in other sciences, especially derivation chains with text; Ex: Millikan oil-drop experiment; but *physics* is not the only model of good science. History of science? I don't know. I think it has helped me.
Publish or perish pathology is terrible for people and for science. Proxmire had a point. Evaluation of publications should rely more on *Science Citation Index* than mere paper count.

Nevitt Sanford study of achievers: Syndrome of self-doubt

Some important, meaningful scientific questions can't be answered at a given time, lack auxiliary theories or good instruments; Crick & Watson needed (a) exact weights, (b) X-ray, (c) quantitative details of hydrogen bonds, molecular distances.

Probability concept

Important in social sciences, our laws are often stochastic

Epistemic versus object-language use of 'probability'

Carnap: probability₁ (confirmation) and probability₂ (relative frequency):

physical or social; relative frequency of event or attribute in a specified physical class.

Stated in object language. One we use in statistics courses.

Theory of probability began with gambling problems

Ratio of favorable to total "equally likely" ways; Pascal and Fermat invented it; limit of relative frequency as definition introduced by von Mises, Reichenbach

von Mises theory

Collective: random sequence, relative frequency converges to a limit; all "place selections" yield same limit

Not "limit" in usual sense of math, rather "stochastic limit"

Relative frequency can't be defined by "equally likely ways"

(e.g., insurance tables don't list equally likely causes of death)

Probability₂ always leads to a number

Probability₁ is metalinguistic concept; epistemic; logical relation between hypothesis and evidence; relation between propositions, beliefs, statements; prima facie, seems not to be about relative frequencies.

Probability (cont.); repeat end of last lecture to clarify:

Probability we usually use in psychology is probability₂: Proportion in a class, percentage, relative frequency of events or properties; a decimal value $0 \leq p \leq 1$ closed interval; “certainty” = ($p = 1$), but not conversely

Object language. Properties of physical objects or events in a domain (genetics, chemistry, psychology, economics)
19th century, John Venn; and Ellis defined by relative frequency

Kolmogoroff, axiomatized probability calculus. P not *defined* by reference to relative frequency. Only 3 postulates about probability numbers. All abstract. Have to coordinate linkages of the p -numbers with empirical proportions. I don’t know how to do that.

Popper. Probability is a *propensity* (= tendency, disposition), and a formal axiom system.
Thinks you need more axioms than 3.

Fisher is a frequentist, but not the Mises-Reichenbach kind. Introduce π by axiom, then *prove* it relates to a frequency.

Psychology students think frequency is the only kind of probability there is, from way statistics is taught. Another kind of probability (life, law courts, even science) doesn’t *look* like a relative frequency.

Historical fact (e.g., Katyn massacre) on evidence, doesn’t look like any kind of frequency.

Kaspar Hauser son of a prince? How express a probability of that as a frequency? Only one such person.

Wegener theory of continental drift. If no other planets exist, still meaningful to say “Wegener’s theory is probable on the evidence.”

Facts may be statistical or not. But that doesn’t make the evidentiary relation *between* facts and theory statistical.

Schizophrenia is a neurological disorder, on the evidence $p(T/F)$ has no algorithm to get a p -number. Bayesian subjectivists *extract* a p -number by forcing people to bet. It works.

But those subjective betting odds are not reached by computing a frequency.

Start with Carnap’s probability₁ and probability₂ *prima facie* distinction; *then* inquire whether they can be identified, or how related if distinct. Probability₁ is about relation between beliefs, statements, propositions—rather than relations between events or properties of physical events

Can always avoid facts (Flat Earth Society)

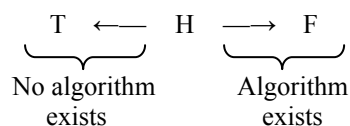
Cardinal Newman’s book *Grammar of Assent* is great on probability₁

Example: Evidence that Hauptman killed Lindbergh baby [passout]

Example: Snyder’s genetics text (pre-1953) that genes are located on chromosomes [passout]

Any juror has to estimate probability₁ without algorithm

Scientific theories are probability₁ on evidence (Piaget, Darwin, Freud, Big Bang) not numerified by an algorithm



probability₁ called “logical probability”

Carnap worked on a probability₁ algorithm

Most don't think it worked

Perfect ideal language of state-descriptives

“Principle of Indifference” or “Principle of Insufficient Reason” applied to state descriptions can give an algorithm [Grover Maxwell story on state descriptions]

Probability₁ and probability₂ prima facie different, and we need both kinds

“Probability is the guide of life” (Bishop Butler)

Query: Yet how are they basically the same? How come same term used for both (in most languages)?

[Randomness of von Mises collectives is called “Principle of Impossibility of a Gambling System”]

von Mises said shouldn't use term ‘probability’ at all, for probability₁.

Reichenbach said only one meaning, limit of relative frequency, for both kinds.

Hard to defend. Example: Probability₁ of scientific theories would really be relative frequency of truth for theories having certain properties.

Some of our most fundamental concepts are fuzzy. Example: Probability; causality.

Why one word? Carnap: “Fair betting odds.” Subjective Bayesians even *define* probability that way.

If truth-frequency in long run doesn't match a purported probability₁ algorithm, algorithm is defective. Example: 1000 murder cases where truth is known; a jury algorithm for evaluating evidence should agree; otherwise it's no good however logically plausible it seems. In that sense relative frequency has a basic status in all probability concepts. So identity theorists like Reichenbach have a point.

de Finetti, Savage “Dutch Book” argument.

Conceptually distinct, yet probability₂ and probability₁ should tend to agree in long run.

Clinical vs statistical prediction

Autobiographical note: My Freudian interests vs 1938–1945 Minnesota department (anti-Freudian, behaviorist, statistical)

Examples: Law school admissions; criminal parole; suicide risk; EST or pills. Serious matters.

(“Help,” “change” presupposes prediction.)

Doing nothing is a form of action, a decision, based on estimate of effects of options. All such predictions are probabilistic and will remain so. Some physical sciences also stochastic (e.g., meteorology).

Meteorology predictions only 15% better than “persistence” prediction

Closed outcome set: Defined predictive task

Almost all such judgments are made by “informal” method (reflect, discuss, vote, or chief decide). On any given day in USA, 99.9% of all decisions about human behavior are made informally.

Kind of data vs mode of combining them

Other way: Once data encoded, a mechanical way, algorithm, puts data together → prediction. “Actuarial” “statistical” But also can make a mechanical rule non-actuarially (armchair)

Genes on chromosomes: Sometimes isomorphism supports identity

Data vs mode of combining, people keep mixing up

Usual prediction situation in psychology or sociology has both psychometric and non-psychometric data, to be combined somehow. Test scores, ratings, school record, etc.

Informal method may rely partly on statistical data. How much theory do we rely on? Varies widely among clinicians (e.g., Michigan vs Minnesota)

Genetic statistics of diagnosis may be only slightly theoretical. Phenotypic trait: Content (semantic) resemblance of dispositions + empirical covariation of them

“Mechanical” combining: By a clerk. Doesn’t assume any statistical form (e.g., linearity). Algorithm, regression equation, actuarial box (function-free), nomogram, computer program.

Configural and powers combinations generally subject to sampling error capitalization

Lykken Actuarial Box: No math function inferred. Glueck delinquency prediction tables.

Mechanical ≠ Actuarial. Actuarial is based on empirical data, tallied frequencies. Can do a mechanical “subjectively,” based on my experience but not tallied—memory, impressions, clinical experience.

Delphi Method: Expert opinions but no meetings. Convergence of opinion? One way to proceed mechanically without actuarial tallying, rely on clinicians.

All actuarial prediction is mechanical, but not conversely.

“But want to predict for this individual, not about groups.” Unsound.

Ex: Should I have drastic surgery?

Ex: Why buy life insurance?

Ex: Russian roulette. Do you prefer a gun with one live, or one empty chamber?

More facts specify a narrower subclass. Insurance actuaries do this. Each added attribute narrows class. “Decrease extension by increasing intension.” Reichenbach rule: “Use smallest reference class for which you have stable relative frequencies.” Basically sound although some technicalities in applying. Each *p*-value is “correct.” But the narrowest is the one to rely on for prediction.

Almost everyone assumes clinical, informal method, the usual one, is the best, “obviously.” People say “*Obviously* you can predict better by understanding the *individual*.”

First comparison, 1928, Burgess on parole violation prediction.

Crude, unweighted sum of 21 factors beat out all 3 prison psychiatrists.

Sarbin (1942) prediction of college grades.

Ex: Wittman predicting response to shock therapy.

Skilled clinician to make some of the ratings (e.g., anal/oral). But how combine?

[Computers still not good in pattern recognition]

Ex: Apostolakos and Martin on diagnosing jaundice.

Meehl study of 29 clinicians vs 6 actuarial methods, neurosis/psychosis, MMPI.

Will Grove survey of studies. Expects 150 before he’s done.

No controversy in social science where studies pile up so clearly in some direction.

Parole or recidivism predictor, same set of predictors work: How many crimes, age of first crime, school level, horizontal mobility, chemical dependence, associates, longest job in private sector, IQ, Porteus Maze Q-score, MMPI.

Many studies show adding more information lowers accuracy. Information overload.

Goldberg Paradox: Clinicians do worse than an equation based on predicting their predictions. (Because clinicians don't apply their own weights consistently.)

Actuarial method is atheoretical. This bothers people.

For theory to work in predictions,

1. Theory has high verisimilitude
2. Accurate measuring instruments

We do not meet either condition.

Similar debate in meteorology as to how much theory to use vs pure blind actuarial method, statistical equation on data.

When algorithm omits a factor so potent it countervails everything else. Meehl's broken leg case. When do you have a broken leg case? This judgment itself is often poorly made. If clinician can spot broken leg case, he will beat the equation. Since he doesn't, we know he over-identifies broken leg cases that aren't there. High school algebra proves this.

Train clinicians to have higher threshold for calling broken leg case. Then they might do better than equation.

Maybe organic medicine can do better. State of theory good enough? Howard Horns glutamic acid example.

"Two methods complement each other, shouldn't set up a conflict." Dumb.

"All these years people have been making judgments." Dumb.

P.E.Meehl: Philosophical Psychology Seminar
Tape 11 of 12 Lecture 11 5/25/89

Psychoanalysis

[Outline enclosed separately]

P.E.Meehl: Philosophical Psychology Seminar
Tape 12 of 12 Lecture 12 6/1/89
(poor sound in spots)

[Psychoanalysis, cont. from previous lecture:

Why psychoanalytic session is better evidence than experiments or correlational studies. If the free association method is not a good data source, then psychoanalysis is probably erroneous. The core problem of psychoanalysis is the session inferences.

Distinction between 2 things: (1) No experimental or statistical evidence on this; (2) Evidence against but I believe it anyway. Rely on clinical experience given (1). Not if (2) is against your clinical impressions. Not just scientific error—*moral* wrong.]

Appraising a scientific theory if we realize H_0 is poor way to do it:

Weak and strong use of significance tests

Epistemological risk \neq Statistical risk. That α is small, doesn't mean theory risk is small.

“Numerify” is word, weaker meaning.

Popperian risk. Salmonean coincidence. “Almost hits,” “near misses” can sometimes corroborate theory strongly.

Index must combine risk with closeness.

Case studies can't settle metatheory arguments. Anecdotes all refute statements never made. Ex: Prout's hypothesis. Ex: Popper's example (Bohr-Kramers-Slater)

If metatheory is inherently stochastic, way to study it is actuarially.

Many properties of theories besides their factual performance (as Laudan shows).

Empirical fit: 2 components: How narrow tolerance, how close we come. *Spielraum*.

Any method must relate the theory tolerance to the *Spielraum*. Ex: Baby elephant trunks.

Index Numbers problem

$$C = \left(1 - \frac{D}{S}\right) \left(1 - \frac{I}{S}\right) \quad [C^* \text{ handout \& \#147}]$$

Function forms without parameters. Curve fitting of constants, still can predict function form.

15–20 functions make up 99% of those occurring in science.

“Closeness”? Not a residual sum of squares.

 Pure error (influenced by size of individual differences)

 Lack of fit (this is what we want)

4 theoretical track record aspects

 Point predictions

 Function form

 Reducibility (Comte Pyramid)

 Qualitative diversity of experiments

I prove verisimilitude and track record are highly correlated