Science as Social Choice

Conor Mayo-Wilson
Department of Philosophy
University of Washington

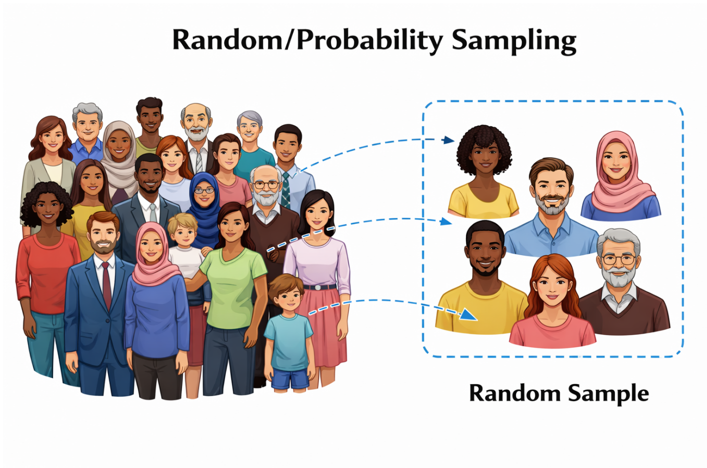
March 6, 2026

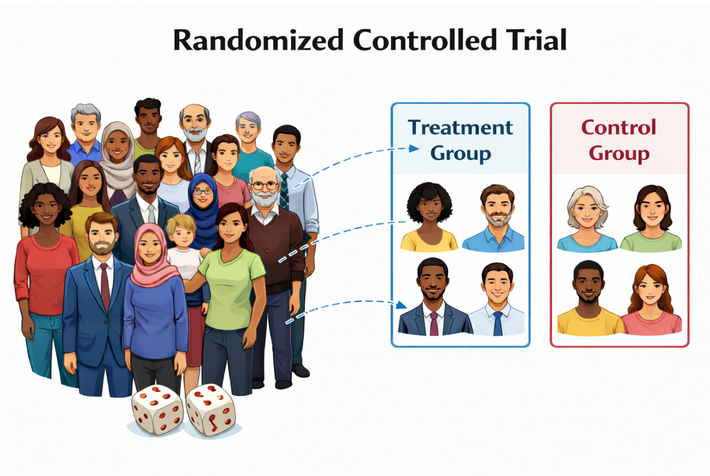
Thesis 1: According to orthodox, *individualistic* rational choice theory:

- Common methods for designing surveys and experiments are **never rationally obligatory**, and they are often **rationally prohibited** (or **irrational** for short).
- Some common statistical estimators and hypothesis testing procedures are simply irrational.

Today's Focus:

- Randomized designs (e.g., random sampling and randomized experiments)
- Maximal confirmation estimators (e.g., maximum likelihood estimator)





Question: What is a maximal confirmation estimator (MCE)?

Answer:

- Theories of evidence rank which hypotheses are supported by the data.
  - **Example:** likelihoodism says a piece of evidence E favors hypothesis θ_1 over θ_2 if $P(E|\theta_1) > P(E|\theta_2)$.
- The MCE for a theory of evidence says that we should endorse the hypothesis best supported by the evidence when such a hypothesis exists and is unique.
  - **Example:** The maximum likelihood estimator (MLE) selects the unique hypothesis θ such that $P(E|\theta) > P(E|\eta)$ for all $\eta \neq \theta$ (if such a hypothesis θ exists and is unique!)

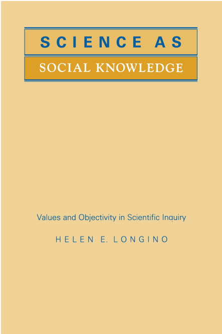
Slightly more precisely than last time:

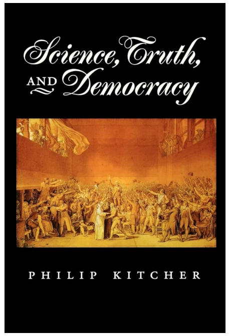
Thesis 1:

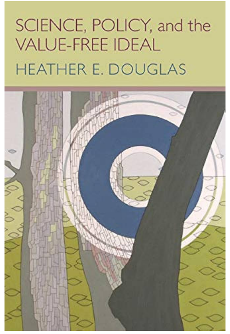
- Randomized designs (e.g., random sampling and randomized experiments) are never rationally obligatory, and they are often rationally forbidden.
- Maximal confirmation estimators (e.g., the maximum likelihood estimator) are irrational in rather simple settings.

Thesis 2: Frameworks for evaluating *collective* decisions (e.g., social choice theory) allow us to justify those same methods for survey and experimental design.

More work is needed to assess whether they plausibly justify the maximal confirmation estimators, except in a somewhat trivial way that I'll discuss.



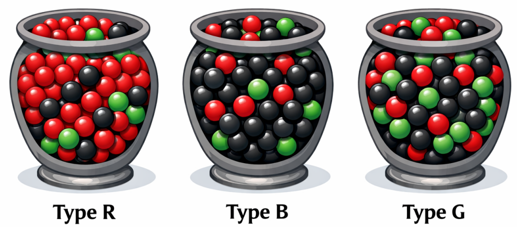




[Longino, 1990, 2001], [Kitcher, 2003] , [Douglas, 2009]

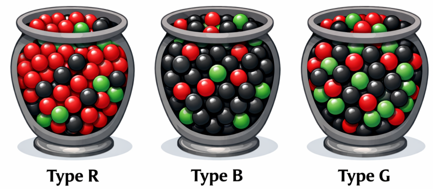
Orthodox Decision Theory

	Type R	Type B	Type G
Red	80	10	40
Black	10	80	40
Green	10	10	20



Task: You must bet what color ball will be drawn from an Acme Urn. You win $100 if you guess correctly. Sadly, you don't know which type of Acme Urn will be used for the drawing.

	Type R	Type B	Type G
Red	80	10	40
Black	10	80	40
Green	10	10	20



	Type R	Type B	Type G
Bet Red	$80	$10	$40
Bet Black	$10	$80	$40
Bet Green	$10	$10	$20

Question: Which color(s) would it be **irrational** to bet on? And why?

	Type R	Type B	Type G
Bet Red	$80	$10	$40
Bet Black	$10	$80	$40
Bet Green	$10	$10	$20

Do **not** bet on Green.

- Betting on Red **dominates** betting on Green.
- I.e., No matter the type of Urn used for the drawing, betting on red has a higher expected value.
- We say betting on Green is **inadmissible** in a case like this.

	Type R	Type B	Type G
Bet Red	$80	$10	$40
Bet Black	$10	$80	$40

Question: But should you bet on Red or on Black?

Answer: Intuitively, the answer depends on how confident that the Urn is of Type R vs Type B.

That intuitive answer is a consequence of Bayesian expected utility theory.

	Type R	Type B	Type G
Bet Red	$80	$10	$40
Bet Black	$10	$80	$40

According to expected utility theory, an agent is rational if she behaves as if she followed **three steps**

Expected Utility

	Type R ($P(R)=1/2$)	Type B ($P(B)=1/5$)	Type G ($P(G)=3/10$)
Bet Red	$80	$10	$40
Bet Black	$10	$80	$40

Step 1: Assigns probabilities to the three types of urns, representing how confident she is that said urn will be used in the drawing.

Expected Utility

	Type R ($P(R)=1/2$)	Type B ($P(B)=1/5$)	Type G ($P(G)=3/10$)	**Expected Value**
Bet Red	\$80	\$10	\$40	\$54
Bet Black	\$10	\$80	\$40	\$23

Step 2: Calculates the **subjective/Bayesian** expected utility of the various actions:

- The value of betting on Red equals

$$\$54 = \frac{1}{2} \cdot \$80 + \frac{1}{5} \cdot \$10 + \frac{3}{10} \cdot \$40$$

- The value of betting on Black equals

$$\$23 = \frac{1}{2} \cdot \$10 + \frac{1}{5} \cdot \$80 + \frac{3}{10} \cdot \$40$$

Expected Utility

	Type R ($P(R)=1/2$)	Type B ($P(B)=1/5$)	Type G ($P(G)=3/10$)	**Expected Value**
Bet Red	\$80	\$10	\$40	\$54
Bet Black	\$10	\$80	\$40	\$23

Step 3: Chooses the act/bet that maximizes her expected earnings (Here, Red)

Expected Utility of Mixed Strategies

	Type R ($P(R)=1/2$)	Type B ($P(B)=1/5$)	Type G ($P(G)=3/10$)	**Expected Value**
Bet Red	\$80	\$10	\$40	\$54
Bet Black	\$10	\$80	\$40	\$23
Random bet	\$45	\$45	\$40	\$77/2

In orthodox decision theory, the expected utility of a randomized strategy is a (weighted) average of the expected utilities of "pure"/non-randomized acts.

Averages are never greater than maxima.

Conclusion:

- Randomized acts never uniquely maximize expected utility. Hence, they are never **obligatory** according to orthodox (Bayesian) expected utility theory.
- They are **permissible** if and only if you randomize only among non-randomized acts that are all optimal.

Even more precisely than last time:

Thesis 1:

- For all theories of evidence I know, the MCE is inadmissible in some rather mundane setting.
- From the orthodox decision theoretic view, scientists' randomized designs are just randomized acts. Thus, randomized designs are **never rationally obligatory**, and they are permissible only if your randomize among optimal deterministic designs.
  - But most scientists think randomization is **required**!

Question: Wait. How do we apply the decision-theoretic apparatus to experimental design and statistical estimation?

Experimental Design, Estimation, and Decision Theory

	Type R	Type B	Type G
Bet Red	\$80	\$10	\$40
Bet Black	\$10	\$80	\$40

A decision problem contains three components:

- States of the world that are beyond the decision-maker's control (e.g., the type of urn used in the problem),
- Acts within the decision-maker's control (e.g., which bet to make), and
- Outcomes, whose expected value can be quantified.
  - Statisticians typically talk about "loss" rather than "utility", and they call those expected loss **risk**.

Example: Acme Inc. makes coins that have biases between $1/4$ and $3/4$ [The bias of a coin is the probability that it lands heads when tossed]. You have an Acme coin, but you don't know its bias. You get to flip the coin once, and imagine your guess will be scored by the squared distance to the true bias.

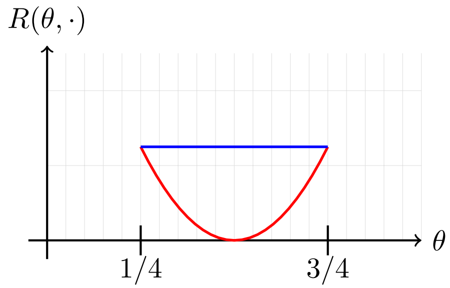
	$\theta = 1/4$	$\theta = 1/2$	$\theta = 3/4$
δ_1	$R(1/4, \delta_1)$	...	$R(3/4, \delta_1)$
δ_2	$R(1/4, \delta_2)$	...	$R(3/4, \delta_2)$

- State = A possible value for the unknown bias of the coin, i.e., a number between $1/4$ and $3/4$,
- Act = An estimator δ, which is a function specifying your planned response $\delta(H)$ if Heads/H is observed, and $\delta(T)$ if Tails/T is observed.
- Outcomes = The risk (expected loss) $R(\theta, \delta)$ of your estimator, where you lose $(\theta - r)^2$ if you guess the bias equals r when in reality it is θ.

	$\theta = 1/4$	$\theta = 1/2$	$\theta = 3/4$
δ_1	$1/16$	0	$1/16$
δ_2	$1/16$	$1/16$	$1/16$

Example: Just like in the urn example, we can calculate the risk of various estimators in the coin case and put the results in a decision table like the above.

- Above, δ_1 is the estimator that conjectures the bias is $1/2$ regardless of whether Heads or Tails is observed.
- δ_2 conjectures the bias is $3/4$ if flip lands Heads and conjectures $1/4$ otherwise.
- Notice that δ_2 seems to be dominated by δ_1. Although the table above shows only three states, that dominance holds across all biases ...



- We can plot the risk of the two estimators for all possible biases of the coin to confirm that δ_1 dominates δ_2.
  - In general, the risk of an estimator δ in a state θ equals:

$$R(\theta, \delta) = \theta \cdot (\delta(H) - \theta)^2 + (1 - \theta) \cdot (\delta(T) - \theta)^2$$

- Important: Notice that δ_2 is the maximum likelihood estimator (MLE)! So this example shows that the MLE is inadmissible in a very simple setting.

Objection: This a weird artifact of using squared-error as your loss function.

Response: No, it's not ...

Theorem

Suppose Θ is a finite set of hypotheses and L is a loss function taking at least three distinct values. Then there is some (conceptually possible) experiment in which the MLE is inadmissible with respect to the loss function L.

Confirmation Measure	Proponent
$P(H\|E) - P(H)$	Carnap [1950]
$P(H\|E) - P(H\|\neg E)$	Christensen [1999]
$P(E\|H) - P(E)$	Mortimer [1988]
$P(E\|H) - P(E\|\neg H)$	Nozick [1981]
$P(E\&H) - P(E) \cdot P(H)$	Carnap [1950]
$\frac{P(H\|E)}{P(H)} - 1$	Finch [1960]
$1 - \frac{P(\neg H\|E)}{P(\neg H)}$	Rips [2001]
$\frac{P(E\|H) - P(E\|\neg H)}{P(E\|H) + P(E\|\neg H)}$	Kemeny and Oppenheim [1952]

Table: Confirmation Measures

Objection: Likelihoodism is a bad theory of evidence. There are lots of measures of confirmation that we could used instead!

Response: We can prove a similar inadmissibility theorem for the MCEs for both of Carnap's measures and for Christensen's measure. Give us a few months and we'll have inadmissibility results for the rest of the confirmation measures in the list.

Philosophical Upshot: Scientists seem to think that they should follow the evidence. But from the standpoint of orthodox decision theory, no matter your loss function and no matter your theory of evidence, following the evidence is sometimes irrational.

- In the coin case, you passively observe the data available.
- But sometimes we need to choose what data to gather.
  - If we're sampling from a large population, the acts available to the researcher are pairs (δ, S), where δ is an estimator and S is a sample.
  - If we're conducting an experiment and dividing subjects into treatment and control groups, then an act is a triple (δ, T, C), where δ is an estimator and T is the set of units assigned to treatment, and C is the units assigned to control.
- No matter what, the choice to use a randomized design (random sampling or randomized experiment) amounts to the use of a randomized/mixed strategy, which is never rationally obligatory.

Science as Social Choice

A bit more precisely than last time:

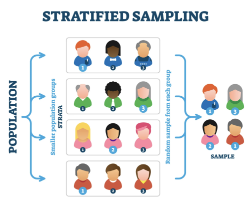
Thesis 2:

- According to many frameworks for *collective* decisions (e.g., social choice theory and bargaining theory), randomized designs are often the **unique/required** collective choice.
- Maximal confirmation estimators are often rationally permissible, though in a somewhat trivial way that I'll discuss.

Social choice and bargaining theory help to address questions about **fair division**.

When a good is **divisible** (like a cake), social choice and bargaining theories often require dividing the good evenly.

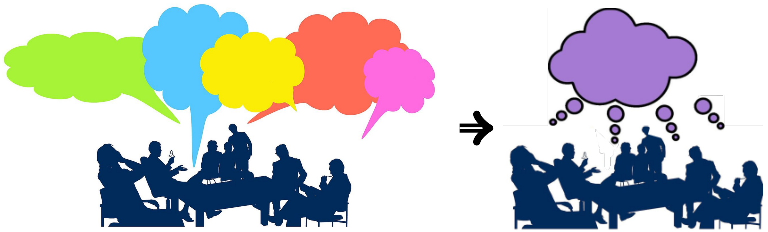
When a good is **indivisible**, various social choice and bargaining theories recommend random allocation.



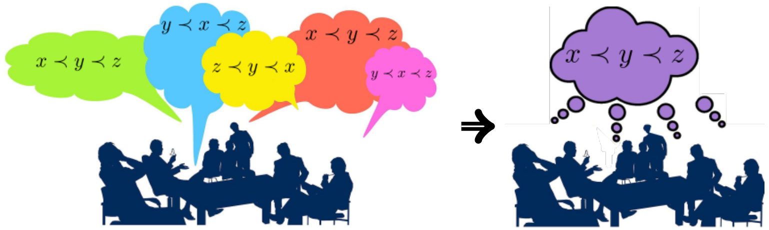
Analogy:

- If different researchers have diverse preferences about which **strata** to include in a sample (e.g., some prefer overrepresenting women, some prefer overrepresenting men), then stratifying is like "fairly dividing" the sample. Similar remarks apply to blocking.
- If researchers have different preferences about which **units** to include, then random sampling is like allocating an indivisible good.

Social Choice of Randomized Designs



Social choice theory = How individual attitudes are and should be related to collective decisions.



Most Common Application: How individual preference orderings should be related to collective preference orderings.

Our Application

- "Voters" = Those affected by a research project.
- Ranked Options =
  - Survey designs (e.g., samples, or sampling methods) or
  - Experimental designs (e.g.,, methods for assigning units to treatment and control groups)

Important: Social choice $\neq$ voting.

I am not advocating that researchers and trial participants vote on designs.

Claim: If those who will be affected by a study have sufficiently **diverse** preferences about which survey design or exprimental design to implement, then fair social rules require randomized designs.

Here's a toy model to explain the argument.

Social Choice

Adel	Bianca	Conor	Social Choice
$A \succ B$	$B \succ C$	$C \succ A$	(P_A, P_B, P_C)

What should the probabilities $P_A, P_B,$, and P_C be? In particular, when are all those probabilities (a) non-zero and (b) equal to one another?

Characteristics of Democratic Decision-Making

Democratic decision-making is often thought to be constrained by at least two principles:

1. Voters should be treated "equally."
   - E.g., "One person one vote."
2. Options should be treated equally.
   - E.g., Incumbents should not be chosen by default when a race is close. Republicans should not be automatically favored over Democrats, or vice versa. Etc.

Slogans

Anonymity: Voters are treated equally.

Neutrality: Options are treated equally.

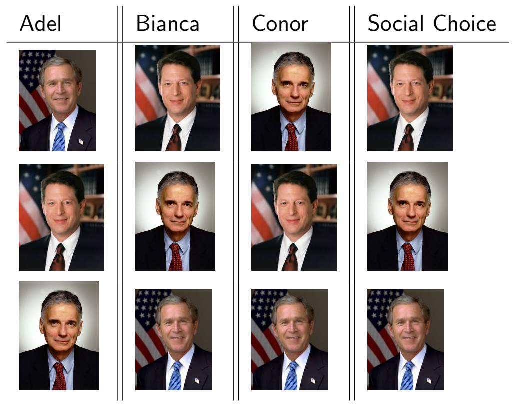
Formalizing Equality

- There are many ways of trying to capture the idea of “treating equally.”
- To capture one such idea, the notion of a **permutation** is extremely helpful ...

Permutations

- As you all know, a permutation is a **rearrangement**.

Anonymity



Anonymity

Bianca	*Conor*	*Adel*	Social Choice

Anonymity: If we permute the voters' names on the ballots, then the social choice/ranking remains the same.

Anonymity

Adel	Bianca	Conor	Social Choice
$A \succ B$	$B \succ C$	$C \succ A$	$(\frac{1}{2}, \frac{3}{8}, \frac{1}{8})$

Bianca	Conor	Adel	Social Choice
$A \succ B$	$B \succ C$	$C \succ A$	$(\frac{1}{2}, \frac{3}{8}, \frac{1}{8})$

Anonymity: If we permute the voters' names on the ballots, then probabilities assigned to the designs remains the same.

Neutrality

Adel	Bianca	Conor	Social Choice

Neutrality

Adel	Bianca	Conor	Social Choice

Neutrality: If we swap two options in every voters' ranking, then those options are swapped in the social choice.

Neutrality

Adel	Bianca	Conor	Social Choice
$A \succ B$	$B \succ C$	$C \succ A$	$(\frac{1}{2}, \frac{3}{8}, \frac{1}{8})$

Adel	Bianca	Conor	Social Choice
$\mathbf{B} \succ \mathbf{C}$	$\mathbf{C} \succ \mathbf{A}$	$\mathbf{A} \succ \mathbf{B}$	$(\frac{\mathbf{3}}{\mathbf{8}}, \frac{\mathbf{1}}{\mathbf{8}}, \frac{\mathbf{1}}{\mathbf{2}})$

Neutrality: If we permute the options in every voters' ranking, then the probabilities of those options are permuted in the social choice.

Disclaimer

Disclaimer:

- Neither Anonymity nor Neutrality is plausible as a constraint on choice of experimental design.
- Neutrality is way too strong, especially in the physical sciences where experiments do not involve sentient subjects.

However: There are plausible related principles. So I'll explore the consequences of the two.

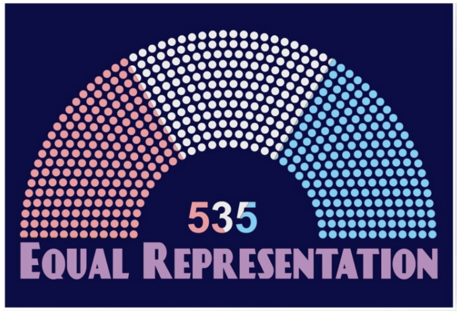
Randomization from Anonymity and Neutrality

Adel	Bianca	Conor	Social Choice
$A \succ B$	$B \succ C$	$C \succ A$	(P_A, P_B, P_C)

Fact: In this toy example, Anonymity and Neutrality entail that $P_A = P_B = P_C$.

Question: When do Anonymity and Neutrality entail that some set of options/samples should be assigned equal probability?

Answer: It suffices for two conditions to hold ...



- Call a group of voters who have the same preferences an interest group.
- **Condition 1 (Equal Representation):** All interest groups are the same size (i.e., contain equal numbers of voters).
  - This is trivially satisfied when everyone has different preferences.

- **Condition 2 (Diverse Interests):** For any voter and any arbitrary way of permuting the options, there is a second voter whose preferences match the permuted ranking of the first voter.

Randomized Experiments

Example: 100 people with cancer sign up for a clinical trial. Each prefers to be given the novel treatment, and no participant has preferences about who else should be in the treatment group. If researchers limit the size of the treatment group to n people, where $n < 100$, then the above two conditions are met and choosing a treatment group of size n at random is required by Anonymity and Neutrality.

Question: What about *researchers*' preferences?

Answer: If researchers' preferences are diverse in the appropriate way, then same argument applies.

- I doubt that researchers' preferences (esp. in the natural sciences) typically **are** diverse in the necessary way to justify the most common randomized designs.
- I suspect that our standards for experimental justification require that designs balance the preferences of all **reasonable** researchers, which might include future or even hypothetical researchers.
- If you doubt that reasonable researchers could have sufficiently diverse preferences to justify common randomized designs, then you haven't uncovered a problem with my argument; you've uncovered a problem with current prescriptions to randomize. Why?

- When my argument doesn't justify typical forms of randomization, existing practices **require** either
  - Randomizing among a collection of designs that all researchers judge to be equally good, OR
  - Assigning positive probability to choosing a sample (or to performing an intervention) that no one prefers.
- In the former case, randomizing seems permissible, not obligatory. In the latter, it seems impermissible.

Social choice of maximal confirmation estimators?

Question: Can we view MCEs as a "fair" collective choice (among various estimators) given researchers with diverse preferences and/or beliefs?

Answer: Yes. BUT

- Here, we need **bargaining theory**, which allows us to consider the cardinal/numerical strength of researchers preferences, AND
- I don't think this line of argument is actually what justifies MCEs, except in a rather trivial way . . .

Estimates vs. Experiments

Estimates $\neq$ Experiments

- An experimenter can publish multiple estimates. In general, she can't conduct multiple experiments (e.g., because she can't assign the same patient to both treatment and control).
- Estimates, therefore, should be viewed as **data summaries** that many researchers can use for various purposes.
- An estimate, therefore, should be how well it balances the tradeoff between (i) *compressing* the data for use, and (ii) communicating enough information to allow researchers with diverse interests to use the data.

- A sufficient statistic is a data summary that contains the "same information" as the full sample.
  - In the Bayesian sense, a summary $S(X)$ of a full data set X is sufficient if $P(H|X) = P(H|S(X))$ for all hypotheses H and all prior probability functions P (i.e., no matter your degrees of belief).
  - **Example:** If you flip a coin 10 times, say, the number of heads is a sufficient statistic.
- A minimal sufficient statistic is the shortest sufficient statistic, if one exists.
- In many settings, a maximum likelihood estimate is a minimal sufficient statistic.

Future Work: Lots of other statistical procedures (e.g., unbiased estimators) do not minimize expected loss relative to priors and loss functions that researchers might plausibly endorse.

When are such estimators like minimal sufficient statistics? When can they be seen as "fair" collective choices according to various principles of social choice and bargaining theories?

Thanks.

Questions? Comments? Devastating objections?

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, second edition.

Christensen, D. (1999). Measuring confirmation. *The Journal of Philosophy*, 96(9):437–461.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, Pittsburgh (Pa.), first edition edition.

Finch, H. A. (1960). Confirming power of observations metricized for decisions among hypotheses, Part II. *Philosophy of Science*, 27(4):391–404.

Kemeny, J. G. and Oppenheim, P. (1952). Degree of factual support. *Philosophy of Science*, 19(4):307–324.

Kitcher, P. (2003). *Science, Truth, and Democracy*. Oxford University Press, Oxford.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Longino, H. E. (2001). *The Fate of Knowledge*. Princeton University Press, Princeton, N.J.

Mortimer, H. (1988). *The logic of induction*. Halsted Press.

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Rips, L. J. (2001). Two Kinds of Reasoning. *Psychological Science*, 12(2):129–134.