A Statistical Guide for the Ethically Perplexed

Lawrence Hubert and Howard Wainer

Department of Psychology
The University of Illinois
and
National Board of Medical Examiners

33rd Annual National Institute on the Teaching of Psychology
January 3-6, 2011

Two Quotes From Notable Women

The true foundation of theology is to ascertain the character of God. It is by the art of Statistics that law in the social sphere can be ascertained and codified, and certain aspects of the character of God thereby revealed. The study of Statistics is thus a religious service.
– Florence Nightingale

Math class is tough. Let's go shopping!
– Barbie

But First, A Few Words About PowerPoint

We note the observations of Edward Tufte on the ubiquity of PowerPoint (PP) for presenting quantitative data, and the degradation it produces in our ability to communicate (*The Cognitive Style of PowerPoint*, Tufte, 2006, p. 26, his italics):

"*The PP slide format has the worst signal/noise ratio of any known method of communication on paper or computer screen.* Extending PowerPoint to embrace paper and internet screens pollutes those display methods."

Generally, PowerPoint is poor at presenting statistical evidence, and is no replacement for more detailed technical reports, data handouts, and the like. It is now part of our "pitch culture," where, for example, we are sold on what drugs to take, but are not provided with the type of detailed numerical evidence we should have for an informed decision about benefits and risks.

In commenting on the incredible obscuration of important data that surrounded the use of PowerPoint-type presentation in the briefings of the Rogers Commission investigating the first Shuttle accident of Challenger in 1986, Richard Feynman noted (reported in Tufte, 2006, p. 17):

"Then we learned about 'bullets' – little black circles in front of phrases that were supposed to summarize things. There was one after another of these little goddamn bullets in our briefing books and on slides."

PowerPoint is also detrimental to the decision making processes within the ongoing war effort in Afghanistan – see the "page-one, above-the-fold" article: *We Have Met the Enemy and He is PowerPoint* (Elisabeth Bumiller, *The New York Times*, April 26, 2010)

A Touch of Probability Theory

We speak of events represented by capital letters, such as A, and the probability of the event as some number in the range from 0 to 1, written as $P(A)$.

Two events are independent whenever the probability of the joint event, $P(A \text{ and } B)$, factors as the product of the individual probabilities, $P(A)P(B)$.

The definition of conditional probability plays a central role in all our uses of probability theory; in fact, most misapplications of statistical/probabilistic reasoning involve confusions of some sort regarding conditional probabilities.

Conditional Probability

The conditional probability of some event A given that B has already occurred, denoted $P(A|B)$, is defined as

$$P(A|B) = P(A \text{ and } B)/P(B) \ .$$

When A and B are independent,

$$P(A|B) = P(A \text{ and } B)/P(B) = P(A)P(B)/P(B) = P(A) \ .$$

So, knowing that B has occurred does not alter the probability of A occurring.

If $P(A|B) > P(A)$, we will say that B is "facilitative" of A; when $P(A|B) < P(A)$, B is said to be "inhibitive" of A.

Facilitation and Inhibition

Suppose A is the event of receiving a basketball scholarship; B, the event of being seven feet tall; and C, the event of being five feet tall.

B is facilitative of A: $P(A|B) > P(A)$

C is inhibitive of A: $P(A|C) < P(A)$

The size and sign of the difference between $P(A|B)$ and $P(A)$ is a raw descriptive measure of how much the occurrence of B is associated with an increased or decreased probability of A, with a value of zero corresponding to statistical independence.

The Case of Sally Clark

The obverse idea that if two events are *not* independent, the joint probability cannot be generated by a simple product of the individual probabilities, lead directly to "one of the great miscarriages of justice in modern British legal history."

The case is that of Sally Clark, a British solicitor convicted and sentenced to life imprisonment for the murder of her two sons in 1999.

The purveyor of statistical misinformation in this case was Sir Roy Meadow, famous for Meadow's Law: "one sudden infant death is a tragedy, two is suspicious, and three is murder."

Roy Meadow testified that the probability of two children from an affluent family suffering sudden infant death syndrome was 1 in 73 million. He obtained this number by squaring 1 in 8500 – supposedly, this latter value is the probability of one 'cot' death in similar circumstances.

We quote from a Press Release from the Royal Statistical Society (the whole press release is in your packet):

"In the recent highly-publicised case of R v. Sally Clark, a medical expert witness drew on published studies to obtain a figure for the frequency of sudden infant death syndrome (SIDS, or 'cot death') in families having some of the characteristics of the defendant's family. He went on to square this figure to obtain a value of 1 in 73 million for the frequency of two cases of SIDS in such a family.

This approach is, in general, statistically invalid. It would only be valid if SIDS cases arose independently within families, an assumption that would need to be justified empirically."

Sally Clark was convicted not only because the number 1 in 73 million was small, but it was then misinterpreted by the court and jury as the probability that she was innocent.

This latter confusion is referred to as the "Prosecutor's Fallacy."

Let A be the event of Sally Clark's innocence –

Let B be the event of two 'cot' deaths –

Assuming that the number of 1 in 73 million is correct (which it is not), it was explicitly meant to be for $P(B|A)$ – the probability of two 'cot'deaths if Sally Clark were innocent. This value was then misinterpreted to be for $P(A|B)$ – the probability of innocence given that two 'cot' deaths have occurred.

Quoting again from the Royal Statistical Society Press Release:

"Aside from its invalidity, figures such as the 1 in 73 million are very easily misinterpreted. Some press reports at the time stated that this was the chance that the deaths of Sally Clark's two children were accidental. This (mis-)interpretation is a serious error of logic known as the Prosecutor's Fallacy."

Where Are They Today

Sally Clark's conviction was overturned in 2003, and she was released from prison. Sally Clark died of acute alcohol poisoning in her home, four years later in 2007, at the age of 42.

Roy Meadow (1933 –) is still an active British pediatrician. He rose to fame for his 1977 academic paper in *Lancet* on Munchausen Syndrome by Proxy (MSbP) – he is the person who coined the name.

He spent his whole career crusading and testifying against parents, especially mothers, who supposedly wilfully harmed or killed their children.

We quote from Lord Howe, the opposition spokesman for health, speaking in the House of Lords on MSbP (February of 2003):

... “one of the most pernicious and ill-founded theories to have gained currency in childcare and social services in the past 10 to 15 years. It is a theory without science. There is no body of peer-reviewed research to underpin MSbP. It rests instead on the assertions of its inventor. When challenged to produce his research papers to justify his original findings, the inventor of MSbP stated, if you please, that he had destroyed them.”

A Bonus Slide on the Prosecutor's Fallacy

The Prosecutor's Fallacy (also called the Fallacy of the Transposed Conditional [Probability]), is responsible for the common misinterpretation that a p-value is the "probability that the null hypothesis is true."

The p-value is the probability of seeing some result as extreme or more than you actually did, when the null hypothesis is true – denoted as $P(\text{data}|H_o)$

To confuse the later with $P(H_o|\text{data})$, is to commit the Fallacy of the Transposed Conditional – and to say that a p-value is the probability of the null hypothesis being true given the data.

Screening for Rare Events: Bayes Theorem

To understand the implications for screening for rare events (such as mammograms for breast cancer or using the P.S.A. for prostate cancer), we have to ultimately reach what is called Bayes Theorem. But before we get that far, we need to introduce some terms, and will do so in a generic screening context.

Suppose we have a test that assesses some relatively rare quantity (e.g., disease, ability, talent, terrorism propensity, drug/steriod usage, antibody presence, being a liar [where the test is a polygraph], fetal haemoglobin, and so forth).

Let B be the event that the test says the person has "it," whatever that may be; A is the event that the person really does have "it."

Two "reliabilities" are needed (note: a "bar" over an event denotes the complement):

a) the probability, $P(B|A)$, that the test is positive if the person has "it"; this is called the *sensitivity* of the test;

b) the probability, $P(\bar{B}|\bar{A})$, that the test is negative if the person doesn't have "it"; this is the *specificity* of the test.

The conditional probability to be used (eventually) in the denominator of Bayes rule, $P(B|\bar{A})$, is merely $1 - P(\bar{B}|\bar{A})$, and is the probability of a "false positive."

The quantity of prime interest, called the *positive predictive value* (PPV), is the probability that a person has "it" given that the test says so, $P(A|B)$, and is obtainable from Bayes rule using the specificity, sensitivity, and prior probability, $P(A)$:

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))} .$$

To understand how well the test does, the facilitative effect of B on A needs interpretation, i.e., a comparison of $P(A|B)$ to $P(A)$, plus an absolute assessment of the size of $P(A|B)$ by itself.

The situation is usually dismal whenever $P(A)$ is small (i.e., screening for a relatively rare quantity), and the sensitivity and specificity are not perfect. Although $P(A|B)$ will generally be greater than $P(A)$, and thus, B facilitative of A, the absolute size of $P(A|B)$ is commonly so small that the value of the screening may be questionable.

A Mammogram Example

As an example, consider the efficacy of mammograms in detecting breast cancer. In the United States, 180,000 women are found to have breast cancer each year from among the 33.5 million women who annually have a mammogram.

Thus, the probability of a tumor is 180,000/33,500,000 = .0054. Mammograms are no more than 90% accurate, implying

$$P(\text{positive mammogram} \mid \text{tumor}) = .90;$$

$$P(\text{negative mammogram} \mid \text{no tumor}) = .90.$$

Because we do not know whether a tumor is present, all we know is whether the test is positive, Bayes theorem must be used to calculate the probability we really care about, the PPV.

Using Bayes rule, the PPV of the test is .047:

$$P(\text{tumor} \mid \text{positive mammogram}) =$$

$$\frac{.90(.0054)}{.90(.0054) + .10(.9946)} = .047 \ ,$$

which is obviously greater than the prior probability of .0054, but still very small in magnitude, i.e., more than 95% of the positive tests that arise turn out to be incorrect.

The Mammogram Example Using Natural Frequencies

Gigerenzer and colleagues have argued for the use of "natural frequencies" rather than actual probabilities substituted into Bayes rule. Based on an assumed population of 10,000, we have the following 2×2 table:

	tumor	no tumor	
+ mammogram	49	995	1044
− mammogram	5	8951	8956
	54	9946	10,000

The PPV is then simply $49/1044 = .047$, using the frequency value of 49 for the cell (+ mammogram, tumor) and the + mammogram row sum of 1044.

Two References on Screening

There are two items in your packet that discuss cancer screening, one for males and one for females:

The Great Prostate Mistake (Richard J. Ablin, *The New York Times*, March 10, 2010)

The Breast Brouhaha (Gail Collins, *The New York Times*, November 19, 2009)

Bonus Slides on Signal Detection Theory

In the terminology of signal detection theory and the general problem of yes/no diagnostic decisions, a plot of sensitivity (true positive probability) (on the y-axis) against 1− specificity (on the x-axis) as the decision boundary criterion point varies, is called an ROC curve (for Receiver Operating Characteristic).

This ROC terminology originates from World War II when the issue was to detect enemy planes by radar from the noise generated by random interference.

The ROC curve is bowed from the origin of (0, 0) at the lower left corner to (1.0, 1.0) at the upper right – it indicates the trade-off between increasing the probability of true positives and the increase of false positives. .

Generally, the adequacy of a particular diagnostic decision strategy is measured by the area under the ROC curve, with areas closer to 1.0 being better, i.e., steeper bowed curves hugging the left wall and the top border of the square box.

For a more comprehensive introduction to diagnostic processes, the inaugural issue of *Psychological Science in the Public Interest* contains a review article by Swets, Dawes, and Monahan with the descriptive title, *Psychological Science Can Improve Diagnostic Decisions* (2000, *1*, 1–26).

Complete Enumeration Versus Sampling in the Census

By a small sample we may judge the whole piece.
– Cervantes (1547 – 1616)

The Law of Large Numbers states that as the sample size increases, a sample mean converges to the population mean.

This implies that to get a sense of the value for a population mean, we only need to take a sample,

Moreover, we can make our estimate as precise as necessary by just increasing sample size. And there are fairly easy formulas for indicating what sample size is needed for a given desired precision – the latter are based on the familiar standard error formula for a sample mean.

Stated in other words for the U.S. Census, complete enumeration (as required by the Constitution) is really unnecessary (and an unnecessary expense as well).

Sampling would also help alleviate the decennial issues of dealing with the 'undercount'.

Your packet contains a short article by David Stout (*The New York Times*, April 3, 2009): *Obama's Census Choice Unsettles Republicans*.

Voodoo Correlations in Social Neuroscience

A recent article (Vul et al. 2009) in a journal from the Association for Psychological Science, *Perspectives on Psychological Science*, has the intriguing title of *Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition* (renamed from the earlier and more controversial, *Voodoo Correlations in Social Neuroscience*).

These authors comment on the extremely high (e.g., $> .8$) correlations reported in the literature between brain activation and personality measures, and point out the fallaciousness of how they were obtained.

Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) are reported. It is tautological that these correlations selected for size must then be large in their average value.

With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best. Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, are inappropriate.

In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

In your packet is an amusing critique of fMRI studies that fail to correct for multiple comparisons and control false positives. It involves the scan of a dead salmon's brain and its response to human emotions (*Trawling the Brain*, Laura Sanders, December 19, 2009, *Science News*).

Doubling-Dipping More Generally

Whenever coincidences are culled or "hot spots" identified from some search of available information, the probabilities that are then regenerated for these situations may not be valid. There are several ways of saying this:

when some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion. In Bayesian terms, you don't get the posterior probabilities from the same information that gave you the prior probabilities.

Alternatively said, it makes no sense to do formal hypothesis assessment (by finding estimated probabilities) when the data themselves have suggested the hypothesis in the first place. Some cross-validation strategy is necessary, e.g., collecting independent data.

Generally, when some process of search or optimization has been used to identify an unusual situation:
– when a "good" regression equation is found through a step-wise procedure;
– when data are "mined" and unusual patterns identified; when DNA databases are searched for "cold-hits" against evidence left at a crime scene;
– when geographic "hot spots" are identified for, say, some particularly unusual cancer;
– when the whole human genome is searched for clues to common diseases;
the same methods for assigning probabilities before the particular situation was identified, are generally no longer appropriate post-hoc.

A particularly problematic case of culling or locating "hot spots," is that of residential cancer-cluster identification. A readable account is given by Atul Gawande, *The Cancer-Cluster Myth*, *The New Yorker*, February 8, 1999.

For the probability issues that arise in searching the whole human genome for clues to some condition, see *Nabbing Suspicious SNPS: Scientists Search the Whole Genome for Clues to Common Diseases* (Regina Nuzzo, *Science News*, June 21, 2008.

Seinfeld Double-Dipping

As a humorous way of relating this admonition that it is always unwise to double-dip, both statistically and from a communal dip bowl at a party, witness the following dialogue from a Seinfeld episode involving George Constanza and Timmy, the brother of George's then current girlfriend:

[George, attending a wake, takes a large tortilla chip, dips it into a bowl of what appears to be sour cream, takes a bite, dips it into the bowl again, and then eats the remainder of the chip.]

Timmy: What are you doing?
George: What?
Timmy: Did, did you just double dip that chip?
George: Excuse me?
Timmy: You double dipped a chip!
George: Double dipped? What, what, what are you talking about?

Timmy: You dipped a chip. You took a bite. And you dipped again.
George: So?
Timmy: That's like putting your whole mouth right in the dip. From now on, when you take a chip, just take one dip and end it.
George: Well, I'm sorry, Timmy, but I don't dip that way.
Timmy: Oh, you don't, huh?
George: You dip the way you want to dip. I'll dip the way I want to dip.

[George grabs another chip, dips it, takes a bite and begins to reach for the dip as Timmy grabs his hand.]

Timmy: Gimme the chip!

[An all-out brawl breaks out between George and Timmy.]

Actuarial versus Clinical Prediction

Paul Meehl in his classic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction.

Individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it. Combining such selected information in some more mechanical manner will generally do better than the person choosing such information in the first place.

If we formally model the predictions of experts using the same chosen information, we can generally do better than the experts themselves. Such formal representations of what a judge does, are called "paramorphic."

In an influential review paper, Dawes (1979) discussed what he called proper and improper linear models, and argued for the "robust beauty of improper linear models."

A proper linear model is one obtained by some optimization process, usually least-squares. Improper linear models are not "optimal" in this latter sense, and typically have their weighting structures chosen by a simple mechanism, e.g., random or unit weighting.

Again, improper linear models generally outperform clinical prediction, but even more surprisingly, improper models typically outperform proper models in cross-validation.

The idea that statistical optimality may not lead to the best predictions, seems counterintuitive, but as argued by Roberts and Pashler (2000), just the achievement of a good fit to observations does not necessarily mean we have found a good model. In fact, because of the overfitting of observations, choosing the model with the absolute best fit is apt to result in poorer predictions.

The more flexible the model, the more likely it is to capture not only the underlying pattern but unsystematic patterns such as noise. A single general purpose tool with many adjustable parameters is prone to instability and greater prediction error.

An observation by John von Neumann is particulary germane: "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk."

This notion that "less-is-more" is difficult to get one's head around, but as Gigerenzer and others have argued, it is clear that simple heuristics can at times be more accurate than complex procedures (even though we won't go as far as Gigerenzer and colleagues in labeling this observation about simple heuristics, such as "take the best," as one of the major discoveries of the last decade).

All of the work emanating from the idea of the "robust beauty of improper linear models" *et sequelae* may force some reassessment of what the normative ideals of rationality might be. Most reduce to simple cautions about overfitting one's observations, and then hoping for better predictions because an emphasis has been placed on immediate optimality instead of the longer-run goal of cross-validation.

Henry A. Wallace and the Modeling of Expert Judgements

There are several interesting historical connections between Henry A. Wallace, one of Franklin D. Roosevelt's Vice-Presidents (1940–1944), and the formal (paramorphic) modeling of the prediction of experts, and applied statistics more generally.

Wallace wrote a paper (1923) in the *Journal of the American Society of Agronomy* (*13*, 300–304), entitled: *What Is In the Corn Judge's Mind*? The data used in this study were ratings of possible yield for some 500 ears of corn from a number of experienced corn judges.

In addition to the ratings, measurements were taken on each ear of corn over six variables: length of ear; circumference of ear; weight of kernel; filling of the kernel at the tip (of the kernel); blistering of kernel; starchiness.

Also, because all the ears were planted in 1916, one ear to a row, the actual yields for the ears were available as well.

For the outcome of this study, see the piece written by Hubert and Wainer in your packet – *Henry A. Wallace and the Modeling of Expert Judgements*.

I might just add that there is a lot more to say about Wallace: he ran for President under the Progressive Party in 1948; was a world-class Applied Statistician who made seminal contributions to computational statistics in the 1930s; and was one of the earliest civil rights activists in a position of political power.

Psychopathy and Construct Validation

As discussed by Cronbach and Meehl in their classic 1955 article, *Construct Validity in Psychological Tests* (*Psychological Bulletin*, *52*, 281–282), the most elusive form of validity to establish is construct validity. In lay terms, a validation process has to be put into place to argue effectively that we are really measuring the construct we think we are measuring.

A recent example of the difficulties inherent in construct validation has appeared in the popular media, and involves the notion of psychopathy – a personality disorder indicated by a pattern of lying, exploitativeness, heedlessness, arrogance, sexual promiscuity, low self-control and lack of empathy and remorse; and all of this combined with an ability to appear normal.

The usual diagnostic manuals (e.g., the *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*) do not include psychopathy as part of their classification scheme for personality disorders. In fact, the notion of psychopathy has been defined *de facto* by one specific 20-item instrument developed by Robert Hare, the Psychopathy Checklist–Revised.

The issue now being raised about the PCL-R is the degree to which criminal behavior is a component crucial to psychopathy. Or, as put by the Australian psychiatrist, John Ellard in 1998: "Why has this man done these terrible things? Because he is a psychopath. And how do you know that he is a psychopath? Because he has done these terrible things."

An article is in your packet from *The New York Times* by Benedict Carey, *Academic Battle Delays Publication by 3 Years* (June 11, 2010), traces the current sordid saga.

Simpson's Paradox

An example of Simpson's Paradox will be given using data on the differential imposition of a death sentence depending on the race of the defendant and the victim. These data are from twenty Florida counties during 1976-7:

Defendant	Death:Yes	Death:No
White	19 (12%)	141
Black	17 (10%)	149

Because 12% of White defendants receive the Death penalty and only 10% of Blacks, at this aggregate level there appears to be no bias against Blacks. But when the data are disaggregated, the situation appears to change:

Victim	Defendant	Death:Yes	Death:No
White	White	19 (13%)	132
White	Black	11 (17%)	52
Black	White	0 (0%)	9
Black	Black	6 (6%)	97

To summarize, when aggregated over victim race, there is a higher proportion of White defendants (12%) receiving the death penalty than Black defendants (10%), so apparently, there is a slight race bias against Whites. But when looking within the race of the victim, in both cases the Black defendant has the higher probability of receiving the death sentence compared to the White defendant (17% to 13% for White victims; 6% to 0% for Black victims).

A more recent study given in your packet is from *The New York Times*, Friday, April 20, 2001: Fox Butterfield, *Victims' Race Affects Decisions on Killers' Sentence, Study Finds*. There is also a short section from Hubert and Wainer on Simpson's Paradox.

Context and Framing in Data Presentation

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest*. One recent issue was from Gerd Gigerenzer and colleagues, entitled: *Helping Doctors and Patients Make Sense of Health Statistics*. It details some issues of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally.

We begin with a quote from Rudy Giuliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican Presidential nomination:

"I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it, in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine."

Not only did Giuliani not receive the Republican Presidential nomination, he was just plain wrong on survival chances for prostate cancer. The problem is a confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

five-year survival rate = (number of diagnosed patients alive after five years)/(number of diagnosed patients);

annual mortality rate = (number of people who die from a disease over one year)/(number in the group).

The inflation of a five-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed.

Moreover, such screening, particularly for cancers such as prostate, leads to an *overdiagnosis bias* — the detection of a pseudodisease that will never progress to cause symptoms in a patient's lifetime.

Besides inflating five-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (e.g., incontinence, impotence, and other health related problems).

A major area of concern in the clarity of reporting health statistics, is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter. We give examples that present the same information:

relative risk reduction — if you have this test every two years, it will reduce your chance of dying from the disease by about one third over the next ten years.

absolute risk reduction — if you have this test every two years, it will reduce your chance of dying from the disease from 3 in 1000 to 2 in 1000, over the next ten years.

We also have a useful variant on absolute risk reduction given by its reciprocal, the *number needed to treat* — if 1000 people have this test every two years, one person will be saved from dying from the disease every ten years.

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm.

It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically. The latter has been called "mismatched framing" by Gigerenzer and colleagues.

Your packet contains two items relevant to reporting data both from Hubert and Wainer: one discusses the Gigerenzer ideas in more detail, and the second gives the classic Tversky and Kahnemann example on the influence of framing in decision making.

Some Resources

Hubert, L., & Wainer, H. (2011). A statistical guide for the ethically perplexed. In A. T. Panter and S. K. Sterba (Eds.), *Handbook of Ethics in Quantitative Methodology* (pp. xxx–xxx). New York: Taylor & Francis.

Also, Hubert and Wainer are expanding this chapter into a book that will have a large collection of Supplementary Readings available, mostly from the New York media (*The New York Times*, *The New Yorker*, *The New York Review of Books*).