Chapter 1

Prediction

The race is not always to the swift nor the battle to the strong — but that's the way to lay your bets.

– Damon Runyon

The attempt to predict the values on some dependent variable by a function of independent variables is typically approached by simple or multiple regression, for one and more than one predictor, respectively. The most common combination rule is a linear function of the independent variables obtained by least-squares, i.e., the linear combination that minimizes the sum of the squared residuals between the actual values on the dependent variable and those predicted from the linear combination. In the case of simple regression, scatterplots again play a major role in assessing linearity of the relationship, the possible effects of outliers on the slope of the least-squares line, and the influence of individual objects in its calculation. Regression slopes, in contrast to the correlation, are neither scale invariant nor symmetric in the dependent and independent variables. One usually interprets the least-squares line as one of expecting, for each unit change in the independent variable, a regression slope change in the dependent variable.[1]

There are several topics in prediction that arise continually when we attempt to reason ethically with fallible multivariable data. We discuss five such areas in the subsections to follow: regression toward the mean; methods involved in using regression for prediction that incorporate corrections for unreliability; differential prediction effects in selection based on tests; interpreting and making inferences from regression weights; and the distinction between actuarial (statistical) and clinical prediction.

1.1 Regression Toward the Mean

Regression toward the mean is a phenomenon that will occur whenever dealing with (fallible) measures with a less-than-perfect correlation. The word "regression" was first used by Sir Francis Galton in his 1886 paper, *Regression Toward Mediocrity in Hereditary Stature*, where he showed that heights of children from very tall or short parents would regress toward mediocrity (i.e., toward the mean) — exceptional scores on one variable (parental height) would not be matched with such exceptionality on the second (child height). This observation is purely due to the fallibility for the various measures, and the concomitant lack of a perfect correlation between the heights of parents and their children.

Regression toward the mean is a ubiquitous phenomenon, and given the name "regressive fallacy" whenever cause is ascribed where none exists. Generally, interventions are undertaken if processes are at an extreme, e.g., a crackdown on speeding or drunk driving as fatalities spike; treatment groups formed from individuals who are seriously depressed; individuals selected because of extreme behaviors, both good or bad; and so on. In all such instances, whatever remediation is carried out will be followed by some lessened value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened; I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (i.e., the "sophomore jinx"); although he hadn't had a hit in some time, he was "due," and the coach played him; and on and on. More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness-of-fit or prediction we might calculate need to be cross-validated either on new data or by a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures. The degree of "shrinkage" we see in our measures based on this cross-validation, is an indication of the fallibility of our measures and the (in)adequacy of the

given sample sizes.

The misleading interpretive effects engendered by regression toward the mean are legion, particularly when we wish to interpret observational studies for some indication of causality. There is a continual violation of the old adage that "the rich get richer and the poor get poorer," in favor of "when you are at the top, the only way is down." Extreme scores are never quite as extreme as they first appear. Many of these regression artifacts are explicated in the cautionary source, *A Primer on Regression Artifacts* (Campbell and Kenny, 2002), including the various difficulties encountered in trying to equate intact groups by matching or analysis-of-covariance. Statistical equating creates the illusion but not the reality of equivalence. As summarized by Campbell and Kenny, "the failure to understand the likely direction of bias when statistical equating is used, is one of the most serious difficulties in contemporary data analysis."

There are a variety of phrases that seem to get attached whenever regression toward the mean is probably operative. We have the "winner's curse," where someone is chosen from a large pool (e.g., of job candidates), who then doesn't live up to expectation; or when we attribute some observed change to the operation of "spontaneous remission." As Campbell and Kenny note: "many a quack has made a good living from regression toward the mean." Or, when a change of diagnostic classification results upon repeat testing for an individual given subsequent one-on-one tutoring (after being placed, for example, in a remedial context). More personally, there is "editorial burnout" when someone is chosen to manage a prestigious journal at the apex of a career, and things go quickly downhill from that point forward.

1.2 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. The enormous amount of literature produced since the appearance of this seminal contribution, has uniformly supported this general observation; simi-

larly, so have the extensions suggested for combining data in ways other than by multiple regression, e.g., by much simpler unit weighting schemes (Wainer, 1976), or those using other prior weights. It appears that 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. This conclusion can be pushed further: 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."[2]

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. What seems to be the reason, is the notorious instability of regression weights with correlated predictor variables, even if sample sizes are very large. Generally, we know that simple averages are more reliable than individual observations, so it may not be so surprising that simple unit weights are likely to do better on cross-validation than those found by squeezing "optimality" out of a sample. Given that the *sine qua non* of any prediction system is its ability to cross-validate, the lesson may be obvious — statistical optimality with respect to a given sample may not be the best answer when we wish to predict well.

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 gen-

eral purpose tool with many adjustable parameters is prone to instability and greater prediction error as a result of high error variability. 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." More generally, this notion that "less-is-more" is difficult to get one's head around, but as Gigerenzer and others have argued (e.g., see Gigerenzer and Brighton, 2009), 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 Brighton [2009] in labeling this observation about simple heuristics, such as "take the best," 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.

Appendix: Henry A. Wallace and the modeling of expert judgements

There are several 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.

The method of analysis for modeling both the expert judgements of yield and actual yield was through the new method of path coefficients just developed by Sewall Wright in 1921 (*Correlation and Causation*, *Journal of*

Agricultural Research, *20*, 557–585). The results were final "scorecards" for how the judges and the actual yield values could be assessed by the six factors (each was normalized to a total of 100 "points"):

JUDGES' SCORE CARD:
Length – 42.0
Circumference – 13.6
Weight of kernel – 18.3
Filling of kernel at tip – 13.3
Blistering of kernel – 6.4
Absence of starchiness – 6.4
Total – 100.00
ACTUAL YIELD SCORE CARD:
Length – 7.7
Circumference – 10.0
Weight of kernel – 50.0
Filling of kernel at tip – 18.0
Blistering of kernel – 9.0
Absence of starchiness – 5.3
Total – 100.00

In rather understated conclusion(s), Wallace comments:

It is interesting to note that while the simple correlation coefficients indicate that the judges took into account blistering of kernel as a damaging factor, the path coefficients indicate that they looked on blistering as beneficial. The long ears with heavy kernels for which the judges had such a fondness tended to be freer from blistering than the short ears with light kernels and for that reason it appears on the surface that the judges did not like blistering. But when other factors are held constant, it is found that there is a slight tendency for the judges to favor blistering. Doubtless this was carelessness on the part of these particular judges.

...

The contrast between the yield score card and the judges' score card is interesting.

It will be noted that the tendency of the judges is to emphasize more than anything else, length of ear, whereas Mother Nature, judging merely from these two years' work with one variety of corn, lays her outstanding emphasis on weight of kernel. Over a period of years it may be that the judges are well warranted in making it a prime requisite that a seed ear in the central part of the Corn Belt should at least be eight inches long. But in case of an emergency, in a season when seed corn is scarce, it is probable that so far as that particular year is concerned, length of ear can be disregarded altogether. The important thing would

seem to be to discard those ears carrying light kernels, especially if they have pointed tips, are blistered, and are starchy.

That the corn judges did not know so very much about the factors which make for yield is indicated by the fact that their scores were correlated with yield to the extent of only .2. The difficulty seems to be that they laid too much emphasis on length of ear and possibly also on some fancy points, which caused them to neglect placing as much emphasis on sound, healthy kernel characteristics as they should.

By using Wright's methods of path coefficients, it should be possible in the future to work out in very definite fashion, what really is in the minds of experienced corn judges. It is suggested that the things which really are in their minds are considerably different from the professed score card. It is realized of course that when the judges are working on samples of corn all of which is of show quality, that length of ear will not be so large a factor as it was in the case of this study when the ears were field run, varying from less than five inches to more than ten inches in length. It should be interesting to make another study to determine just what is in the minds of the corn judges when they are judging single ear samples at a corn show.

That corn judging is to some extent a profession with recognized standards is indicated by the fact that the correlation coefficient between the scores of different judges working on the same 500 ears of field, run corn averaged around .7. Inasmuch as corn judging still has a vogue in some of our Corn Belt states, it would seem to be worth while to determine just what is in different corn judges' minds. It would be especially interesting to have corn judges from central Iowa, central Illinois, and central Indiana work on the same 500 ears and then make up by means of path coefficients their true score cards.

1.3 Incorporating Reliability Corrections in Prediction

There are two aspects of variable unreliability in the context of prediction that might have consequences for ethical reasoning. One is in estimating a person's true score on a variable; the second is in how regression might be handled when there is measurement error in the independent and/or dependent variables. In both of these instances, there is an implicit underlying model for how any observed score, X, might be constructed additively from a true score, T_X, and an error score, E_X, where E_X is typically assumed uncorrelated with T_X: $X = T_X + E_X$. When we consider the distribution of an observed variable over, say, a population of individuals, there are two sources of variability present in the true and the error scores. If we are interested primarily in structural models among true scores, then some correction must be made because the common regression models implicitly assume that

variables are measured without error.

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920's (see Kelley, 1947), with a reliance on the algebraic equivalence that the squared correlation between observed and true score is the reliability. If we let $\hat{\rho}$ be the estimated reliability, Kelley's equation can be written as

$$\hat{T}_X = \hat{\rho}X + (1 - \hat{\rho})\bar{X} \ ,$$

where $\bar{X}$ is the mean of the group to which the individual belongs. In other words, depending on the size of $\hat{\rho}$, a person's estimate is partly due to where they are in relation to the group — upwards if below the mean; downwards if above. The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive, that this equation has been called "Kelley's Paradox" (Wainer, 2005, Chapter 10). We might note that this notion of being somewhat punitive of performances better than the group to which one supposedly belongs, was not original with Kelley, but was known at least 400 years earlier. In the words of Miguel de Cervantes (1547–1616): "Tell me what company you keep and I'll tell you what you are."

In addition to obtaining a true score estimate from an obtained score, Kelly's regression model also provides a standard error of estimation (which in this case is now called the standard error of measurement). An approximate 95% confidence interval on an examinee's true score is given by

$$\hat{T}_X \pm 2\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}}) \ ,$$

where $\hat{\sigma}_X$ is the (estimated) standard deviation of the observed scores. By itself, the term $\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}})$, is the standard error of measurement, and is generated from the usual regression formula for the standard error of estimation but applied to Kelly's model predicting true scores. The standard error of measurement most commonly used in the literature is not Kelly's but rather $\hat{\sigma}_X\sqrt{1 - \hat{\rho}}$, and a 95% confidence interval taken as the observed score plus or minus twice this standard error. An argument can be made that this latter procedure leads to "reasonable limits" (after Gulliksen, 1950), whenever $\hat{\rho}$ is reasonably high, and the obtained score is not extremely deviant

from the reference group mean. Why we should assume these latter preconditions and not use the more appropriate procedure to begin with, reminds us of a Bertrand Russell quote: "The method of postulating what we want has many advantages; they are the same as the advantages of theft over honest toil."

In the topic of errors-in-variables regression, we try to compensate for the tacit assumption in regression that all variables are measured without error. Measurement error in a response variable does not bias the regression coefficients per se, but it does increase standard errors, and thereby reduces power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms. Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient. Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables. Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

1.4 Differential Prediction Effects in Selection

One area in which prediction is socially relevant is in selection based on test scores, whether for accreditation, certification, job placement, licensure, educational admission, or other high-stakes endeavors. We note that most of these discussions about fairness of selection need to be phrased in terms of regression models relating a performance measure to a selection test; and whether the regressions are the same over all identified groups of relevance, e.g., ethnic, gender, age, and so on. Specifically, are slopes and intercepts the same? If so or if not, how does this affect the selection mechanism being implemented, and whether it can be considered fair? It is safe to say that depending on the pattern of data within groups, all sorts of things can happen. Generally, an understanding of how a regression/selection model works

with this kind of variation, is necessary for a numerically literate discussion of its intended or unintended consequences. To give a greater sense of the complications that can arise, an extended but redacted quote is given below from Allen and Yen (2001; Chapter 4.4, *Bias in Selection*):

> When regression equations are used in selection procedures and regression lines differ across groups, questions of fairness can arise. For example, suppose that, in predicting a criterion such as college grades for two groups of examinees, the regression lines are found to be parallel but generally higher for one group than the other. ... The regression equation that was produced in the combined group, when compared with within-group regression equations, consistently overpredicts criterion scores for group 2 and underpredicts criterion scores for group 1. In effect, the combined regression equation favors the low-scoring group rather than the high-scoring group. This effect suggests that, if there are group differences in the level of criterion scores in the regression problem, using a combined-group or the higher group's regression equation can help the 'disadvantaged' group.
>
> If there are group differences in the level of predictor scores, a combined-group regression equation can underpredict the lower group's criterion scores. ... The combined group regression line, when compared with within-group predictions, overpredicts criterion scores for most members of group 1 and underpredicts criterion scores for most members of group 2. Using the combined-group regression line in this situation would hurt the disadvantaged group (that is, the group with lower predictor scores).
>
> When regression equations differ across groups, we cannot state (without causing an argument) which procedure is more fair — the use of different regression lines for the two groups or the use of the regression line based on the combined group. If different equations are used, examinees in one group can complain that, to be attributed with the same criterion scores, they need a higher predictor score than those in the other group. In other words, two examinees with the same high school grades could have different predicted college grades solely because they belong to different groups. If the regression equation based on the combined groups is used, some examinees can complain that group membership is a valid part of the prediction and their criterion scores are being underpredicted.

The practical consequences of these differential prediction effects were made evident in a employment discrimination suit in which the plaintiffs claimed that women were underpaid. The evidence supporting this claim was a regression analysis in which salary, on the y-axis, was regressed on a composite index of job qualifications, on the x-axis. The regression line for men was higher than that for women, indicating that for equally qualified candidates, men were paid more. The defendants countered by reversing the regression, conditioning on salary, and showed that for the same salary, the employer could get a more qualified man (Conway & Roberts, 1983). Would the judge have been able to reach an ethically and scientifically correct deci-

sion without a deep understanding of regression? Well, yes. He ruled for the plaintiffs because the law protects against unequal pay by sex. But it does not protect an employer's "right" to get the most for their salary dollars.

1.5 Interpreting and Making Inferences From Regression Weights

Mathematics has given economics rigor, but alas, also mortis.
– Robert Heilbroner

An all too common error in multivariable systems is to over-interpret the meaning of the obtained regression weights. For example, in a model that predicts freshmen college grades from Scholastic Aptitude Test (SAT) scores and High School Grade Point Average (HSGPA) (two highly correlated predictor variables), it has often been argued that because the regression weight was much higher for HSGPA than for SAT, the latter could just be eliminated. In fact, both variables are correlated about 0.7 with freshmen grades, and because of their high intercorrelation, their regression weights are massively unstable. Another instructive and evocative example of the dangers of such interpretations grew out of a large regression analysis done during World War II (personal communication to HW by John Tukey, January 20, 2000). To understand the various aspects of their flights affecting the accuracy of Allied bombers on missions over Germany, a large regression analysis was used to predict accuracy of bombing as a function of many variables. After gathering large amounts of data, a model was built. Variables showing no relation were elided and a final model derived and tested successfully on a neutral data sample. The final model showed a positive regression weight on the variable "number of interceptors encountered." Or in other words, when the Germans sent up fighters to intercept the bombers, their accuracy improved! Some wished to interpret this result in a causal way, but wiser heads prevailed. A deeper examination showed that when the weather was overcast, bomber visibility was so impaired that they couldn't hit anything. In such situations the Germans didn't bother sending up interceptors, relying on ground fire entirely. It was only when the weather was clear that interceptors were launched. When a new variable, "visibility," was added

to the model, the regression weight associated with the variable "number of interceptors encountered" changed sign, and became negative. The lesson to be drawn from all of this, is that we can never be sure all relevant variables are included, and when we add one, the size and even the direction of the regression weights can change.

Although multiple regression can be an invaluable tool in many arenas, the interpretive difficulties that result from the interrelated nature of the independent variables must always be kept in mind. As in the World War II example just described, depending on what variables are (or are not) included, the structural relationship among the variables can change dramatically. At times, this malleability can be put to either good or ill usage. For example, in applying regression models to argue for employment discrimination (e.g., in pay, promotion, hiring, and so on), the multivariable system present could be problematic in arriving at a "correct" analysis. Depending on what variables are included, some variables may "act" for others (as "proxies"), or be used to hide (or at least, to mitigate) various effects. If a case for discrimination rests on the size of a coefficient for some "dummy" variable that indicates group membership (according to race, sex, age, and so on), it may be possible to change its size depending on what variables are included or excluded from the model, and their relationship to the dummy variable. In short, based on how the regressions are performed and one's own (un)ethical predilections, different conclusions could be produced from what is essentially the same data set.[3]

In considering regression in econometric contexts, interests are typically not in obtaining any cosmic understanding of the interrelations among the independent variables, or in the story that might be told. The goal is usually more pragmatic, and phrased in terms of predicting a variable reflecting value and characterized in some numerical way (e.g., as in money or performance statistics). The specific predictor variables used are of secondary importance; what is central is that they "do the job." One recent example of success for quantitative modeling is documented by Michael Lewis in *Moneyball*, with its focus on data-driven decision making in baseball. Instead of relying on finding major league ball players using the hordes of fallible scouts visiting interminable high-school and college games, one adopts quantitative measures

of performance, some developed by the quantitative guru of baseball, Bill James. *Moneyball* relates the story of the Oakland Athletics and their general manager, Billy Beane, and how a successful team, even with a limited budget, could be built on the basis of statistical analysis and insight, and not on intuitive judgements from other baseball personnel (e.g., from coaches, scouts, baseball writers, and so on).

A contentious aspect of using regression and other types of models to drive decision making, arises when "experts" are overridden (or their assessments second-guessed and discounted, or their livelihoods threatened) – by replacing their judgements with those provided by an equation. One particularly entertaining example is in the prediction of wine quality in the Bordeaux or elsewhere. Here, we have wine experts such as Robert Parker (of the *Wine Advocate*), matched against econometricians such as Orley Ashenfelter (of Princeton). One good place to start is with the *Chance* article by Ashenfelter, Ashmore, and LaLonde, *Bordeaux Wine Vintage Quality and the Weather* (*8*, 1995, 7–14). As the article teaser states: "Statistical prediction of wine prices based on vintage growing-season characteristics produces consternation among wine 'experts'." We also note an earlier article from *The New York Times* by Peter Passell (March 4, 1990), with the cute double-entendre title: *Wine Equation Puts Some Noses Out of Joint.* There is also a short (and amusing) letter to the Editor (March 18, 1990) by Frederick R. Waugh: *Keep Those Economists Out of the Vineyards.*

1.6 The (Un)reliability of Clinical Prediction

Prosecutors in Dallas have said for years – any prosecutor can convict a guilty man. It takes a great prosecutor to convict an innocent man.

Melvyn Bruder (*The Thin Blue Line*)

The last section in this chapter on prediction is much longer than most. It concerns the (un)reliability of clinical (behavioral) prediction, particularly for violence, and will include two extensive redactions in appendices at the end of the section: one is the majority opinion in the Supreme Court case of Barefoot v. Estelle (Decided, July 6, 1983) and an eloquent Justice Blackmun dissent; the second is an *Amicus Curiae* brief in this same case from the

American Psychiatric Association on the accuracy of clinical prediction of future violence. Both of these documents are detailed and self-explanatory, and highly informative about our current (in)abilities to make clinical assessments that would lead to accurate and reliable predictions of future behavior. To set the background for the Barefoot v. Estelle case, the beginning part of the *Amicus Curiae* brief follows; parts of the rest of the brief, as already noted, are given in an appendix at the end of the section.

American Psychiatric Association, *Amicus Curiae* Brief: Barefoot v. Estelle

Petitioner Thomas A. Barefoot stands convicted by a Texas state court of the August 7, 1978 murder of a police officer – one of five categories of homicides for which Texas law authorizes the imposition of the death penalty. Under capital sentencing procedures established after this Court's decision in Furman v. Georgia, the "guilt" phase of petitioner's trial was followed by a separate sentencing proceeding in which the jury was directed to answer three statutorily prescribed questions. One of these questions – and the only question of relevance here – directed the jury to determine: whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society. The jury's affirmative response to this question resulted in petitioner being sentenced to death.

The principle evidence presented to the jury on the question of petitioner's "future dangerousness" was the expert testimony of two psychiatrists, Dr. John T. Holbrook and Dr. James Grigson, both of whom testified for the prosecution. Petitioner elected not to testify in his own defense. Nor did he present any evidence or testimony, psychiatric or otherwise, in an attempt to rebut the state's claim that he would commit future criminal acts of violence.

Over defense counsel's objection, the prosecution psychiatrists were permitted to offer clinical opinions regarding petitioner, including their opinions on the ultimate issue of future dangerousness, even though they had not performed a psychiatric examination or evaluation of him. Instead, the critical psychiatric testimony was elicited through an extended hypothetical question propounded by the prosecutor. On the basis of the assumed facts stated in the hypothetical, both Dr. Holbrook and Dr. Grigson gave essentially the same testimony.

First, petitioner was diagnosed as a severe criminal sociopath, a label variously defined as describing persons who "lack a conscience," and who "do things which serve their own purposes without regard for any consequences or outcomes to other people." Second, both psychiatrists testified that petitioner would commit criminal acts of violence in the future. Dr. Holbrook stated that he could predict petitioner's future behavior in this regard "within reasonable psychiatric certainty." Dr. Grigson was more confident, claiming predictive accuracy of "one hundred percent and absolute."

The prosecutor's hypothetical question consisted mainly of a cataloguing of petitioner's

past antisocial behavior, including a description of his criminal record. In addition, the hypothetical question contained a highly detailed summary of the prosecution's evidence introduced during the guilt phase of the trial, as well as a brief statement concerning petitioner's behavior and demeanor during the period from his commission of the murder to his later apprehension by police.

In relevant part, the prosecutor's hypothetical asked the psychiatrists to assume as true the following facts: First, that petitioner had been convicted of five criminal offenses – all of them nonviolent, as far as the record reveals – and that he had also been arrested and charged on several counts of sexual offenses involving children. Second, that petitioner had led a peripatetic existence and "had a bad reputation for peaceful and law abiding citizenship" in each of eight communities that he had resided in during the previous ten years. Third, that in the two-month period preceding the murder, petitioner was unemployed, spending much of his time using drugs, boasting of his plans to commit numerous crimes, and in various ways deceiving certain acquaintances with whom he was living temporarily. Fourth, that petitioner had murdered the police officer as charged, and that he had done so with "no provocation whatsoever" by shooting the officer in the head "from a distance of no more than six inches." And fifth, that subsequent to the murder, petitioner was observed by one witness, "a homosexual," who stated that petitioner "was not in any way acting unusual or that anything was bothering him or upsetting him ..."

Testimony of Dr. Holbrook:

Dr. Holbrook was the first to testify on the basis of the hypothetical question. He stated that the person described in the question exhibited "probably six or seven major criterias (sic) for the sociopath in the criminal area within reasonable medical certainty." Symptomatic of petitioner's sociopathic personality, according to Dr. Holbrook, was his consistent "antisocial behavior" from "early life into adulthood," his willingness to take any action which "serves [his] own purposes" without any regard for the "consequences to other people," and his demonstrated failure to establish any "loyalties to the normal institutions such as family, friends, politics, law or religion."

Dr. Holbrook explained that his diagnosis of sociopathy was also supported by petitioner's past clinical violence and "serious threats of violence," as well as an apparent history of "escaping or running away from authority" rather than "accepting a confrontation in the legal way in a court of law." And finally, Dr. Holbrook testified that petitioner had shown a propensity to "use other people through lying and manipulation ... " According to Dr. Holbrook, by use of such manipulation the sociopath succeeds in "enhancing [his] own ego image ... It makes [him] feel good."

After stating his diagnosis of sociopathy, Dr. Holbrook was asked whether he had an "opinion within reasonable psychiatric certainty as to whether or not there is a probability that the Thomas A. Barefoot in that hypothetical will commit criminal acts of violence in the future that would constitute a continuing threat to society?" Without attempting to explain the implied clinical link between his diagnosis of petitioner and his prediction of future dangerousness, Dr. Holbrook answered simply: "In my opinion he will."

Testimony of Dr. Grigson:

On the basis of the prosecutor's hypothetical question, Dr. Grigson diagnosed petitioner as "a fairly classical, typical, sociopathic personality disorder" of the "most severe category." The most "outstanding characteristic" of persons fitting this diagnosis, according to Dr. Grigson, is the complete "lack of a conscience." Dr. Grigson stated that such persons "repeatedly break the rules, they con, manipulate and use people, [and] are only interested in their own self pleasure [and] gratification."

Although Dr. Grigson testified that some sociopathic individuals do not pose a continuing threat to society, he characterized petitioner as "your most severe sociopath." Dr. Grigson stated that persons falling into this special category are "the ones that ... have complete disregard for another human being's life." Dr. Grigson further testified that "there is not anything in medicine or psychiatry or any other field that will in any way at all modify or change the severe sociopath."

The prosecutor then asked Dr. Grigson to state his opinion on the ultimate issue – "whether or not there is a probability that the defendant ... will commit criminal acts of violence that would constitute a continuing threat to society?" Again, without explaining the basis for his prediction or its relationship to the diagnosis of sociopathy, Dr. Grigson testified that he was "one hundred percent" sure that petitioner "most certainly would" commit future criminal acts of violence. Dr. Grigson also stated that his diagnosis and prediction would be the same whether petitioner "was in the penitentiary or whether he was free."

The Psychiatrist, James Grigson, featured so prominently in the opinions for Barefoot v. Estelle and the corresponding American Psychiatric Association *Amicus* brief, played the same role repeatedly in the Texas legal system. For over three decades before his retirement in 2003, he would testify when requested at death sentence hearings to a high certainty as to "whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society." An affirmative answer by the sentencing jury imposed the death penalty automatically, as it was on Thomas Barefoot; he was executed on October 30, 1984. When asked if he had a last statement to make, he replied:

Yes, I do. I hope that one day we can look back on the evil that we're doing right now like the witches we burned at the stake. I want everybody to know that I hold nothing against them. I forgive them all. I hope everybody I've done anything to will forgive me. I've been praying all day for Carl Levin's wife to drive the bitterness from her heart because that bitterness that's in her heart will send her to Hell just as surely as any other sin. I'm sorry for everything I've ever done to anybody. I hope they'll forgive me.

James Grigson was expelled in 1995 from the American Psychiatric Association and the Texas Association of Psychiatric Physicians for two chronic

ethics violations: making statements in testimony on defendants he had not actually examined, and for predicting violence with 100% certainty. The press gave him the nickname of "Dr. Death." The role he played was similar to Roy Meadow, our earlier villain, who crusaded against mothers supposedly abusing their children (remember, the Münchausen Syndrome by Proxy). In Grigson's case, it was sociopaths he wanted put to death, as opposed to receiving just life imprisonment (without parole).

There is another connection in Barefoot v. Estelle to our earlier discussions about the distinctions between actuarial and clinical prediction, and where the former is commonly better than the latter. There is some evidence mentioned in the APA brief that actuarial predictions of violence carried out by statistically informed laymen might be better than those of a clinician, because of the absence of bias that psychiatrists might (unsuspectingly) have in over-predicting violence, whether because of the clients they see or for other reasons related to their practice. There is a pertinent passage from the APA brief (not given in our redactions):

> That psychiatrists actually may be less accurate predictors of future violence than laymen, may be due to personal biases in favor of predicting violence arising from the fear of being responsible for the erroneous release of a violent individual. It also may be due to a tendency to generalize from experiences with past offenders on bases that have no empirical relationship to future violence, a tendency that may be present in Grigson's and Holbrook's testimony. Statistical prediction is clearly more reliable than clinical prediction — and prediction based on statistics alone may be done by anyone.

The two psychiatrists mentioned in Barefoot v. Estelle, James Grigson and John Holbrook, appeared together repeatedly in various capital sentencing hearings in Texas during the later part of the 20th century. Although Grigson was generally the more outrageous of the two with predictions of absolute certitude based on a sociopath diagnosis, Holbrook was similarly at fault ethically. This Frick and Frack of Texas death penalty fame might well be nicknamed "Dr. Death" and "Dr. Doom." They were both culpable in the famous exoneration documented in the award winning film by Errol Morris, *The Thin Blue Line*.

A later chapter of this monograph discusses in detail the Federal Rules of Evidence and the admissibility of expert witnesses and scientific data. The central case discussed in that chapter is Daubert v. Merrell Dow Pharmaceu-

tical (1993) that promulgates what is called the Daubert standard for admitting expert testimony in federal courts. The majority opinion in Daubert was written by Justice Blackman, the same Justice writing the dissent in Barefoot v. Estelle. The court stated that Rule 702 of the Federal Rules of Evidence was the governing standard for admitting scientific evidence in trials held in federal court (and now in most state courts as well). Rule 702, Testimony by Experts, states:

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

We give a short redaction of that part of the Wikipedia article on Daubert v. Merrell devoted to the discussion of the Daubert standard governing expert testimony. Needless to say, we have doubts that clinical predictions of violence based on a sociopath diagnosis would be admissible under the Daubert standard.

The Standard Governing Expert Testimony: Three key provisions of the Rules governed admission of expert testimony in court. The first was scientific knowledge. This means that the testimony must be scientific in nature, and that the testimony must be grounded in "knowledge." Of course, science does not claim to know anything with absolute certainty; science "represents a process for proposing and refining theoretical explanations about the world that are subject to further testing and refinement." The "scientific knowledge" contemplated by Rule 702 had to be arrived at by the scientific method.

Second, the scientific knowledge must assist the trier of fact in understanding the evidence or determining a fact in issue in the case. The trier of fact is often either a jury or a judge; but other fact-finders may exist within the contemplation of the federal rules of evidence. To be helpful to the trier of fact, there must be a "valid scientific connection to the pertinent inquiry as a prerequisite to admissibility." Although it is within the purview of scientific knowledge, knowing whether the moon was full on a given night does not typically assist the trier of fact in knowing whether a person was sane when he or she committed a given act.

Third, the Rules expressly provided that the judge would make the threshold determination regarding whether certain scientific knowledge would indeed assist the trier of fact in the manner contemplated by Rule 702. "This entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue." This preliminary assessment can turn on whether something has been tested, whether an idea has been subjected to scientific peer review or published in scientific journals, the rate of error

involved in the technique, and even general acceptance, among other things. It focuses on methodology and principles, not the ultimate conclusions generated.

The Court stressed that the new standard under Rule 702 was rooted in the judicial process and intended to be distinct and separate from the search for scientific truth. "Scientific conclusions are subject to perpetual revision. Law, on the other hand, must resolve disputes finally and quickly. The scientific project is advanced by broad and wide-ranging consideration of a multitude of hypotheses, for those that are incorrect will eventually be shown to be so, and that in itself is an advance." Rule 702 was intended to resolve legal disputes, and thus had to be interpreted in conjunction with other rules of evidence and with other legal means of ending those disputes. Cross examination within the adversary process is adequate to help legal decision-makers arrive at efficient ends to disputes. "We recognize that, in practice, a gate-keeping role for the judge, no matter how flexible, inevitably on occasion will prevent the jury from learning of authentic insights and innovations. That, nevertheless, is the balance that is struck by Rules of Evidence designed not for the exhaustive search for cosmic understanding but for the particularized resolution of legal disputes."

As noted in the various opinions and *Americus* brief given in Barefoot v. Estelle, the jury in considering whether the death penalty should be imposed, has to answer affirmatively one question: whether there was a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society. The bare use of the word "probability" without specifying any further size seems odd to say the least, but Texas courts have steadfastly refused to delimit it any further. So, presumably a very small probability of future violence would be sufficient for execution if this small probability could be proved "beyond a reasonable doubt."

The point of much of this section has been to emphasize that actuarial evidence about future violence is all there really is in making such predictions. More pointedly, the assignment of a clinical label, such as "sociopath," adds nothing to an ability to predict, and to suggest that it does is to use the worst "junk science," even though it may be routinely assumed true in the larger society. All we have to rely on is the usual psychological adage that the best predictor of future behavior is past performance. Thus, the best predictor of criminal recidivism is a history of such behavior, and past violence suggests future violence. The greater the amount of past criminal behavior or violence, the more likely that such future behavior or violence will occur (a form of a "dose-response" relationship). At its basis, this is statistical evidence of such a likely occurrence and no medical or psychological diagnosis is needed (or useful).

Besides the specious application of a sociopath diagnosis to predict future violence, after the Supreme Court decision in Estelle v. Smith, such a usage had to be made on the basis of a hypothetical question and not on an actual psychological examination of the defendant. In addition to a 100% incontrovertible assurance of future violence, offering testimony without actually examining a defendant proved to be Grigson's eventual downfall and one reason for the expulsion from his professional psychiatric societies. This prevention of an actual examination of a defendant by the Supreme Court case, Estelle v. Smith, also involved James Grigson. Ernest Smith, indicted for murder, had been examined by Grigson in jail and who determined he was competent to stand trial. In the psychiatric report on Smith, Grigson termed him "a severe sociopath" but gave no other statements as to future dangerousness. Smith was sentenced to death based on the sociopath label given by Grigson. In Estelle v. Smith (1981), the Supreme Court held that because of the famous case of Miranda v. Arizona (1966), the state could not force a defendant to submit to a psychiatric examination for the purposes of sentencing – it violated a defendant's Firth Amendment rights against self-incrimination and the Sixth Amendment right to counsel. Thus, the examination of Ernest Smith was inadmissible at sentencing. From that point on, predictions of violence were made solely on hypothetical questions and Grigson's belief that a labeling as a sociopath was sufficient to guarantee future violence on the part of a defendant, and therefore, the defendant should be put to death.

Appendix: Continuation of the American Psychiatric Association, *Amicus Curiae* Brief: Barefoot v. Estelle

INTRODUCTION AND SUMMARY OF ARGUMENT

The questions presented in this case are the logical outgrowth of two prior decisions by this Court. In the first, Jurek v. Texas, the Court dealt with the same Texas capital sentencing procedure involved here. The Court there rejected a constitutional challenge to the "future dangerousness" question, ruling that the statutory standard was not impermissibly vague. Although recognizing the difficulty inherent in predicting future behavior, the Court held that "[t]he task that [the] jury must perform ... is basically no different from the task performed countless times each day throughout the American system of criminal justice." The Jurek Court thus upheld the use of the Texas statutory question, but did not consider the types of evidence that could be presented to the jury for purposes of this determination.

Subsequently in Estelle v. Smith, the Court again dealt with the Texas sentencing scheme – this time in the context of a psychiatric examination to determine the defendant's competency to stand trial. The Court held that the Fifth Amendment's privilege against self-incrimination applied to such psychiatric examinations, at least to the extent that a prosecution psychiatrist later testifies concerning the defendant's future dangerousness. The Court reasoned that although a defendant has no generalized constitutional right to remain silent at a psychiatric examination properly limited to the issues of sanity or competency, full Miranda warnings must be given with respect to testimony concerning future dangerousness because of "the gravity of the decision to be made at the penalty phase ... " The Smith decision thus enables a capital defendant to bar a government psychiatric examination on the issue of future dangerousness.

The [present] case raises the two issues left unresolved in Jurek and Smith. These are, first, whether a psychiatrist, testifying as an expert medical witness, may ever be permitted to render a prediction as to a capital defendant's long-term future dangerousness. The second issue is whether such testimony may be elicited on the basis of hypothetical questions, even if there exists no general prohibition against the use of expert psychiatric testimony on the issue of long-term future dangerousness. *Amicus* believes that both of these questions should be answered in the negative.

I. Psychiatrists should not be permitted to offer a prediction concerning the long-term future dangerousness of a defendant in a capital case, at least in those circumstances where the psychiatrist purports to be testifying as a medical expert possessing predictive expertise in this area. Although psychiatric assessments may permit short-term predictions of violent or assaultive behavior, medical knowledge has simply not advanced to the point where long-term predictions – the type of testimony at issue in this case – may be made with even reasonable accuracy. The large body of research in this area indicates that, even under the best of conditions, psychiatric predictions of long-term future dangerousness are wrong in at least two out of every three cases.

The forecast of future violent conduct on the part of a defendant in a capital case is, at bottom, a lay determination, not an expert psychiatric determination. To the extent such predictions have any validity, they can only be made on the basis of essentially actuarial data to which psychiatrists, qua psychiatrists, can bring no special interpretative skills. On the other hand, the use of psychiatric testimony on this issue causes serious prejudice to the defendant. By dressing up the actuarial data with an "expert" opinion, the psychiatrist's testimony is likely to receive undue weight. In addition, it permits the jury to avoid the difficult actuarial questions by seeking refuge in a medical diagnosis that provides a false aura of certainty. For these reasons, psychiatric testimony on future dangerousness impermissibly distorts the fact-finding process in capital cases.

II. Even if psychiatrists under some circumstances are allowed to render an expert medical opinion on the question of future dangerousness, *amicus* submits that they should never be permitted to do so unless they have conducted a psychiatric examination of the defendant. It is evident from the testimony in this case that the key clinical determination relied upon by both psychiatrists was their diagnosis of "sociopathy" or "antisocial personality disorder."

However, such a diagnosis simply cannot be made on the basis of a hypothetical question. Absent an in-depth psychiatric examination and evaluation, the psychiatrist cannot exclude alternative diagnoses; nor can he assure that the necessary criteria for making the diagnosis in question are met. As a result, he is unable to render a medical opinion with a reasonable degree of certainty.

These deficiencies strip the psychiatric testimony of all value in the present context. Even assuming that the diagnosis of antisocial personality disorder is probative of future dangerousness – an assumption which we do not accept – it is nonetheless clear that the limited facts given in the hypothetical fail to disprove other illnesses that plainly do not indicate a general propensity to commit criminal acts. Moreover, these other illnesses may be more amenable to treatment – a factor that may further reduce the likelihood of future aggressive behavior by the defendant.

...

Appendix: Opinion and Dissent in the U.S. Supreme Court, Barefoot v. Estelle

(a) There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community. To accept such an argument would call into question predictions of future behavior that are constantly made in other contexts. Moreover, under the generally applicable rules of evidence covering the admission and weight of unprivileged evidence, psychiatric testimony predicting dangerousness may be countered not only as erroneous in a particular case but also as generally so unreliable that it should be ignored. Nor, despite the view of the American Psychiatric Association supporting petitioner's view, is there any convincing evidence that such testimony is almost entirely unreliable, and that the factfinder and the adversary system will not be competent to uncover, recognize, and take due account of its shortcomings.

(b) Psychiatric testimony need not be based on personal examination of the defendant, but may properly be given in response to hypothetical questions. Expert testimony, whether in the form of an opinion based on hypothetical questions or otherwise, is commonly admitted as evidence where it might help the factfinder do its job. Although this case involves the death penalty, there is no constitutional barrier to applying the ordinary rules of evidence governing the use of expert testimony.

...

Justice Blackmun dissenting:

I agree with most of what Justice Marshall has said in his dissenting opinion. I, too, dissent, but I base my conclusion also on evidentiary factors that the Court rejects with some emphasis. The Court holds that psychiatric testimony about a defendant's future dangerousness is admissible, despite the fact that such testimony is wrong two times out of three. The Court reaches this result – even in a capital case – because, it is said, the testimony is subject to cross-examination and impeachment. In the present state of psychiatric knowl-

edge, this is too much for me. One may accept this in a routine lawsuit for money damages, but when a person's life is at stake – no matter how heinous his offense – a requirement of greater reliability should prevail. In a capital case, the specious testimony of a psychiatrist, colored in the eyes of an impressionable jury by the inevitable untouchability of a medical specialist's words, equates with death itself.

To obtain a death sentence in Texas, the State is required to prove beyond a reasonable doubt that "there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society." As a practical matter, this prediction of future dangerousness was the only issue to be decided by Barefoot's sentencing jury.

At the sentencing hearing, the State established that Barefoot had two prior convictions for drug offenses and two prior convictions for unlawful possession of firearms. None of these convictions involved acts of violence. At the guilt stage of the trial, for the limited purpose of establishing that the crime was committed in order to evade police custody, the State had presented evidence that Barefoot had escaped from jail in New Mexico where he was being held on charges of statutory rape and unlawful restraint of a minor child with intent to commit sexual penetration against the child's will. The prosecution also called several character witnesses at the sentencing hearing, from towns in five States. Without mentioning particular examples of Barefoot's conduct, these witnesses testified that Barefoot's reputation for being a peaceable and law-abiding citizen was bad in their respective communities.

Last, the prosecution called Doctors Holbrook and Grigson, whose testimony extended over more than half the hearing. Neither had examined Barefoot or requested the opportunity to examine him. In the presence of the jury, and over defense counsel's objection, each was qualified as an expert psychiatrist witness. Doctor Holbrook detailed at length his training and experience as a psychiatrist, which included a position as chief of psychiatric services at the Department of Corrections. He explained that he had previously performed many "criminal evaluations," and that he subsequently took the post at the Department of Corrections to observe the subjects of these evaluations so that he could "be certain those opinions that [he] had were accurate at the time of trial and pretrial." He then informed the jury that it was "within [his] capacity as a doctor of psychiatry to predict the future dangerousness of an individual within a reasonable medical certainty,", and that he could give

"an expert medical opinion that would be within reasonable psychiatric certainty as to whether or not that individual would be dangerous to the degree that there would be a probability that that person would commit criminal acts of violence in the future that would constitute a continuing threat to society."

Doctor Grigson also detailed his training and medical experience, which, he said, included examination of "between thirty and forty thousand individuals," including 8,000 charged with felonies, and at least 300 charged with murder. He testified that, with enough information, he would be able to "give a medical opinion within reasonable psychiatric certainty as to the psychological or psychiatric makeup of an individual," and that this skill was "particular to the field of psychiatry, and not to the average layman."

Each psychiatrist then was given an extended hypothetical question asking him to assume

as true about Barefoot the four prior convictions for nonviolent offenses, the bad reputation for being law-abiding in various communities, the New Mexico escape, the events surrounding the murder for which he was on trial and, in Doctor Grigson's case, the New Mexico arrest. On the basis of the hypothetical question, Doctor Holbrook diagnosed Barefoot "within a reasonable psychiatr[ic] certainty," as a "criminal sociopath." He testified that he knew of no treatment that could change this condition, and that the condition would not change for the better but "may become accelerated" in the next few years. Finally, Doctor Holbrook testified that, "within reasonable psychiatric certainty," there was "a probability that the Thomas A. Barefoot in that hypothetical will commit criminal acts of violence in the future that would constitute a continuing threat to society," and that his opinion would not change if the "society" at issue was that within Texas prisons, rather than society outside prison.

Doctor Grigson then testified that, on the basis of the hypothetical question, he could diagnose Barefoot "within reasonable psychiatric certainty" as an individual with "a fairly classical, typical, sociopathic personality disorder." He placed Barefoot in the "most severe category of sociopaths (on a scale of one to ten, Barefoot was "above ten"), and stated that there was no known cure for the condition. Finally, Doctor Grigson testified that whether Barefoot was in society at large or in a prison society there was a "one hundred percent and absolute" chance that Barefoot would commit future acts of criminal violence that would constitute a continuing threat to society.

On cross-examination, defense counsel questioned the psychiatrists about studies demonstrating that psychiatrists' predictions of future dangerousness are inherently unreliable. Doctor Holbrook indicated his familiarity with many of these studies, but stated that he disagreed with their conclusions. Doctor Grigson stated that he was not familiar with most of these studies, and that their conclusions were accepted by only a "small minority group" of psychiatrists – "[i]t's not the American Psychiatric Association that believes that.

After an hour of deliberation, the jury answered "yes" to the two statutory questions, and Thomas Barefoot was sentenced to death.

The American Psychiatric Association (APA), participating in this case as *amicus curiae*, informs us that "[t]he unreliability of psychiatric predictions of long-term future dangerousness is by now an established fact within the profession." The APA's best estimate is that two out of three predictions of long-term future violence made by psychiatrists are wrong. The Court does not dispute this proposition, and indeed it could not do so; the evidence is overwhelming. For example, the APA's Draft Report of the Task Force on the Role of Psychiatry in the Sentencing Process (1983) states that

"[c]onsiderable evidence has been accumulated by now to demonstrate that long-term prediction by psychiatrists of future violence is an extremely inaccurate process."

John Monahan, recognized as "the leading thinker on this issue" even by the State's expert witness at Barefoot's federal habeas corpus hearing, concludes that

"the 'best' clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior,"

even among populations of individuals who are mentally ill and have committed violence in the past. Another study has found it impossible to identify any subclass of offenders

"whose members have a greater-than-even chance of engaging again in an assaultive act." Yet another commentator observes:

"In general, mental health professionals ... are more likely to be wrong than right when they predict legally relevant behavior. When predicting violence, dangerousness, and suicide, they are far more likely to be wrong than right."

Neither the Court nor the State of Texas has cited a single reputable scientific source contradicting the unanimous conclusion of professionals in this field that psychiatric predictions of long-term future violence are wrong more often than they are right.

The APA also concludes, as do researchers that have studied the issue, that psychiatrists simply have no expertise in predicting long-term future dangerousness. A layman with access to relevant statistics can do at least as well, and possibly better; psychiatric training is not relevant to the factors that validly can be employed to make such predictions, and psychiatrists consistently err on the side of overpredicting violence. Thus, while Doctors Grigson and Holbrook were presented by the State and by self-proclamation as experts at predicting future dangerousness, the scientific literature makes crystal clear that they had no expertise whatever. Despite their claims that they were able to predict Barefoot's future behavior "within reasonable psychiatric certainty," or to a "one hundred percent and absolute" certainty, there was, in fact, no more than a one in three chance that they were correct.[4]

It is impossible to square admission of this purportedly scientific but actually baseless testimony with the Constitution's paramount concern for reliability in capital sentencing.[5] Death is a permissible punishment in Texas only if the jury finds beyond a reasonable doubt that there is a probability the defendant will commit future acts of criminal violence. The admission of unreliable psychiatric predictions of future violence, offered with unabashed claims of "reasonable medical certainty" or "absolute" professional reliability, creates an intolerable danger that death sentences will be imposed erroneously.

The plurality in Woodson v. North Carolina, stated:

"Death, in its finality, differs more from life imprisonment than a 100-year prison term differs from one of only a year or two. Because of that qualitative difference, there is a corresponding difference in the need for reliability in the determination that death is the appropriate punishment in a specific case."

The Court does not see fit to mention this principle today, yet it is as firmly established as any in our Eighth Amendment jurisprudence. Only two weeks ago, in Zant v. Stephens, the Court described the need for reliability in the application of the death penalty as one of the basic "themes ... reiterated in our opinions discussing the procedures required by the Constitution in capital sentencing determinations." (capital punishment must be "imposed fairly, and with reasonable consistency, or not at all"). State evidence rules notwithstanding, it is well established that, because the truth-seeking process may be unfairly skewed, due process may be violated even in a noncapital criminal case by the exclusion of evidence probative of innocence, or by the admission of certain categories of unreliable and prejudicial evidence ("[i]t is the reliability of identification evidence that primarily determines its admissibility"). The reliability and admissibility of evidence considered by a capital sentencing factfinder is

obviously of still greater constitutional concern.

The danger of an unreliable death sentence created by this testimony cannot be brushed aside on the ground that the "jury [must] have before it all possible relevant information about the individual defendant whose fate it must determine." Although committed to allowing a "wide scope of evidence" at presentence hearings, the Court has recognized that "consideration must be given to the quality, as well as the quantity, of the information on which the sentencing [authority] may rely." Thus, very recently, this Court reaffirmed a crucial limitation on the permissible scope of evidence: "[s]o long as the evidence introduced ... do[es] not prejudice a defendant, it is preferable not to impose restrictions." The Court all but admits the obviously prejudicial impact of the testimony of Doctors Grigson and Holbrook; granting that their absolute claims were more likely to be wrong than right, the Court states that "[t]here is no doubt that the psychiatric testimony increased the likelihood that petitioner would be sentenced to death.". Indeed, unreliable scientific evidence is widely acknowledged to be prejudicial. The reasons for this are manifest. "The major danger of scientific evidence is its potential to mislead the jury; an aura of scientific infallibility may shroud the evidence, and thus lead the jury to accept it without critical scrutiny."[6]

Where the public holds an exaggerated opinion of the accuracy of scientific testimony, the prejudice is likely to be indelible. There is little question that psychiatrists are perceived by the public as having a special expertise to predict dangerousness, a perception based on psychiatrists' study of mental disease. It is this perception that the State in Barefoot's case sought to exploit. Yet mental disease is not correlated with violence, and the stark fact is that no such expertise exists. Moreover, psychiatrists, it is said, sometimes attempt to perpetuate this illusion of expertise, and Doctors Grigson and Holbrook – who purported to be able to predict future dangerousness "within reasonable psychiatric certainty," or absolutely – present extremely disturbing examples of this tendency. The problem is not uncommon.

Furthermore, as is only reasonable, the Court's concern in encouraging the introduction of a wide scope of evidence has been to ensure that accurate information is provided to the sentencing authority without restriction. The joint opinion announcing the judgment in Gregg explained the jury's need for relevant evidence in these terms:

"If an experienced trial judge, who daily faces the difficult task of imposing sentences, has a vital need for accurate information ... to be able to impose a rational sentence in the typical criminal case, then accurate sentencing information is an indispensable prerequisite to a reasoned determination of whether a defendant shall live or die by a jury of people who may never before have made a sentencing decision."

So far as I am aware, the Court never has suggested that there is any interest in providing deceptive and inaccurate testimony to the jury. Psychiatric predictions of future dangerousness are not accurate; wrong two times out of three, their probative value, and therefore any possible contribution they might make to the ascertainment of truth, is virtually nonexistent (psychiatric testimony not sufficiently reliable to support finding that individual will be dangerous under any standard of proof). Indeed, given a psychiatrist's prediction that an individual will be dangerous, it is more likely than not that the defendant will not commit

further violence. It is difficult to understand how the admission of such predictions can be justified as advancing the search for truth, particularly in light of their clearly prejudicial effect. Thus, the Court's remarkable observation that "[n]either petitioner nor the [APA] suggests that psychiatrists are always wrong with respect to future dangerousness, only most of the time," misses the point completely, and its claim that this testimony was no more problematic than "other relevant evidence against any defendant in a criminal case," is simply incredible. Surely, this Court's commitment to ensuring that death sentences are imposed reliably and reasonably requires that nonprobative and highly prejudicial testimony on the ultimate question of life or death be excluded from a capital sentencing hearing.

Despite its recognition that the testimony at issue was probably wrong and certainly prejudicial, the Court holds this testimony admissible because the Court is

"unconvinced ... that the adversary process cannot be trusted to sort out the reliable from the unreliable evidence and opinion about future dangerousness."

One can only wonder how juries are to separate valid from invalid expert opinions when the "experts" themselves are so obviously unable to do so. Indeed, the evidence suggests that juries are not effective at assessing the validity of scientific evidence.

There can be no question that psychiatric predictions of future violence will have an undue effect on the ultimate verdict. Even judges tend to accept psychiatrists' recommendations about a defendant's dangerousness with little regard for cross-examination or other testimony. The American Bar Association has warned repeatedly that sentencing juries are particularly incapable of dealing with information relating to "the likelihood that the defendant will commit other crimes," and similar predictive judgments. Relying on the ABA's conclusion, the joint opinion announcing the judgment in Gregg v. Georgia, recognized that,

"[s]ince the members of a jury will have had little, if any, previous experience in sentencing, they are unlikely to be skilled in dealing with the information they are given."

But the Court in this case, in its haste to praise the jury's ability to find the truth, apparently forgets this well-known and worrisome shortcoming.

As if to suggest that petitioner's position that unreliable expert testimony should be excluded is unheard of in the law, the Court relies on the proposition that the rules of evidence generally

"anticipate that relevant, unprivileged evidence should be admitted and its weight left to the factfinder, who would have the benefit of cross-examination and contrary evidence by the opposing party."

But the Court simply ignores hornbook law that, despite the availability of cross-examination and rebuttal witnesses,

"opinion evidence is not admissible if the court believes that the state of the pertinent art or scientific knowledge does not permit a reasonable opinion to be asserted."

Because it is feared that the jury will overestimate its probative value, polygraph evidence, for example, almost invariably is excluded from trials despite the fact that, at a conservative estimate, an experienced polygraph examiner can detect truth or deception correctly about 80 to 90 percent of the time. In no area is purportedly "expert" testimony admitted for the jury's consideration where it cannot be demonstrated that it is correct more often than

not. "It is inconceivable that a judgment could be considered an expert' judgment when it is less accurate than the flip of a coin." The risk that a jury will be incapable of separating "scientific" myth from reality is deemed unacceptably high.[7]

The Constitution's mandate of reliability, with the stakes at life or death, precludes reliance on cross-examination and the opportunity to present rebuttal witnesses as an antidote for this distortion of the truthfinding process. Cross-examination is unlikely to reveal the fatuousness of psychiatric predictions because such predictions often rest, as was the case here, on psychiatric categories and intuitive clinical judgments not susceptible to cross-examination and rebuttal. Psychiatric categories have little or no demonstrated relationship to violence, and their use often obscures the unimpressive statistical or intuitive bases for prediction.[8] The APA particularly condemns the use of the diagnosis employed by Doctors Grigson and Holbrook in this case, that of sociopathy:

"In this area confusion reigns. The psychiatrist who is not careful can mislead the judge or jury into believing that a person has a major mental disease simply on the basis of a description of prior criminal behavior. Or a psychiatrist can mislead the court into believing that an individual is devoid of conscience on the basis of a description of criminal acts alone. ... The profession of psychiatry has a responsibility to avoid inflicting this confusion upon the courts, and to spare the defendant the harm that may result. ... Given our uncertainty about the implications of the finding, the diagnosis of sociopathy ... should not be used to justify or to support predictions of future conduct. There is no certainty in this area."

It is extremely unlikely that the adversary process will cut through the facade of superior knowledge. The Chief Justice [Burger] long ago observed:

"The very nature of the adversary system ... complicates the use of scientific opinion evidence, particularly in the field of psychiatry. This system of partisan contention, of attack and counterattack, at its best is not ideally suited to developing an accurate portrait or profile of the human personality, especially in the area of abnormal behavior. Although under ideal conditions the adversary system can develop for a jury most of the necessary fact material for an adequate decision, such conditions are rarely achieved in the courtrooms in this country. These ideal conditions would include a highly skilled and experienced trial judge and highly skilled lawyers on both sides of the case, all of whom, in addition to being well-trained in the law and in the techniques of advocacy, would be sophisticated in matters of medicine, psychiatry, and psychology. It is far too rare that all three of the legal actors in the cast meet these standards."

Another commentator has noted:

"Competent cross-examination and jury instructions may be partial antidotes ... but they cannot be complete. Many of the cases are not truly adversarial; too few attorneys are skilled at cross-examining psychiatrists, laypersons overweigh the testimony of experts, and, in any case, unrestricted use of experts promotes the incorrect view that the questions are primarily scientific. There is, however, no antidote for the major difficulty with mental health 'experts' – that they simply are not experts. ... In realms beyond their true expertise, the law has little special to learn from them; too often, their testimony is ... prejudicial."

Nor is the presentation of psychiatric witnesses on behalf of the defense likely to remove

the prejudicial taint of misleading testimony by prosecution psychiatrists. No reputable expert would be able to predict with confidence that the defendant will not be violent; at best, the witness will be able to give his opinion that all predictions of dangerousness are unreliable. Consequently, the jury will not be presented with the traditional battle of experts with opposing views on the ultimate question. Given a choice between an expert who says that he can predict with certainty that the defendant, whether confined in prison or free in society, will kill again, and an expert who says merely that no such prediction can be made, members of the jury, charged by law with making the prediction, surely will be tempted to opt for the expert who claims he can help them in performing their duty, and who predicts dire consequences if the defendant is not put to death.[9]

Moreover, even at best, the presentation of defense psychiatrists will convert the death sentence hearing into a battle of experts, with the Eighth Amendment's well-established requirement of individually focused sentencing a certain loser. The jury's attention inevitably will turn from an assessment of the propriety of sentencing to death the defendant before it to resolving a scientific dispute about the capabilities of psychiatrists to predict future violence. In such an atmosphere, there is every reason to believe that the jury may be distracted from its constitutional responsibility to consider "particularized mitigating factors," in passing on the defendant's future dangerousness.

One searches the Court's opinion in vain for a plausible justification for tolerating the State's creation of this risk of an erroneous death verdict. As one Court of Appeals has observed:

"A courtroom is not a research laboratory. The fate of a defendant ... should not hang on his ability to successfully rebut scientific evidence which bears an 'aura of special reliability and trustworthiness,' although, in reality, the witness is testifying on the basis of an unproved hypothesis ... which has yet to gain general acceptance in its field."

Ultimately, when the Court knows full well that psychiatrists' predictions of dangerousness are specious, there can be no excuse for imposing on the defendant, on pain of his life, the heavy burden of convincing a jury of laymen of the fraud.[10]

The Court is simply wrong in claiming that psychiatric testimony respecting future dangerousness is necessarily admissible in light of Jurek v. Texas, or Estelle v. Smith. As the Court recognizes, Jurek involved "only lay testimony." Thus, it is not surprising that "there was no suggestion by the Court that the testimony of doctors would be inadmissible," and it is simply irrelevant that the Jurek Court did not "disapprov[e]" the use of such testimony. In Smith, the psychiatric testimony at issue was given by the same Doctor Grigson who confronts us in this case, and his conclusions were disturbingly similar to those he rendered here. The APA, appearing as *amicus curiae*, argued that all psychiatric predictions of future dangerousness should be excluded from capital sentencing proceedings. The Court did not reach this issue, because it found Smith's death sentence invalid on narrower grounds: Doctor Grigson's testimony had violated Smith's Fifth and Sixth Amendment right. Contrary to the Court's inexplicable assertion in this case, Smith certainly did not reject the APA's position. Rather, the Court made clear that "the holding in Jurek was guided by recognition that the inquiry [into dangerousness] mandated by Texas law does not require

resort to medical experts." If Jurek and Smith held that psychiatric predictions of future dangerousness are admissible in a capital sentencing proceeding as the Court claims, this guiding recognition would have been irrelevant.

The Court also errs in suggesting that the exclusion of psychiatrists' predictions of future dangerousness would be contrary to the logic of Jurek. Jurek merely upheld Texas' substantive decision to condition the death sentence upon proof of a probability that the defendant will commit criminal acts of violence in the future. Whether the evidence offered by the prosecution to prove that probability is so unreliable as to violate a capital defendant's rights to due process is an entirely different matter, one raising only questions of fair procedure.[11] Jurek's conclusion that Texas may impose the death penalty on capital defendants who probably will commit criminal acts of violence in no way establishes that the prosecution may convince a jury that this is so by misleading or patently unreliable evidence.

Moreover, Jurek's holding that the Texas death statute is not impermissibly vague does not lead ineluctably to the conclusion that psychiatric testimony is admissible. It makes sense to exclude psychiatric predictions of future violence while admitting lay testimony, because psychiatric predictions appear to come from trained mental health professionals, who purport to have special expertise. In view of the total scientific groundlessness of these predictions, psychiatric testimony is fatally misleading. Lay testimony, frankly based on statistical factors with demonstrated correlations to violent behavior, would not raise this substantial threat of unreliable and capricious sentencing decisions, inimical to the constitutional standards established in our cases; and such predictions are as accurate as any a psychiatrist could make. Indeed, the very basis of Jurek, as I understood it, was that such judgments can be made by laymen on the basis of lay testimony.

Our constitutional duty is to ensure that the State proves future dangerousness, if at all, in a reliable manner, one that ensures that "any decision to impose the death sentence be, and appear to be, based on reason rather than caprice or emotion." Texas' choice of substantive factors does not justify loading the factfinding process against the defendant through the presentation of what is, at bottom, false testimony.

Notes

[1]There are several topics involving prediction that do not (necessarily) concern linear regression, and because of this, no extended discussion of these is given in this chapter. One area important for the legal system is Sex Offender Risk Assessment, and the prediction of recidivism for committing another offense. Several such instruments are available, with most relying on some simple point counting system based on descriptive information about the subject and previous crimes (and readily available in a subject's "jacket"). One of the easiest to implement is called the Rapid Risk Assessment for Sexual Offender Recidivism (or the more common acronym, RRASOR, and pronounced "razor"). It is based on four items:

Prior Sex Offense Convictions: 0, 1, 2, or 3 points for 0, 1, 2, or 3+ prior convictions, respectively; Victim Gender: only female victims (0 points); only male victims (1 point); Relationship to Victim: only related victims (0 points); any unrelated victim (1 point); Age at Release: 25 or more (0 points); 18 up to 25 (1 point).

As the RRASOR author Hanson (1997) notes in validation work, those with a score of zero had a recidivism rate of 6.5% after 10 years; for those who scored 5, the rate was 73% after 10 years – R. K.

Hanson (1997), *The Development of a Brief Actuarial Risk Scale for Sexual Offense Recidivism*, Ottawa: Solicitor General of Canada.

Another approach to prediction that we do not develop, is in the theory behind chaotic systems, such as the weather. A hallmark of such dynamic prediction problems is an extreme sensitivity to initial conditions, and a general inaccuracy in prediction even over a relatively short time frame. The person best known for chaos theory is Edward Lorenz and his "butterfly effect" – very small differences in the initial conditions for a dynamical system (e.g., a butterfly flapping its wings somewhere in Latin America), may produce large variations in the long term behavior of the system.

[2]For a popular and fairly recent discussion of the clinical/expert versus statistical/computer comparison, see: *Maybe We Should Leave That Up to the Computer* (Douglas Meingartner, *The New York Times*, July 18, 2006).

[3]To give an anecdote about multiple regression not always being the silver bullet some think it is, one of our colleagues some many years ago, was working on a few creative schemes for augmenting the income of his poverty-stricken quantitative graduate students, and came up with the following idea: he would look for opportunities for statistical consulting anywhere he could, with the offer of him doing the consulting for free as long as the client would then pay the graduate students to carry out the suggested analyses. The first client advertised in the local newspaper for someone with quantitative expertise to assist in a project they had in conjunction with a law enforcement agency. The project involved trying to predict blood alcohol level from various dichotomous behavioral indicators for a driver during the operation of a car. Our colleague got $500 for the graduate student to analyze the data set, which contained an objective blood alcohol level plus a number of 0/1 variables (because of the proprietary nature of the project, the actual meaning of the variables was unknown; they were numbered arbitrarily).

The graduate student ran a version of stepwise regression on a mainframe version of BMDP. Three variables seemed to have some ability to predict blood alcohol level; all the others were more-or-less wash-outs. Our colleague communicated the variable numbers that seemed to have some explanatory power to the project head. Thereupon, a variety of expletives (none deleted) were used. How dare he just come up with these three, when there was obviously such a treasure-trove of other subtle behavioral information available in the data set just waiting to be found. Phrases such as "statistical charlatan" were freely used, but the graduate student still got to keep the $500.

The three dichotomous variables with some explanatory power were:

whether (or not) the driver hit a building; whether (or not) the driver waved at a police officer when passing by; whether (or not) the driver was asleep at the side of the road with the car running. Although not subtle, these seem pretty strong indicators to us.

[4]Like the District Court and the Court of Appeals, the Court seeks to justify the admission of psychiatric testimony on the ground that

"[t]he majority of psychiatric experts agree that where there is a pattern of repetitive assaultive and violent conduct, the accuracy of psychiatric predictions of future dangerousness dramatically rises."

The District Court correctly found that there is empirical evidence supporting the common sense correlation between repetitive past violence and future violence; the APA states that

"[t]he most that can be said about any individual is that a history of past violence increases the probability that future violence will occur."

But psychiatrists have no special insights to add to this actuarial fact, and a single violent crime cannot provide a basis for a reliable prediction of future violence.

The lower courts and this Court have sought solace in this statistical correlation without acknowledging its obvious irrelevance to the facts of this case. The District Court did not find that the State demonstrated any pattern of repetitive assault and violent conduct by Barefoot. Recognizing the importance of giving some credibility to its experts' specious prognostications, the State now claims that the "reputation" testimony adduced at the sentencing hearing "can only evince repeated, widespread acts of criminal violence." This is simply absurd. There was no testimony worthy of credence that Barefoot had committed acts of violence apart from the crime for which he was being tried; there was testimony only of a bad reputation for peaceable and law-abiding conduct. In light of the fact that each of Barefoot's prior convictions was for a nonviolent

offense, such testimony obviously could have been based on antisocial but nonviolent behavior. Neither psychiatrist informed the jury that he considered this reputation testimony to show a history of repeated acts of violence. Moreover, if the psychiatrists or the jury were to rely on such vague hearsay testimony in order to show a "pattern of repetitive assault and violent conduct," Barefoot's death sentence would rest on information that might "bear no closer relation to fact than the average rumor or item of gossip," and should be invalid for that reason alone. A death sentence cannot rest on highly dubious predictions secretly based on a factual foundation of hearsay and pure conjecture.

[5]Although I believe that the misleading nature of any psychiatric prediction of future violence violates due process when introduced in a capital sentencing hearing, admitting the predictions in this case – which were made without even examining the defendant – was particularly indefensible. In the APA's words, if prediction following even an in-depth examination is inherently unreliable,

"there is all the more reason to shun the practice of testifying without having examined the defendant at all. ... Needless to say, responding to hypotheticals is just as fraught with the possibility of error as testifying in any other way about an individual whom one has not personally examined. Although the courts have not yet rejected the practice, psychiatrists should."

Such testimony is offensive not only to legal standards; the APA has declared that "[i]t is unethical for a psychiatrist to offer a professional opinion unless he/she has conducted an examination." The Court today sanctions admission in a capital sentencing hearing of "expert" medical testimony so unreliable and unprofessional that it violates the canons of medical ethics.

[6]There can be no dispute about this obvious proposition:

"Scientific evidence impresses lay jurors. They tend to assume it is more accurate and objective than lay testimony. A juror who thinks of scientific evidence visualizes instruments capable of amazingly precise measurement, of findings arrived at by dispassionate scientific tests. In short, in the mind of the typical lay juror, a scientific witness has a special aura of credibility."

"Scientific ... evidence has great potential for misleading the jury. The low probative worth can often be concealed in the jargon of some expert ... ". This danger created by use of scientific evidence frequently has been recognized by the courts. Speaking specifically of psychiatric predictions of future dangerousness similar to those at issue, one District Court has observed that, when such a prediction

"is proffered by a witness bearing the title of 'Doctor,' its impact on the jury is much greater than if it were not masquerading as something it is not."

In United States v. Addison, the court observed that scientific evidence may "assume a posture of mystic infallibility in the eyes of a jury of laymen." Another court has noted that scientific evidence "is likely to be shrouded with an aura of near infallibility, akin to the ancient oracle of Delphi."

[7]The Court observes that this well-established rule is a matter of evidence law, not constitutional law. But the principle requiring that capital sentencing procedures ensure reliable verdicts, which the Court ignores, and the principle that due process is violated by the introduction of certain types of seemingly conclusive, but actually unreliable, evidence, which the Court also ignores, are constitutional doctrines of long standing. The teaching of the evidence doctrine is that unreliable scientific testimony creates a serious and unjustifiable risk of an erroneous verdict, and that the adversary process, at its best, does not remove this risk. We should not dismiss this lesson merely by labeling the doctrine nonconstitutional; its relevance to the constitutional question before the Court could not be more certain.

[8]In one study, for example, the only factor statistically related to whether psychiatrists predicted that a subject would be violent in the future was the type of crime with which the subject was charged. Yet the defendant's charge was mentioned by the psychiatrists to justify their predictions in only one-third of the cases. The criterion most frequently cited was "delusional or impaired thinking."

[9]"Although jurors may treat mitigating psychiatric evidence with skepticism, they may credit psychiatric evidence demonstrating aggravation. Especially when jurors' sensibilities are offended by a crime, they may seize upon evidence of dangerousness to justify an enhanced sentence." Thus, the danger of jury deference to expert opinions is particularly acute in death penalty cases. Expert testimony of this sort may permit juries to avoid the difficult and emotionally draining personal decisions concerning rational and just punishment. Doctor Grigson himself has noted both the superfluousness and the misleading effect of his testimony: "I

think you could do away with the psychiatrist in these cases. Just take any man off the street, show him what the guy's done, and most of these things are so clear-cut he would say the same things I do. But I think the jurors feel a little better when a psychiatrist says it – somebody that's supposed to know more than they know."

[10]The Court is far wide of the mark in asserting that excluding psychiatric predictions of future dangerousness from capital sentencing proceedings "would immediately call into question those other contexts in which predictions of future behavior are constantly made." Short-term predictions of future violence, for the purpose of emergency commitment or treatment, are considerably more accurate than long-term predictions. The APA, discussing civil commitment proceedings based on determinations of dangerousness, states that, in light of the unreliability of psychiatric predictions, "[c]lose monitoring, frequent follow-up, and a willingness to change one's mind about treatment recommendations and dispositions for violent persons, whether within the legal system or without, is the only acceptable practice if the psychiatrist is to play a helpful role in these assessments of dangerousness." In a capital case, there will be no chance for "follow-up" or "monitoring." A subsequent change of mind brings not justice delayed, but the despair of irreversible error.

[11]The Court's focus in the death penalty cases has been primarily on ensuring a fair procedure: "In ensuring that the death penalty is not meted out arbitrarily or capriciously, the Court's principal concern has been more with the procedure by which the State imposes the death sentence than with the substantive factors the State lays before the jury as a basis for imposing death, once it has been determined that the defendant falls within the category of persons eligible for the death penalty."

Additional Chapter Epigrams

The only function of economic forecasting is to make astrology look good.

– John Kenneth Galbraith

If all else fails, immortality can always be assured by spectacular error.

– John Kenneth Galbraith

I like also the men who study the Great Pyramid, with a view to deciphering its mystical lore. Many great books have been written on this subject, some of which have been presented to me by their authors. It is a singular fact that the Great Pyramid always predicts the history of the world accurately up to the date of publication of the book in question, but after that date it becomes less reliable.

– Bertrand Russell

The list of studies in which the regression factor has been neglected grows monotonous, as well as distressing.

– Philip Rulon (1941)

Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.

– Hillaire Belloc (*The Silence of the Sea*)

Years ago a statistician might have claimed that statistics deals with the processing of data ... today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty.

– H. Chernoff and L. E. Moses, *Elementary Decision Theory* (1959)

Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex.

– Fritz Machlup

When a true genius appears in this world, you may know him by this sign, that the dunces are all in confederacy against him.

– Jonathan Swift

If your mother says she loves you, check it out.

– Adage from the Chicago City News Bureau

It's tough to make predictions – especially about the future.

– Yogi Berra

I would not say that the future is necessarily less predictable than the past. I think the past was not predictable when it started.

– Donald Rumsfeld

Chapter 2

The (Questionable) Use of Statistical Models

The form of statistical practice most commonly carried out by those with a mathematical bent (and in contrast to those more concerned with simple manifest forms of data analysis and visualization), is through the adoption of a stochastic model commonly containing (unobserved) latent variables. Here, some data generating mechanism is postulated, characterized by a collection of parameters and strong distributional assumptions (e.g., (conditional) independence, normality, homogeneous variability, and so on). Based on a given data set, the parameters are estimated, and usually, the goodness-of-fit of the model assessed by some statistic. We might even go through a ritual of hoping for non-significance in testing a null hypothesis that the model is true (generally through some modified chi-squared statistic heavily dependent on sample size). The cautionary comments of Roberts and Pashler (2000) should be kept in mind that the presence of a good fit does not imply a good (or true) model. Moreover, models with (many) parameters are open to the problems engendered by over-fitting and of a subsequent failure to cross-validate. We provide the abstract of the Roberts and Pashler (2000) article, *How Persuasive Is a Good Fit? A Comment on Theory Testing* (*Psychological Review*, *107*, 358–367):

Quantitative theories with free parameters often gain credence when they closely fit data. This is a mistake. A good fit reveals nothing about the flexibility of the theory (how much it cannot fit), the variability of the data (how firmly the data rule out what the theory cannot fit), or the likelihood of other outcomes (perhaps the theory could have fit any plausible result), and a reader needs all three pieces of information to decide how much the fit should

increase belief in the theory. The use of good fits as evidence is not supported by philosophers of science nor by the history of psychology; there seem to be no examples of a theory supported mainly by good fits that has led to demonstrable progress. A better way to test a theory with free parameters is to determine how the theory constrains possible outcomes (i.e., what it predicts), assess how firmly actual outcomes agree with those constraints, and determine if plausible alternative outcomes would have been inconsistent with the theory, allowing for the variability of the data.

A model-based approach is assiduously avoided throughout this monograph. It seems ethically questionable to base interpretations about some given data set and the story that the data may be telling, through a model that is inevitably incorrect, at least at the periphery if not at its core. As one highly cherished example in the behavioral sciences, it is now common to frame questions of causality through structural equation (or path) models, and to perform most data analysis tasks through the fitting of various highly parameterized latent variable models. In a devastating critique of this type of practice, David Freedman in a *Journal of Educational Statistics* article (*As Others See Us: A Case Study in Path Analysis*; 1987, *12*, 101–128) ends with this paragraph:

My opinion is that investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models. Estimating nonexistent parameters cannot be very fruitful. And it must be equally a waste of time to test theories on the basis of statistical hypotheses that are rooted neither in prior theory nor in fact, even if the algorithms are recited in every statistics text without caveat.

The late Leo Breiman took on the issue directly of relying on stochastic models (or, as he might have said, "hiding behind"), in (almost) all of contemporary statistics. What Breiman advocates is the adoption of optimization in place of parameter estimation, and of methods that fall under the larger rubric of (supervised or unsupervised) statistical learning theory. Currently, this approach is best exemplified by the comprehensive text, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition, 2009) (T. Hastie, R. Tibshirani, and J. Friedman). We give two quotes from Breiman. The first is from an invited discussion of a paper from *Statistical Science* by Jim Ramsay (*Monotone Regression Splines in Action* (1988, *3*, 425–441). It reacts to Ramsay's preference for maximum likelihood estimation or a Bayesian approach to the fitting of splines, both of which require

distributional assumptions:

In describing the fitting of the yarn data, the author states that "The fitting criterion could be least squares, but this is not desirable when the dependent variable is being transformed," and he opts for maximum likelihood or Bayesian approaches; in this instance for maximum likelihood. The reason why least squares is not desirable is not stated. Least squares is an old and reliable friend. Maximum likelihood or Bayesian approaches always impose distributional assumptions on the data which are usually difficult to verify. If you clap when Tinkerbell asks "do you believe in fairies" then fine. If you are in doubt, as I often am with real data, then use an earthy friend

The second quote is the abstract from Leo Breiman's *Statistical Science* piece, *Statistical Modeling: The Two Cultures* (2001, *16*, 199–215):[1]

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

The view of statistics to be followed in this monograph is, to use a Breiman term, "earthy." We might go so far as to consider what (linear) regression models can (or cannot) do, or the implications of a basic sampling model, but we go no further than least-squares treated as an algorithmic optimization process, and a suggestion to adopt various sample reuse methods to gauge stability and assess cross-validation. Remembering the definition of a *deus ex machina* – a plot device in Greek drama whereby a seemingly insoluble problem is suddenly and abruptly solved with the contrived and unexpected intervention of some new character or god – we will not invoke any statistical *deus ex machina* analogues.[2]

Stochastic data models do have a place but not when that is only as far as it goes. When we work solely within the confines of a closed system given by the model, and base all inferences and conclusions under that rubric alone (e.g., we claim a causal link because some path coefficient is positive and significant), the ethicality of such a practice is highly questionable. George Box has famously said that "all models are wrong, but some are useful" (or

Henri Theil's similar quip: "models are to be used, but not to be believed"). Box was referring to the adoption of a model heuristically to guide a process of fitting data; the point being that we only "tentatively entertain a model," with that model then subjected to diagnostic testing and reformulation, and so on iteratively. The ultimate endpoint of such a process is to see how well the fitted model works, for example, on data collected in the future. Once again, some type of (cross-)validation is essential, which should be the *sine qua non* of any statistical undertaking.

Notes

[1]One of us (LH) has a personal story about how statisticians believe you can't do anything without an explicitly stated stochastic model. He was giving a paper in the 1990s (in Switzerland, no less) on how an additive-tree representational structure could be fit to a proximity matrix based on an L_1 criterion (i.e., the sum of absolute differences between the given proximities and the obtained additive-tree distances was minimized). It was an elegant approach (or, so he thought), based on a recursive dynamic programming algorithm guaranteeing global optimality for the obtained solution. In fact, he believed this to be quite the coup given that local optimality has plagued proximity representations forever. The first question from the audience after he was done, and from a close colleague at one of his academic stops: "and do you have a stochastic model for the generation of the proximities, where the additive-tree structure could be characterized through parameters that could then be estimated, preferably, through maximum likelihood?" This was before the time algorithmic modelers could be emboldened by Leo Breiman's "Two Cultures" paper. So, LH softly said "no" – and quietly slinked off the stage and back into his seat.

[2]A (slightly) amusing story told in some of our beginning statistics sequences reflects this practice of postulating a *deus ex machina* to carry out statistical interpretations. Three academics – a philosopher, an engineer, and a statistician – are walking in the woods toward a rather large river that needs to be crossed The pensive philosopher stops, and opines about whether they really need to cross the river; the engineer pays no attention to the philosopher and proceeds immediately to chop down all the trees in sight to build a raft; the statistician yells to the other two: "stop, assume a boat."

Additional Chapter Epigrams

As a single atom man is an enigma; as a whole he is a mathematical problem. As an individual he is a free agent; as a species the offspring of necessity.

– Winwood Reade, *The Martyrdom of Man* (1872)

It is commonly believed that anyone who tabulates numbers is a statistician. This is like believing that anyone who owns a scalpel is a surgeon.

– R. Hooke, *How to Tell the Liars From the Statisticians* (1983)

There is no more common error than to assume that, because prolonged and accurate mathematical calculations have been made, the application of the result to some fact of nature is absolutely certain.

– Alfred North Whitehead

Humanist as I claim to be, I do not deplore this [quantification] trend so far as it has gone. Quantification has increased, is increasing, and in my opinion ought not to be diminished but to stay — quantitatively just about where it is now. ... This is, like it or not, the Quantified Age. Better to ride the waves, if one has sufficient finesse, than to strike attitudes of humanistic defiance and end ... in the dustbin of history.

– O.H.K. Spate (1960)

Golomb's don'ts of mathematical modelling:

Don't believe in the 33rd order consequences of a 1st order model. Catch Phrase: 'Cum grano salis.'

Don't extrapolate beyond the region of fit. Catch Phrase: 'Don't go off the deep end.'

Don't apply any model until you understand the simplifying assumptions on which it is based, and can test their applicability. Catch Phrase: 'Use only as directed.'

Don't believe the model is reality. Catch Phrase: 'Don't eat the menu.'

Don't distort reality to fit the model. Catch Phrase: 'The Procustes Method.'

Don't limit yourself to a single model: More than one may be useful for understanding different aspects of the same phenomenon. Catch Phrase: 'Legalise polygamy.'

Don't retain a discredited model. Catch Phrase: 'Don't beat a dead horse.'

Don't fall in love with your model. Catch Phrase: 'Pygmalion.'

Don't apply terminology of subject A to the problems of subject B if it is to the enrichment of neither. Catch Phrase: 'New names for old.'

Don't expect by having named a demon you have destroyed him. Catch Phrase: 'Rumpelstiltskin'

If you think you need a model to have your data tell you what is there, think again –

– Larry Hubert

Statisticians, like artists, have the bad habit of falling in love with their models.

– George Box

... the statistician knows ... that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.

– George Box