Koger et al. v. Reno, 1994 WL 116142, D. D. C., March 15, 1994 (Oberdorfer, J), *aff'd.* 98 F. 3rd 631 (D.C. Cir. 1996)

Older GS-11 United States Marshals complained, under the Age Discrimination In Employment Act (ADEA), that they were under-selected for promotion to GS-12 positions. This case is a classic example of an "expert" not knowing how to model selections, one of the most frequent of litigation subjects. Mary W. Gray, a professor of mathematics at American University (with, in addition, a degree in law), was engaged by plaintiffs. I do not mean to embarrass her, but she is named in the opinions. She does leave, as her legacy, this example of the mistakes one makes when not understanding how defendant institution works, and not modeling her clients' complaint in that context.

To be promoted, a Marshal had to apply to a specific announcement. Applications contained information from which applicants were scored. The highest scoring three applications (or more if the announcement contained more than one position, or if there were ties for third place) were " certified." The selecting official chose the winner from those certified, not from all applicants. This two-stage process is the most common method of competitive hiring and promotion within the federal government. I use the word "competitive" to mean that there is a fixed number of openings and an application closing date. Non-competitive promotion is also often available in the federal government through promotion plans, wherein an employee must achieve certain standards, but cannot be denied a promotion because there is not a "vacancy." An expert analyzing federal government promotions must first determine whether promotions were competitive or non-competitive, because they call for different methods of analysis. See *Harrison v. Lewis* as an example where plaintiffs' expert did not understand this distinction.

Older marshals were under-selected, plaintiffs claimed, by the Service's reliance in scoring their applications upon a physical "fitness" test, the primary component of which was a 1.5 mile run. Secondary reasons also concerned scoring, in particular that education had lower weight the further in the past it was obtained. These complaints go to certification, a stage Dr. Gray did not analyze.

Plaintiffs asked the Marshals Service for data indicating who applied for each vacancy announcement, their scores, and the outcomes of the process. They were provided with photo copies from boxes of vacancy announcements. Each announcement packet contained the announcement itself, scoring sheets, a certification list (indicating also who was selected), and applications. Some of these announcements had been coded and entered by the Marshals Service. Printouts of this coding were provided. Dr. Gray key entered the data from all this paper, herself.

As is unfortunately common, LRA was brought into the case long after the delivery of data to plaintiffs. The Service had made only one copy of vacancy announcement folders, which they gave to plaintiffs. They did not retain even a log of the materials provided to plaintiffs. The boxes they copied from continued to be used (announcements pulled, results updated, etc.) subsequent to the data delivery.

Thus, although plaintiffs did not know it, I did not have information about the same vacancy announcements they had, nor necessarily the same information when I had the announcement. *We did not have the same data.*

Plaintiffs were rightly concerned about their own data, as that is one area in which defendant has a natural advantage. They were relieved when I did not question their data, especially given the casual (and incorrect) way in which they had key-entered it. As my data came from a different collection of vacancy announcements than plaintiffs', the proper procedure would have been to allow me to review plaintiffs' documents, and also to allow plaintiffs to review defendant's current data.[1] Then either the experts would have arrived at a common data base, or we would have had sufficient information to debate the merits of the two data bases before the court.

Gray's data set did not distinguish among vacancy announcements. Not understanding the multiple pools aspect of the selection process, or how to model it, Gray skipped over the certification stage. She then applied a methodology that assumes that every applicant competed for every position. The two experts applied different analytic methods to different data sets, and both blamed the difference in results on the methods. I believe this conclusion was essentially correct, but with only my data, I cannot actually know it.[2] The adversarial process, which is supposed to compel the parties to bring good data into the court room, will not always do so.

[1] In Mary Gray's deposition, defendant requested her computerized data, which she agreed to provide. Subsequently, plaintiffs refused to provide her data, so defendant refused to provide mine. This case was tried under rules of evidence existing in 1991, which did not require a data exchange. However, exchange of computerized data would not have sufficed, as I used information that Gray did not key enter. I needed access to the photocopies.

[2] Of course I could and did replicate Mary Gray's methods on my data. Only in one instance, which I reported, did I get approximately the same result that she obtained.

For each applicant she entered each component score, age, and whether the applicant was selected or not—but not which applicants had been certified.[3] Nor did she enter applications when all applicants were over age 40 or under age 40, mistakenly believing no information was contained in that data. She performed the following analyses:

- Mean Values. Dr. Gray divided her data into yearly groupings. Dr. Gray performed a t-test on the mean ages of winners and losers. She reports "no significant difference in the ages for the two groups," indicating that winners "look like" losers (my language), which would imply that winners are selected without regard to age, at least within the range of two standard deviations.[4]
- Correlation. Dr. Gray correlated age with score components, reporting only those in which she found a "significant" negative correlation.
- Regression. Dr. Gray used the binary variable indicating selection (or not) as her dependent variable in a linear regression. Her independent variables were score components and age. She reported a negative coefficient for age. She did not report other coefficients.

Plaintiffs presented the following "findings:" Older applicants are not under-selected on the face of it, that is, not accounting for any other variables. However, when their scores on other variables are accounted for, they are under-selected. That is, they are under-selected given their characteristics other than age. With plaintiffs agreeing that older applicants were selected in proportion to their rate of application, Judge Oberdorfer granted summary judgment for defendant on plaintiffs' claim of disparate impact.[5]

The case went to trial on plaintiffs' disparate treatment claim: older applicants who otherwise "look like" younger applicants are under selected based on their age.[6] This logic is exactly the opposite of that

[3] As the application process takes some time, age should be determined on some consistent date. I presume the selection date would be relevant under the law, as it is the date of the employer's action. Because selection was determined from a retrospective entry on the certification list, the date of selection was not always known. I used the certification date to calculate age, as each certification list had a date. I also had access to computerized personnel data which contained the date of the promotion, but not the date of the decision to promote. Mary Gray did not use a consistent date, sometimes using an age provided by the Marshals Service without inquiring about its basis.

[4] This summary is a quotation from my report to the court, "Age and Promotion of U.S. Marshals To GS-12 Senior Criminal Investigator," January 25, 1994, pages 6-7.

[5] Order of March 15, 1994.

[6] The written opinion dismissing plaintiffs' disparate impact claim, although "unpublished," is available in FEP Cases and Westlaw. I can find no copy of the more interesting opinion, based on the trial.

posited by plaintiffs in their complaint, that older applicants are harmed by the run test and conventions of scoring. That is the disparate impact complaint plaintiffs lost before trial. Thus Gray was bringing her clients down a path that bore no relationship to their original argument. Where did she go wrong?

Gray appeared to make two serious technical errors. Those errors led her to believe that older applicants fared badly, controlling for score, when in fact, if they scored high enough to be certified, they fared well in selection.

Her first error was that the ADEA does not recognize age variation under age 40. Plaintiffs first must choose an age above which they complain discrimination has occurred. They need not specify age 40, but may not specify an age below 40. In this case, they selected age 40. Above that age, one can still count as "discrimination" the selection of a younger applicant instead of an older applicant. Thus, even if 40 is the demarcation, a 45 year old winner can be evidence of discrimination if the losers were over 45. However, a 30 year old winner does not support a complaint of age discrimination if no loser was over age 40, even if all losers were older than 30. All applicants under the selected age, here 40, have the same age in the eyes of the law.

To my knowledge, this case represents the first instance in which this link between the treatment of age under ADEA and appropriate variable configuration was made. Some analyses of age discrimination are based on binary variables, losing the variation above the denoted age. Other analyses incorrectly allow age variation wherever it occurs. The incorrectness of such analysis is not a difficult point to attorneys, only to experts.[7] When pointed out to Judge Oberdorfer, it was immediately accepted:

> In addition, Dr. Gray performed a regression analysis in which variation in age under age 40 affected her conclusions. However, such a variation is no basis for liability under the Age Discrimination in Employment Act.[8]

The need for the expert to understand the nuances of the law, with regard to the specifics of the statistical analysis, cannot be over-emphasized. Much though lawyers denigrate the expert's struggles with the law, the experienced expert usually knows more about the law than the lawyer knows about statistics. Few lawyers would understand, without being told by an expert, that Mary Gray's regression and correlation analyses were incompatible with applicable law. Obviously Mary Gray's client did not perceive this error. My client, an Assistant United States Attorney, grasped it immediately upon my pointing it out, and comfortably conveyed it to the judge during Mary Gray's cross examination. The judge then confirmed the basic contention with the witness, and excluded further testimony based on analyses utilizing a continuous age variable.

[7] For example, Harriet Zellner argues to use a continuous measure of age (claiming that experts incorrectly use binary variables) in an article "When Is It Really Age Discrimination?" *New York Law Journal*, November 4, 1993. Zellner's continuous variable would be incorrect.

[8] Typescript of June 6, 1994 Memorandum opinion, p. 17.

As Gray had secondarily converted age into a binary variable in which there was no variation either below or above age 40 (i.e., all older applicants appeared to be of the same age, and all younger applicants appeared to be of the same age, but older and younger were distinguished) she was still allowed to present analyses based on that variable. Doing so, however, eliminated those cases in which applicants over 40 were selected while applicants further over 40 were not. Thus not only were Gray's primary analyses fatally flawed through failure to incorporate the law, her secondary analyses were a poor cure.

Gray's second fundamental error was ignoring the institutional phenomenon of vacancy announcements. In her analyses, all applicants were competing for all openings. Indeed, as a Marshal wanting promotion had to apply for each position sought, in Gray's analysis any Marshal who applied for more than one opening would compete against himself for promotion. A statistical analysis of employer decisions should be consistent with the institutional context in which those decisions were being made, as the judge understood:

> Dr. Gray ignored the requirement that selections be made from among those in individual vacancy announcement pools. Thus, her analysis treated as "losers" applicants who had high scores in comparison to the entire pool of applicants for all vacancies but whose scores were not sufficient to gain promotion to the particular position for which the applicants applied.[9]

What is called for is an analysis that calcualtes parameters from individual pools of applicants, but summarizes them over all such pools. Although when I started analyzing employment situations no analyst understood this problem, by March, 1994, I would have expected all analysts to understand the problem *and its solution*. Furthermore, by 1994, most statistical packages (including *Systat*, which Gray used) had appropriate software, though usually only as an add-on option.

[9] Typescript of June 6, 1994 Memorandum opinion, pp. 16-17.

As Gray had analyzed selection from applicants, with and without controls for component scores, I followed her lead, revising her methods to comport with the law. I defined age as age if the applicant was 40 or above, and the mean of all under 40 applicants if the applicant was below age 40. In this measure there was no variation in age among younger applicants, so a younger winner "looked like" a younger loser regardless of their actual ages, whereas older winners and losers kept their real ages.[10] I used multiple pools methods, in which an equation is fit to the data estimating the relationship between age and selection, allowing competitions only within vacancy announcements. The multivariate method is called "conditional logit," a multiple pools form of logit analysis, which Gray should have used (instead of regression) for her multi-variate analysis.[11]

When the applicants are compared only within vacancy announcements, and age variation below 40 is eliminated (and using my data which, as noted above, was not identical to plaintiffs'), I found:

- Older applicants were disproportionately favored in selection.
- Given their application scores, older applicants were selected in the same proportion as younger applicants.

The implication of these findings, which was brought out in my cross examination, was that older applicants had lower scores than younger applicants. When we know their scores, our "expectation" of the success of older applicants is lowered, so success appears proportional, not excessive. Although this finding directly contradicts plaintiffs' own expert's findings, it supports their argument. It is also a correct description of the facts.

The case ended with testimony from the two experts. Both analyses were of promotion from applicants, the framework set out by plaintiffs' expert. I do not hold the attorney responsible for understanding nuances such as the difference between binary and continuous variables, even though most of Gray's work was dismissed at trial for her failure to understand the implication of her continuous age variable in the ADEA. But I do hold attorneys responsible for understanding how the institution at issue works. Gray's most fundamental error, one her client should have understood and corrected, was ignoring the certification stage of the promotion process.

[10] In setting the fixed age of younger applicants at their mean, to avoid arbitrarily picking a number, I let actual ages below 40 affect the scale, the distance between "young" and "old." Judge Oberdorfer suggested setting "young" at an age just below 40, say, 39 or 39.9. That might be a better procedure than mine. Rank statistics might be preferred, in which all applicants under age 40 would be tied for lowest age without specifying what age that was.

[11] Although there is much technically wrong with Gray's approach, logit software is easy to come by and use. I cannot see why she would use regression in this circumstance, although the two procedures almost always produce the same (qualitative) result. One can devise a multiple pools algorithm for regression. Her error was not understanding multiple pools, not her use of regression with a binary dependent variable,

In *Connecticut v. Teal*, 457 U.S. 440 (1982) a "bottom line" analysis of hires from applicants showed that blacks were hired proportionate to their application. However, they failed an initial exam at an excessive rate. The question was whether excessive selection of blacks from the remaining candidates compensated blacks for the disparate impact of the test. The court's answer was that the individual blacks who failed the test may have been discriminated against (unless the test was shown to have been valid), regardless of the outcome of other blacks.

Koger was as close to the facts of *Teal* as one is likely to get. It should have been analyzed by plaintiffs along the lines suggested by *Teal*. Rather than asking "who was selected," Gray should have asked "who was certified?" Doing so probably would have shown that the run test did have a disparate impact favoring younger applicants.

The Marshals Service operated just as anyone familiar with such institutions would have expected. There was an "old boy" network, in the most literal sense. An older applicant who could survive to a certification list was likely to be selected for promotion—more likely than a similarly situated younger applicant. However, just as plaintiffs claimed, older applicants were less likely to be certified, largely because of their failure on the run test.[12] This is *Teal* reincarnate.

I had warned my client that although the Service looked good on a "bottom line" view, it might not if plaintiffs's expert followed *Teal* and tested for who reached the next level. I also told them that Gray had not done so, and probably could not, because she had not key entered certification information. But I advised them to stand ready to defend the run test as valid. They did.

There are several important stories in this case. Older Marshals were indeed disproportionately eliminated from consideration for selection through failure to score well on the physical examination. Plaintiffs' contention, as far as it went, was correct. The next step would have been to ask if fitness is a valid requirement for the position, and if the run test is a valid indicator of fitness. Plaintiffs' failure to pursue these questions would have doomed them even had their expert properly discovered the facts.

With a correct analysis, could plaintiffs have prevailed? The defense did justify the relationship between fitness and job performance on the one hand, and the relationship between the run and relevant fitness on the other, as incidental information, but also probably well enough to assert validity. The judge, who is guarded by Marshals, seemed inclined to prefer that his guards be physically fit. It may thus have been that plaintiffs could not have won this case. But they never gave it a shot.

[12] Plaintiffs' allegations about the scoring of education were not tested. To do so would have required data on those components of education plaintiffs claimed were under-valued, and the proposal of an alternative scoring system. Older applicants did not score noticeably lower than younger applicants in education, but whether they should have scored higher I cannot say. Had Gray understood and followed her constituents' complaints, she might have formulated a very different analysis.

There is not always a single “correct” technical approach to a problem. An expert whose analysis is found inferior by a judge may actually have done capable work. On technical issues the expert is alone. His client relies on him to do his job. The best job may not support the expert’s client, or may not be understood to be superior by the judge. Although these are problems to be explained and understood, they were not problems in this case.

In contrast, an expert who does not perform a *conceptually* correct analysis of the issues presented by his clients has not failed alone. This is the area attorneys and experts should be discussing. Does the expert understand all prevailing law? Has that law been incorporated into the analysis? An attorney who does not spend the time with his expert to determine whether the expert’s analysis correctly portrays the institution, and the law, is not doing his job as an attorney. An expert who does not explain his approach and ask if it comports with the theory of the case and the law as his client understands it, is not doing *his* job.

Finally, it should be noted that Judge Oberdorfer understood and made a correct ruling on one issue I raised. And then the D.C. Circuit court incorrectly reversed him! As far as I know, this is the only time that subject has been adjudicated *de nouveau*, other courts referring to this incorrect ruling to support their own. That issue is whether Gray was correct to eliminate vacancy announcements in which there were only young or old applicants. The appeals court’s rationale was that, if there was no age “competition” (remembering that Gray had eliminated age variation) that application’s data could not inform the court. That would be correct if age were the only variable in the analysis. But when there is at least one other variable (here, certification score), how it affects selection should be determined from all vacancy announcements. In a multi-variate analysis, all competitions should be in the data to be analyzed.

On the published record, plaintiffs’ expert’s failure in *Koger* appears to be one of statistical method. That is not the lesson to learn. The failure on plaintiffs’ side was that neither the expert nor her attorney client understood how to formulate an analysis relevant to the institution and the law.

In the spirit of O Henry, there is a final reason why this case so well demonstrates the failure of the expert, the attorney and the attorney-expert relationship: *Teal v. Connecticut* was cited by plaintiffs in their pleadings! They did not fail to know about the case, and its relevance to *Koger*. They just failed, utterly, to connect it to the work of the expert.