The Strangely Persistent "Transposition Fallacy":
Why "Statistically Significant" Evidence of Discrimination
May Not Be Significant

Kingsley R. Browne*

I. Introduction

Statistical evidence has long played a prominent role in employment discrimination cases. Such evidence is de rigeur in both pattern-and-practice cases¹ and disparate-impact cases² and is sometimes used as an adjunct in individual disparate-treatment cases.³ After decades of experience with statistical proof, many courts seem to have become relatively comfortable with statistical methods. This increased level of comfort has not always been accompanied by an increased level of statistical sophistication, however, and courts routinely make substantial errors in their interpretation of statistical evidence.⁴ The topic of this article is one particular error that courts very frequently make in discrimination cases, what has been dubbed in the statistics literature the transposition fallacy.⁵

As a general matter, the transposition fallacy involves equating "the probability of A given B" with "the probability of B given A." In the discrimination context, one commits the transposition fallacy by

---

*Kingsley R. Browne is a professor at Wayne State University Law School. E-mail: kbrowne@novell.law.wayne.edu. The author gratefully acknowledges the valuable comments of Eric Kades, David Kaye, and Michael McIntyre. Any remaining errors are, of course, solely those of the author.

equating the probability of an observed disparity given random selection with the probability of random selection given the observed disparity.\(^6\) Stated in this way, the mistake may appear to be an arcane technical mistake of interest to statistical purists but of no particular significance to real-world outcomes. In fact, however, commission of the fallacy almost certainly has a substantial impact on litigation outcomes because it causes courts to attribute unwarranted significance to a plaintiff's showing of "statistically significant" workforce disparities.

II. The Transposition Fallacy in Action

There is no better way to come to an understanding of the fallacy than through a simple concrete example. Suppose an employer has a workforce of 400 entry-level employees and that blacks constitute 20 percent of the labor force in the relevant labor market. One would "expect" statistically that 80 (20 percent) of the 400 employees would be black. On the other hand, we would no more be suspicious of the employer's selection practices because it did not have precisely 80 black employees than we would be suspicious of the fairness of a coin simply because it did not yield exactly 50 "heads" and 50 "tails" in a series of 100 tosses. We understand that, over the course of a large number of trials, the results will tend to cluster around the expected number, but in any given trial there may be a small, or even a relatively large, disparity between observed and statistically expected results.

The distribution of results, whether the number of blacks in employers' workforces or the number of heads in a series of coin tosses, can be represented graphically by the familiar "normal"—or "bell-shaped"—curve. Figure 1 is a representation of the results that would be generated by an infinite (or very large) number of employers of 400 employees selecting employees randomly with respect to race from an infinite (or very large) pool of employees in which blacks constitute 20 percent of the pool. The results cluster around the mean of 80 black employees, and as one moves farther away from the mean in either direction the number of employers with such results becomes increasingly smaller. A characteristic feature of the normal curve is that through calculation of the "standard deviation" one can calculate the percentage of the distribution that falls under any given part of the curve. Specifically, here, one can calculate the probability that an employer selecting at random will obtain a particular result. In the example, 68% of employers will have workforces with between 72 and 88 black employees, and less than 5% of employers will have fewer than 64 or more than 96 black employees.

When an employer has fewer blacks than statistically expected, the probability that is typically of interest is the probability of obtaining

\(^6\) Zeisel & Kaye, supra note 5, at 82.
by chance a racial distribution that deviates at least that much from the expected distribution. Thus, the statistical question is whether the disparity is so great that we think it is not just a result of sampling error but is instead attributable to some systematic, nonrandom cause—perhaps, but not necessarily, discrimination. In other words, we want to know if the disparity is "statistically significant." The level at which results will be labeled statistically significant—the "significance level"—should be set prior to conducting the statistical analysis. The most common significance level used in employment litigation is five percent, although sometimes a one-percent level is used.\[7. \text{See generally Ramona L. Paetzold & Steven L. Willborn, The Statistics of Discrimination: Using Statistical Evidence in Discrimination Cases § 2.04 (1994).} \]
\[8. \text{Although determination of "statistical significance" is routine in litigated cases, David Kaye has persuasively argued that because adoption of a particular level of statistical significance is arbitrary, testimony concerning whether results are significant should be supplanted by testimony concerning the size of the p-value. See David H. Kaye, Is Proof of Statistical Significance Relevant?, 61 Wash. L. Rev. 1333, 1343-45, 1362 (1986).} \]
In applying the above analysis, the statistician may use the technique of "hypothesis testing" which would test the null hypothesis that there is no "real" racial difference in the probability of selection—that is, that observed differences in outcome are merely a consequence of sampling error.\footnote{Mikel Aickin, Issues and Methods in Discrimination Statistics, in D. H. Kaye & Mikel Aickin, Statistical Methods in Discrimination Litigation 159, 164-68 (1986).} The alternative hypothesis would be that there is a racial difference in the probability of selection, that is, that the employer's decisions are not random with respect to race.\footnote{It should be emphasized that "not random with respect to race" does not necessarily mean "based on race." If the employer selects on the basis of a factor that is correlated with race—test scores, for example—the decision would not be random with respect to race but it would not be "based on" race.} If the difference in selection rates for the two groups is statistically significant, the statistician would reject the null hypothesis and conclude that the probability of selection for members of the two groups is not the same.\footnote{This discussion assumes that a two-tailed test is being used. Under a two-tailed test, the null hypothesis is that there is no racial difference (i.e., blacks have the same probability of being hired as whites do), and the alternative hypothesis is that the probability of being hired is different, although the difference could run in either direction. See Richard Goldstein, Two Types of Statistical Errors in Employment Discrimination Cases, 26 Jurimetrics J. 32, 45-47 (1985). Under a one-tailed test, the alternative hypothesis is that the difference runs in a particular direction (e.g., blacks have a lower probability of being hired than whites). Because a two-tailed test generally produces probabilities that are twice the size of a one-tailed test, equivalent deviations are more likely to be labeled statistically significant under a one-tailed test than they are under a two-tailed test. David H. Kaye & David A. Freedman, Reference Guide on Statistics, in Federal Judicial Center, Reference Manual on Scientific Evidence 331, 382-83 (1994).} If the difference in rates is not statistically significant, the null hypothesis could not be rejected and the statistician would conclude that the statistics do not demonstrate biased selection.\footnote{It is important to note that one does not "prove" the null hypothesis. That is, the fact that the difference is not statistically significant does not prove that the employer has not discriminated, only that discrimination is not demonstrated by the statistics.}

In our illustration, assume that the employer has just 62 blacks rather than the statistically expected 80. Employing the normal approximation of the binomial distribution,\footnote{See Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers 120-22 (1990).} it turns out that the standard deviation is 8.\footnote{The formula for the standard deviation of the binomial distribution is:
\[ s.d. = \sqrt{n \cdot p \cdot (1-p)} \]
where \( n \) is the number of trials, and \( p \) is the probability of obtaining the result in a given trial, or here:
\[ \sqrt{400 \times .20(1-.20)} = 8 \]} What we want to determine is the \textit{z-score} associated with the employer's distribution, which is the number of standard deviations the employer's distribution differs from the expected. A deviation

\[ \sqrt{400 \times .20(1-.20)} = 8 \]
of 18 employees produces a z-score of \(-2.25\),\(^{15}\) the minus sign denoting that the employer has fewer blacks than expected. Reference to a probability table for the normal distribution, which can be found at the back of any statistics text, reveals that the probability associated with a z-score of \(-2.25\) is approximately .0244 or 2.44%. This probability is commonly referred to as the \(p\)-value.

This article concerns the meaning of the \(p\)-value, a meaning that courts routinely misinterpret. In the example above, less-careful courts would conclude from the statistical showing that there is only a 2.44% chance that the disparity is a consequence of a nondiscriminatory cause; in other words, there is a 97.56% chance that the employer discriminated.\(^{16}\) Such a showing would, of course, amply satisfy the plaintiff's obligation of proof by a preponderance of the evidence. Many courts attempting to be careful would say that there is only a 2.44% chance that the underrepresentation of blacks was a product of random selection but that the statistics do not tell us whether the nonrandom cause was discrimination or something else.\(^{17}\) However, having ruled chance out as the cause, the court would typically call upon the employer to identify the responsible nonrandom cause.\(^{18}\) Although one appreciates the reservation of the more-careful courts, both the more- and the less-careful courts are committing the transposition fallacy.

Why does the 2.44% figure not represent the likelihood that the result was obtained by chance? The answer requires one to reflect on the nature of the distribution described by the curve from which the probability was calculated. It was a distribution of nondiscriminating employers. While the probability figure does represent the probability of an individual nondiscriminating employer's obtaining such a distribution, it does not reveal the likelihood that an employer with such a distribution obtained it by chance. That is, the \(p\)-value does not reveal the probability that our employer did not in fact come from the distribution of nondiscriminating employers from which the \(p\)-value was com-

---

15. The formula for the z-score is:

\[
z = \frac{\text{observed} - \text{expected}}{\text{s.d.}}
\]

or

\[
z = \frac{62 - 80}{8} = -2.25
\]

16. See, e.g., Ivy v. Meridian Coca-Cola Bottling Co., 641 F. Supp. 157, 165 (S.D. Miss. 1986) (stating that "a fluctuation of two or three standard deviations indicates that the result is caused by discriminatory intent rather than chance").

17. See, e.g., Lilly v. Harris-Teeter Supermarket, 720 F.2d 326, 336 (4th Cir. 1983) (stating that disparities of more than two or three standard deviations "conclusively ruled out chance as the cause of the disparity in the termination rates"), cert. denied, 466 U.S. 951 (1984).

puted but rather from a wholly different distribution of discriminating employers.

A simple coin-flipping example may assist in understanding the fallacy. If 100 people each flip a coin 100 times, the distribution of the number of heads obtained by each person will describe a normal curve with a mean of 50 and a standard deviation of 5. Suppose attention focuses on a hapless subject who obtains 61 heads. We may suspect that the coin is not fair, as we have eliminated the possibility that he is misreporting his results by carefully observing his tosses. Using the statistical analysis described above, we would calculate the probability of obtaining 61 or more heads or tails in 100 tosses as .0278,\(^\text{19}\) or 2.78%. What is the likelihood that the coin is not a fair coin? If we employ the transposition fallacy, we would say that the likelihood that the coin is fair is only 2.78%; the likelihood that the coin is unfair would therefore be 97.22%. Yet, we know that some statistically significant deviations are to be expected even if we use a fair coin, and from the facts given we have no way of saying what the probability is that this particular subject was using an unbalanced coin. If we knew that the experimenter had “salted” the pool of coins with some unbalanced coins, our subjective view of the probability might increase, while if we knew that unbalanced coins were very rare, we might discount the likelihood. In either event, however, we would not be guided by the 2.78% probability that the standard analysis would produce.

To take a different kind of example, suppose that a camper enters a park and is told by the ranger at the entry gate to be careful not to pitch his tent under a great sap-oozing oak because of the noxious exudate produced by the tree. The ranger describes the tree but says that the risk is low because out of perhaps 100,000 trees in the park, only approximately 100 are the dangerous tree. The camper looks for a place to pitch his tent away from the sap-oozing but because it is dark by the time he finds a spot he cannot tell if he is under the tree in question. When he wakes up in the morning, he finds that his tent is covered with a sticky viscous liquid that he is sure will never come out of the tent. What is the likelihood that the ranger lied to him about the prevalence of the sap-oozing? Applying the transposition fallacy, we would say that if the ranger was telling the truth, the likelihood of camping under a sap-oozing is 1 in 1,000, and therefore the probability that the ranger was telling the truth is 0.1% and the likelihood that he was lying is 99.9%. Yet, most people would not reason in that way. Instead, they would probably say that the camper was just unlucky; he simply picked the wrong tree. Although we could construct facts that might cause us to increase our estimate of the likelihood that the ranger lied—if, for example, the

---

\(^{19}\) Employing the same formula used before, the z-score is 2.4.
ranger knew that the camper was having an affair with the ranger's wife—the probabilities by themselves would not lead us to suspect the ranger.

The essence of the fallacy, defined at the outset, now becomes more clear. If a court were to conclude that there was only a 2.44% chance that the employer obtained only 62 black employees by chance, it would mistakenly equate the probability of obtaining that result under a system of random selection with the probability of random selection given the observed result. These two probabilities are simply not equivalent.

III. A Graphical Demonstration of the Error of Equating a P-value with the Likelihood of Random Selection

Courts err in concluding that the probability of random selection can be derived from the distribution of randomly selecting employers. Rather, in order to calculate such a probability one would also have to know something that one never knows, which is the distribution of discriminating employers, for the ultimate fact that we want to know is which category—that of discriminating employers or nondiscriminating employers—the particular employer comes from.

To illustrate the lack of meaningful relationship between the significance level and the likelihood of discrimination, consider the following example. Assume that there are 10,000 large employers. The question for any given employer with a statistically significant disparity is "what is the probability that the disparity was obtained by chance?" Assume, first, that none of the employers in the population engages in discrimination (that is, that the "base rate" of discrimination is zero) and that all hire at random with respect to race. By definition, one would expect 500 (5 percent) of them to have disparities that are significant at the 5% level. (See Figure 2). What is the likelihood that any one of those

| Base Rate = 0 |
| % Discriminatory Disparities = 0 |

![Figure 2](image)
employers with a disparity obtained it through discrimination? The answer is 'zero,' because we already know by hypothesis that no employers discriminate.

Now change the assumption so that one percent of all employers engage in systematic discrimination. What is the likelihood that a randomly selected employer with a racial disparity obtained the disparity through discrimination? One hundred (1%) of the employers will have statistically significant disparities due to discrimination, while 495 (5% of the remaining 9900) will have disparities due to chance. (See Figure 3). Thus, 100/995, or approximately 16.8 percent, of the disparities will have been caused by discrimination.

If five percent of all employers engage in systematic discrimination, then 500 employers will have disparities due to discrimination and 475 employers (5% of the remaining 9500) will have chance disparities. Therefore, 500/975, or 51.3%, of employers with disparities will have obtained them by discrimination. Employing the same reasoning, if 10% of employers engage in discrimination, 69.0% of disparities will have been caused by discrimination (see Figure 5), and if 50% of employers engage in discrimination, then 95.2% of them will have been caused by discrimination (see Figure 6).

20. For sake of clarity of discussion, assume that all disparities not caused by chance were caused by discrimination and that all employers who discriminate would have statistically significant disparities.

21. Statistically proficient readers will recognize that the argument provided here relies implicitly on Bayesian inference. See Jonathan J. Koehler & Daniel N. Shavio, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use Of Overtly Probabilistic Evidence and Methods*, 75 Cornell L. Rev. 247, 255-56 (1990). Bayesian methods allow direct calculation of the probability of the hypothesis given the evidence (p(H | E)), rather than the probability of the evidence given the hypothesis (p(E | H)), which is closer to the probability that the standard analysis provides. The formula for calculating p(H | E) is:

\[ p(H | E) = \frac{p(H)p(E | H)}{p(E)} \]
**Base Rate = 5%**

\%

% Discriminatory Disparities = 51.3%

![Pie Chart](chart1)

**Figure 4**

**Base Rate = 10%**

\%

% Discriminatory Disparities = 69.0%

![Pie Chart](chart2)

**Figure 5**

where \( p(E) = p(H)p(E|H) + p(-H)p(E|-H) \). *Id.* at 255 n.27. For simplicity, we consider the evidence to be that the disparity is significant at the .05 level, without regard to the actual \( p \)-value of the disparity for a randomly selected employer.

Using the data from figure 4, where the base rate of employer discrimination was 10%, the relevant terms are:

- \( p(H) \) = the "prior probability"; that is, the probability of the hypothesis that the employer discriminated without taking into account whether the employer in question has a significant disparity. In this example, the prior probability is provided exclusively by the base rate of discrimination = .10
- \( p(-H) \) = the prior probability that the employer did not discriminate = .90
- \( p(E|H) \) = the probability of a significant disparity if the employer discriminates = 1
- \( p(E|-H) \) = the probability of a significant disparity if the employer does not discriminate (the significance level) = .05
- \( p(E) \) = the probability that any random employer will have a significant disparity = \( (.10)(1) + (.9)(.05) = .1 + .045 = .145 \)

Therefore, \( p(H|E) \), which is the probability that the employer discriminated given the observed disparity, equals

\[
\frac{.10(1)}{.145} = .69
\]
Base Rate = 50%
% Discriminatory Disparities = 95.2%

Figure 6

Thus, the probability that a disparity was obtained other than by chance ranges in these examples from zero to 95 percent even though in each case the standard analysis would yield exactly the same 5% probability of random selection. However, only by assuming that half of all employers engage in discrimination as their standard operating procedure can one obtain a figure equivalent to the figure suggested by the conventional analysis.\(^\text{22}\)

which is the same probability calculated using the more intuitive approach described in the text. The actual application of Bayesian methods of inference to legal decision-making is controversial, in large part because of difficulties in specifying the "prior" probabilities, id., and is beyond the scope of this article.

22. Arguably, one might justify continued misuse of the tools of statistical inference on the ground that we should assume that the base rate of discrimination is high. See Michael J. Zimmer et al., Cases and Materials on Employment Discrimination 284 (4th ed. 1994). However, if a plaintiff's statistical analysis rests on an assumption that fifty percent of employers discriminate—which is what would be necessary to justify the level of confidence in the meaning of statistical disparities that courts now display—then the assumption should be made explicit and be empirically supported by the plaintiff, who bears the burden of proof. Empirical support would probably not be forthcoming, however, since the assumption that fifty percent of all employers engage in discrimination as their "standard operating procedure—the regular rather than the unusual practice," International Bhd. of Teamsters v. United States, 431 U.S. 324, 336 (1977), is almost certainly wrong. No doubt, the base rate of discrimination varies by region and by industry and with respect to whether the discrimination is based on race, sex, national origin, religion, age, etc., factors that courts never take into account in interpreting the p-value.

In any event, once it is understood that the overall probability that a disparity was caused by discrimination is a function of the rate of employer discrimination generally, then reliance on raw statistics may seem less appropriate. Put in its starkest form, if one were to assume that 51% of all employers discriminate, then one could "prove" that any given employer "probably" discriminated simply by proving that it was an employer. Such an argument would be analogous to permitting the Internal Revenue Service to impose civil penalties against a taxpayer on a showing that 51 percent of taxpayers underreport their income. There is a large literature on the use of such "naked" statistical evidence—that is, evidence that relies largely or wholly on base-rate evidence. See, e.g., Koehler & Shaviro, supra note 21; David Kaye, The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation, 1982 Am. Bar Found. Res. J. 487.
IV. The Prevalence and Persistence of the Fallacy

The fact that courts sometimes make mistakes with statistical evidence is not by itself of tremendous concern. Our system of civil justice is set up to tolerate a substantial amount of error. The error committed by courts engaging in the transposition fallacy is remarkable, however, for its ubiquity; the vast majority of courts that describe the meaning of a p-value explain it in terms embodying the fallacy.

Leading cases from virtually all the circuits have explicitly committed the transposition fallacy.23 In a characteristic opinion, the Second Circuit observed:

Standard deviation analysis measures the probability that a result is a random deviation from the predicted result—the more standard deviations the lower the probability the result is a random one. Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random and the deviation must be accounted for by some factor other than chance. A finding of two or three standard deviations (one in 384 chance the result is random) is generally highly probative of discriminatory treatment.24

One should not fault these courts too much, for it appears that in many cases they may have been led into their error by expert witnesses. Courts regularly characterize the testimony of statisticians in terms that embody the transposition fallacy.25 Whether the courts are accurately paraphrasing the experts, of course, cannot be determined without resort to the trial transcripts.

There is reason to believe that the courts are not consistently mischaracterizing the experts' testimony, for quite a number of scholarly authorities commit the fallacy in their writings, where they are presumably less likely to be misquoted.26 For example, the leading treatise on employment discrimination law describes a 1% significance level as meaning that there is "no more than one chance in a hundred that the observed disparity occurred by chance."27 Similarly, a treatise on statistical proof of discrimination states:

We start by assuming hypothetically that Employer A was not discriminating. . . . Given this assumption, it is quite likely that Employer A would have hired between 12 and 28 women—some other result would

23. See Browne, supra note 4, footnotes 46-47 (collecting cases).
25. See, e.g., Blum v. Witco Chem. Corp., 829 F.2d 367, 371 (3d Cir. 1987) ("Additionally, plaintiffs produced a statistical expert who testified that the probability that the disparate retention rate was due to some random factor unrelated to age was .0084."); Moore v. McGraw Edison Co., 804 F.2d 1026, 1031 (8th Cir. 1986) (plaintiffs' expert "testified that the chances were two in one thousand that age was not a factor in the terminations"). See also Browne, supra note 4, footnote 48 (collecting cases).
26. See Browne, supra note 4, at 491-95 n.49-51 (collecting examples).
have occurred by chance no more than 1 in 20 times. . . . Employer A hired only 10 women. . . . Although it is possible that Employer A was not discriminating—Employer A may have been that one case out of 20 in which random hiring resulted in hiring a number of women outside of the 12- to 28-range—it seems more likely that our hypothetical assumption was not true and that Employer A was discriminating. . . . In rejecting the hypothetical assumption of no discrimination, one can never be certain that the correct inference is made; a chance of error results from the small probability (one chance out of 20) that employer A was not discriminating when only 10 women were hired. This type of error, rejecting X when, in fact, it is true, is called a Type I error. 28

Lawyers, in addition to judges, seem to come by the transposition fallacy honestly. Casebooks and treatises used by law students routinely instill the fallacy in students of employment discrimination law. 29

Perhaps not surprisingly, the transposition fallacy has led some to connect explicitly what they perceive to be the likelihood of error with the standard of proof. That is, if one believes that the standard five-percent significance level represents the likelihood of error, then one naturally questions why 95% certainty is demanded in a regime that requires proof only by a preponderance of the evidence. After describing the justification for use of the 5% standard in social science research, John Dawson asserts:

---

28. PAETZOLD & WILLBORN, supra note 7, at 2-12 to 2-13 (emphasis added). Interestingly, although in the above quotation Paetzold & Willborn describe the meaning of the p-value in terms of the transposition fallacy, they correctly describe the p-value in other places. They state:

The probability that is explicitly considered in hypothesis testing, the probability of Type I error, is not the probability one would instinctively want in discrimination cases. The Type I error probability (1 out of 20 times, in our example) is the expected proportion of times that the number of women hired would fall outside the range of numbers "close" to 20 in hypothetical repetitions of the same hiring process on the condition that the null hypothesis of no discrimination is true. One would prefer to know the probability that the null hypothesis is true given that we have observed a number of women hired outside the range of numbers "close" to 20. Unfortunately, this model of hypothesis testing cannot produce the preferred kind of probability.

PAETZOLD & WILLBORN, supra, at 2-13 n. 21. Paetzold & Willborn do not take the next step, however, which is to tell us why we should substitute one probability for the other given their lack of equivalence.

29. See, e.g., JOEL W. FRIEDMAN & GEORGE M. STRICKLER, JR., THE LAW OF EMPLOYMENT DISCRIMINATION: CASES AND MATERIALS (4th Ed.) 328 (1997) (stating that "the standard deviation is a way to calculate the likelihood that chance is responsible for the difference between a predicted result and an actual result") (internal quotation marks omitted); ZIMMER ET AL., supra note 22, at 288 (stating that the p-value allows one to decide that "the null hypothesis is incorrect"); MACK A. PLAYER, ELAINE W. SHOEN & RISA L. LIEBERWITZ, EMPLOYMENT DISCRIMINATION LAW: CASES AND MATERIALS (2d Ed.) 236 (1995) (stating that the standard deviation analysis provides a measure of the "odds of error"); MACK A. PLAYER, EMPLOYMENT DISCRIMINATION LAW 344-349 (1987) (stating that if the p-value is .04, "we are 96% sure that the result did not occur by chance"); MACK A. PLAYER, FEDERAL LAW OF EMPLOYMENT DISCRIMINATION (NUTSHELL SERIES) 81-84 (3d Ed. 1992) (stating that the standard deviation "calculates the probabilities that the difference between the expected number and the observed number was a product of chance").
The objective in civil litigation, however, is not to prove a scientific theory, but to decide which of two opposing parties is more likely right. In this context such a standard may be unduly onerous, since a very low risk of false inculpation (as the 5 percent standard is) may be associated with a high risk of false exculpation.\textsuperscript{30}

As we have seen, however, the \( p \)-value is in no way a measure of the likelihood of error,\textsuperscript{31} and therefore establishing a significance level of .05 does not mean the law is demanding a 95\% certainty of discrimination.

Despite its prevalence, numerous commentators have identified the fallacy, no one more clearly than David Kaye. For example, Kaye has explained:

The court's assumption, however, that when the "probability of statistical error is less than 5\%," the "scientific fact is at least 95\% certain" exemplifies a common misunderstanding of the role of statistical tests in scientific inference. . . .

The difficulty is that this interpretation of the result of the hypothesis test is wrong. The test was structured so as to retain the null hypothesis unless the chance of getting the evidence under this hypothesis fell below 5\%. The test focused exclusively on the probability of the evidence given the null hypothesis. Nothing was said about the probability of the hypothesis in the light of the experimental evidence. It may be tempting to call the probability of 0.055 the chance of a coincidence, and to say that the probability of something other than a coincidence—of foul play—must be what is left over, namely 0.945. But this only shows that one can "prove" anything with words.\textsuperscript{32}

Information concerning the proper interpretation of probabilities is not difficult for judges to obtain. The Reference Manual on Scientific Evidence, provided free of charge by the Federal Judicial Center to all members of the federal judiciary, provides as follows:

[S]ince \( p \) is calculated by assuming the null hypothesis, the \( p \)-value cannot give the chance that this hypothesis is true. The \( p \)-value merely gives the chance of getting evidence against the null hypothesis as


\textsuperscript{31} For a more detailed discussion of the lack of a simple relationship between the significance level and the burden of persuasion, see David H. Kaye, Statistical Significance and the Burden of Persuasion, 46 LAW & CONTEMP. PROBS. 13 (Autumn 1983).

\textsuperscript{32} Kaye, supra note 31, at 21-22. See also Paul Meier et al., What Happened in Hazelwood: Statistics, Employment Discrimination and the 80\% Rule, 1984 AM. B. FOUND. RES. J. 139, 148 (pointing out that "it is a serious but very frequent error" to confuse the probability of random selection with the \( p \)-value); Robert Follett & Finis Welch, Testing for Discrimination in Employment Practices, 46 LAW & CONTEMP. PROBS. 171, 174 (Autumn 1983) (stating that "the 0.05 rule does not say that when a difference as large as two standard deviations occurs the probability that the two groups are treated equally is 5\% or less, nor does it say that the probability of unequal treatment is 95\% or more").
strong or stronger than the evidence at hand—assuming the null hypothesis is correct. . . . [T]here is no meaningful way to assign a numerical probability to the null hypothesis, or to any alternative hypothesis, for that matter.\textsuperscript{33}

For the most part, these cautions have been ignored. In a few cases, they have been heard but misunderstood. For example, some commentators have interpreted criticism of commission of the fallacy as being based on an assumption that the 5 percent significance level is too high.\textsuperscript{34} The point, however, is not that a 5% chance of error is too high; indeed, most students of our civil justice system would no doubt believe that an accuracy rate of 95% would be a monumental improvement over the status quo. The point is instead that a significance level of .05 is simply not the same thing as an error rate of five percent. Thus, it is not the particular significance level with which critics of the fallacy take issue; it is the interpretation of the meaning of the significance level generally.

The persistence of the fallacy in discrimination cases is puzzling.\textsuperscript{35} For some reason perhaps having roots deep in human psychology, the fallacy is very easy to fall into and seemingly difficult to extricate oneself from.\textsuperscript{36} While one might be tempted to attribute it to bias in favor of plaintiffs, there is no indication that lawyers and experts for defendants are any less prone to it than those of plaintiffs or that courts finding for plaintiffs are more likely to commit the fallacy than courts finding for defendants.

In many other contexts, the illogic behind the transposition fallacy is obvious. Suppose, for example, that in a given lottery, there is a one-in-a-million chance of winning. In other words, only one out of every million tickets is a valid winner. Asked the prior probability of an honest customer’s purchasing a single valid ticket that turns out to be the winner, most people would unhesitatingly—and correctly respond—“one in a million.” Now suppose that an individual shows up at the

\textsuperscript{33} Kaye & Freedman, supra note 11, at 378-79 (1994) (emphasis added).

\textsuperscript{34} See, e.g., Mark R. Brown, Popularizing Ballot Access: The Front Door to Election Reform, 58 Ohio St. L.J. 1281, 1316 (asserting that “while most sociologists use [the 5% significance level], others would require a greater degree of confidence,” the “others” being Browne, supra note 4, and Kaye, supra note 8). See also Michael J. Zimmer et al., Cases and Materials on Employment Discrimination 245 (3d ed. 1994) (stating that “Kingsley R. Browne asserts that even [a five percent chance of error] is too much for the courts to risk by using hypothesis testing in employment discrimination cases,” citing Browne, supra).

\textsuperscript{35} As an example of the persistence of the fallacy, one source explicitly acknowledges that my description of the fallacy is “theoretically correct,” Zimmer et al., supra note 22, at 284, but then goes on, when describing the meaning of statistically significant results, to engage in the fallacy repeatedly, id. 281-93.

\textsuperscript{36} When psychologists describe the processing of probabilistic information as “counterintuitive,” what they mean is that the human brain is not well-designed to process that kind of information. See Stephen E. Edgell et al., Base Rates, Experience, and the Big Picture, 19 Behav. & Brain Sci. 21 (1996).
lottery office with an apparently valid ticket bearing the winning number. What is the likelihood that he is a valid winner, as opposed to, say, a trafficker in counterfeit lottery tickets? If one responds "one in a million," one commits the transposition fallacy.

The fallacy is far less seductive in the lottery example than it has proved to be in the discrimination context. Most people would correctly recognize that assessment of the likelihood that the ticket is valid should turn on such facts as the general prevalence of counterfeit lottery tickets, the difficulty of counterfeiting tickets, and indicia of honesty of the person presenting the ticket—for example, whether the person is a bishop or a professional swindler. They would likely recognize that these facts should be considered in reaching a subjective judgment about the likelihood of the ticket's validity, however, rather than viewing them as a means to compute a precise probability.

By comparison to discrimination cases, some courts have been relatively quick to recognize the transposition fallacy in the criminal context. In the famous case of People v. Collins, the California Supreme Court reversed criminal convictions that were based upon the logic of the transposition fallacy. Witnesses had seen a blond woman with a ponytail and a black man with a beard and mustache drive away in a yellow car immediately following a robbery, and the defendants met that description. The prosecution's expert testified that the probability that a randomly chosen couple would match this description was one in twelve million. The California Supreme Court found several flaws in the prosecution's statistical evidence, but the one that is relevant to this discussion is that the court recognized that the probability that a random couple would match the description is entirely different from the probability that a couple matching the description is innocent. As the court noted, the prosecutor gave "absolutely no guidance on the crucial issue: of the admittedly few such couples, which one, if any, was guilty of committing this robbery?" Numerous scholars have commented, virtually all favorably, on the Collins court's statistical analysis.

This is not to suggest that criminal courts do not commit the transposition fallacy, for they do, probably most commonly in their interpretation of DNA evidence. In fact, in the criminal literature, the fallacy goes by the name of the "prosecutor's fallacy" because of the recognition that it gives the prosecutor an undue advantage. In such cases, the

37. 68 Cal.2d 319 (1968).
38. Id. at 329.
39. Id. at 330.
41. See Thompson & Schumann, supra note 5.
fallacy is committed when courts interpret testimony that the probability that a sample of DNA would match a randomly selected person is, say, 1 in 10,000, to mean that there is only a 1 in 10,000 chance that when there is a match that it was a coincidental one and therefore a 9,999 in 10,000 chance that the sample came from the suspect. Perhaps because there is greater concern about erroneous results in favor of criminal prosecutors than there is about erroneous results in favor of discrimination plaintiffs, the fallacy is much more widely discussed in the criminal context.

V. The Consequences of the Fallacy: Why It Matters

Readers who have come this far should now see clearly the important effect that the transposition fallacy has on the reasoning of courts in discrimination cases. Courts' misunderstanding of the meaning of the p-value has substantial impact on what they think that plaintiffs have proved. When they reason that there is only a one-in-20 chance that a disparity was randomly caused, they conclude that they are seeing a rare result and that chance has been ruled out as the cause. For example, the Fifth Circuit has characterized a finding of statistical significance as follows: "A variation of two standard deviations would indicate that the probability of the observed outcome occurring purely by chance would be approximately five out of 100; that is, it could be said with a 95% certainty that the outcome was not merely a fluke." Similarly, the Fourth Circuit has stated that disparities of more than two or three standard deviations "conclusively ruled out chance as the cause of the disparity in the termination rates." Many commentators also explicitly link the p-value with the likelihood of making a mistake, as in the statement, "When led to a rejection of the null hypothesis at a level of significance of 0.05, a court can be at least 95% confident that a disparity

42. The error of such reasoning is obvious. In a city of 1,000,000 people, there will be approximately 100 who would match the sample. Based solely on the DNA evidence, then, one might well argue that the chance that the sample came from the suspect—rather than being 9,999 out of 10,000—is 1 out of 100.


45. Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982).

of treatment of the relevant groups exists.\textsuperscript{47} Some courts have equated statistical significance even more directly with discrimination.\textsuperscript{48}

The transposition fallacy creates an erroneous mindset that leads courts to act with a far greater sense of certainty than is warranted. This feeling of certainty leads them to conclude that statistically significant disparities raise a strong inference of discrimination that requires a judgment in favor of the plaintiff unless the employer can rebut it.\textsuperscript{49} Those who encourage courts to commit the fallacy are asking the court to reason, in effect:

The plaintiff has proved that there is only a 5-percent chance that the employer's workforce imbalance was produced by chance; therefore, we can rule out chance as the cause. The employer's workforce distribution is suspicious on its face, and the employer should bear the burden of identifying the nonrandom nondiscriminatory cause of the disparity, if there is one. In the absence of such proof, I am entitled to infer that the nonrandom cause is discrimination.

What the court should reason is very different:

The plaintiff has introduced statistics that would be true for thousands of nondiscriminating employers. I cannot determine from the statistics alone whether chance or some systematic cause is responsible, let alone whether the systematic cause is discrimination. Therefore, if the plaintiff wants me to conclude that discrimination is responsible, he had better give me a lot more than what he has given me so far.

The aura of accuracy, precision, and objectivity of statistical evidence creates a great "temptation to accord disproportionate weight" to the evidence.\textsuperscript{50} An expert statistician's testimony that there is only a 2.44\% chance that the employer's workforce disparity was created by chance is powerful evidence, but, as this article attempts to demonstrate, it is evidence that should be ignored.

Now, one might respond that even if a $p$-value of .05 does not correspond to a five-percent probability of a nonrandom cause, and therefore is not conclusive on the question of discrimination, the results are sufficiently rare that they are at least strongly probative of discrimination. There are two flaws in that argument.

First, statistically significant disparities caused by chance are in fact widespread rather than rare. One out of 20 statistical comparisons of nondiscriminating employers will be statistically significant at the

\textsuperscript{47} Braun, supra note 30, at 87.
\textsuperscript{48} See, e.g., Capac v. Katz & Besthoff, Inc., 711 F.2d 647, 651-52 (5th Cir. 1983) (characterizing the $p$-value as the "probability of unbiased hiring"), cert. denied, 486 U.S. 927 (1984); Hameed v. Iron Workers Local 396, 637 F.2d 506, 512-13 (8th Cir. 1980) (stating that "[i]f tests of statistical significance eliminate chance as a likely explanation for the differential pass rates, courts will presume that the disparate pass rates are attributable to racially discriminatory selection criteria").
\textsuperscript{49} See, e.g., Palmer v. Shultz, 815 F.2d 84, 108 (D.C. Cir. 1987).
level. Not only does this mean that one out of 20 employers will have statistically significant disparities that are caused by chance, but it also means that all employers of substantial size—with multiple departments, job classifications, and so forth—are likely to have numerous statistically significant disparities within their workforces.  

Second, rarity does not logically imply an improper cause or any particular kind of cause at all, for that matter. Although courts sometimes reason that the rarity of a result supports an inference of an illegitimate cause, their reasoning is logically flawed. Some very common consequences result from improper causes, and some very uncommon consequences result from entirely benign causes.

One might also argue that it does not really matter that courts engage in fallacious reasoning in interpreting the plaintiff’s prima facie case because the defendant has an opportunity to rebut the incorrect factual inference drawn from the plaintiff’s statistics. There are at least two responses to that argument, one procedural and the other substantive. First, imposing on employers the obligation to disprove discrimination merely because the plaintiff has introduced evidence that is not very strongly suggestive of discrimination is inappropriate in a regime that places the burden of proof on the plaintiff. Second, how is the employer to prove that chance is the cause of the disparity after the transposition fallacy has led the trier of fact to rule out chance as the cause? One can never prove the null hypothesis in any event, but only the slickest lawyer could exculpate an employer by proving the null hypothesis after it has already been rejected.

The fact that a disparity that is barely significant at the 5% level is not by itself strongly suggestive of an illegal cause or even a nonrandom one does not mean that statistical evidence is irrelevant. As the disparities become larger, the likelihood of a nonrandom cause becomes greater. For example, in EEOC v. Sears, Roebuck & Co.—a case involving allegations of sex discrimination in hiring for commission sales jobs—the z-scores for nationwide comparisons ranged from 11.9 to 45.1, with infinitesimal associated probabilities. Similarly, in Capaci v. Katz &

---

51. The concept of “rarity” does not have a great deal of independent meaning divorced from context. For example, it would not be said of someone who every two weeks gets a speeding ticket during his daily commute that he “rarely” gets speeding tickets. On the other hand, one would probably say that someone who is late to work 19 out of 20 times is “rarely” on time.

52. See, e.g., McGonigal v. Gearhart Indus., Inc., 788 F.2d 321, 327 (5th Cir. 1986) (stating that “[i]f a man with thirty years experience in the munitions field had not previously heard of a grenade with a missing delay column, then it also seems reasonable to conclude that such an error will occur only when the inspector is negligent.”).

53. Zimmer et al., supra note 22, at 284 (arguing that criticism of courts for committing the transposition fallacy does not take into account that the defendant may rebut the showing with “[s]ufficiently strong testimony” that chance was responsible).

54. See Texas Dept. of Community Affairs v. Burdine, 450 U.S. 248, 253 (1981) (stating that “[t]he ultimate burden of persuading the trier of fact that the defendant intentionally discriminated against the plaintiff remains at all times with the plaintiff”).

55. 839 F.2d 302, 323 n.20 (7th Cir. 1988).
Besthoff—a case involving allegations of sex discrimination in entry into management training positions—the court noted that “the highest probability of unbiased hiring was $5.367 \times 10^{-30}$, less than one in a billion billions.” In both of these cases, it would be wrong to conclude that the $p$-value provided a measure of the probability of random hiring, but nonetheless one could have a fairly high degree of subjective confidence that some nonrandom cause, whether discriminatory or not, was responsible. Nonstatistical evidence is necessary in such cases to guide the ultimate conclusion whether the nonrandom cause was a legal or illegal one.

Evidence of statistical disparities necessarily plays an important role in pattern-or-practice cases. In such cases, the plaintiff’s burden is to demonstrate that discrimination is the employer’s standard operating procedure, a showing that would be difficult, although not logically impossible, to make in the absence of a sizable disparity between the employer’s workforce and the statistical expectation. However, the limitations on statistical proof presented here suggest that much statistical evidence should be given less weight than it has been given to date.57

VI. Conclusion

I have argued here that courts have routinely misunderstood, and overestimated, the meaning of statistically significant disparities. Courts interpret the $p$-value to be a measure of the likelihood that an employer’s workforce was obtained by chance, when in fact what the $p$-value represents is the probability that a randomly hiring employer will obtain such a distribution by chance. Although these phrases may sound much alike, there is no relationship between the two probabilities that would allow courts to infer one from the other.

To say that courts have misunderstood the $p$-value does not mean, however, that statistical evidence is wholly without value. It is still true that the greater the disparity between an employer’s workforce and the statistical expectation the more likely it will usually be that discrimination or some other nonrandom cause is responsible.58 What this article establishes is that the probability cannot be quantified in the way that it routinely has been in the past.

Recognition of the transposition fallacy may diminish the appeal of statistical evidence for some. One of the major attractions of such evidence is that it has been thought to provide a quantitative measure of probabilities to replace our intuition. The message here is that we can no longer defer to probability tables in deciding whether we think an employer has engaged in illegal discrimination and must instead, for better or worse, rely on more complex subjective human judgments.

56. 711 F.2d 647, 652 (5th Cir. 1983).
57. See Browne, supra note 4, at 541-549 (arguing that strong anecdotal evidence of discrimination should be required in pattern-or-practice cases).