EXHIBIT G

OMNIBUS BROWN DECLARATION

Reference Manual on Scientific Evidence

Third Edition

Committee on the Development of the Third Edition of the
Reference Manual on Scientific Evidence

Committee on Science, Technology, and Law
Policy and Global Affairs

FEDERAL JUDICIAL CENTER

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

The Federal Judicial Center contributed to this publication in furtherance of the Center's statutory mission to develop and conduct educational programs for judicial branch employees. The views expressed are those of the authors and not necessarily those of the Federal Judicial Center.

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The development of the third edition of the *Reference Manual on Scientific Evidence* was supported by Contract No. B5727.R02 between the National Academy of Sciences and the Carnegie Corporation of New York and a grant from the Starr Foundation. The views expressed in this publication are those of the authors and do not necessarily reflect those of the National Academies or the organizations that provided support for the project.

International Standard Book Number-13: 978-0-309-21421-6
International Standard Book Number-10: 0-309-21421-1

Library of Congress Cataloging-in-Publication Data

Reference manual on scientific evidence. — 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-309-21421-6 (pbk.)
ISBN-10: 0-309-21421-1 (pbk.)
1. Evidence, Expert—United States. I. Federal Judicial Center.
KF8961.R44 2011
347.73′67—dc23

2011031458

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

The *p*-value is the probability of getting data as extreme as, or more extreme than, the actual data—given that the null hypothesis is true. In the example, *p* turns out to be essentially zero. The discrepancy between the observed and the expected is far too large to explain by random chance. Indeed, even if the panel had included 155 women, the *p*-value would only be around 0.02, or 2%.[98] (If the population is more than 50% female, *p* will be even smaller.) In short, the jury panel was nothing like a random sample from the community.

Large *p*-values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance variation. On the other hand, if *p* is very small, something other than chance must be involved: The data are far away from the values expected under the null hypothesis. Significance testing often seems to involve multiple negatives. This is because a statistical test is an argument by contradiction.

With the Dr. Spock example, the null hypothesis asserts that the jury panel is like a random sample from a population that is 50% female. The data contradict this null hypothesis because the disparity between what is observed and what is expected (according to the null) is too large to be explained as the product of random chance. In a typical jury discrimination case, small *p*-values help a defendant appealing a conviction by showing that the jury panel is not like a random sample from the relevant population; large *p*-values hurt. In the usual employment context, small *p*-values help plaintiffs who complain of discrimination—for example, by showing that a disparity in promotion rates is too large to be explained by chance; conversely, large *p*-values would be consistent with the defense argument that the disparity is just due to chance.

Because *p* is calculated by assuming that the null hypothesis is correct, *p* does not give the chance that the null is true. The *p*-value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the *p*-value can therefore be summarized in two lines:

> *p* is the probability of extreme data given the null hypothesis.
> *p* is not the probability of the null hypothesis given extreme data.[99]

[98] With 102 women out of 350, the *p*-value is about $2/10^{15}$, where 10^{15} is 1 followed by 15 zeros, that is, a quadrillion. See *infra* Appendix for the calculations.

[99] Some opinions present a contrary view. *E.g.*, Vasquez v. Hillery, 474 U.S. 254, 259 n.3 (1986) ("the District Court . . . ultimately accepted . . . a probability of 2 in 1000 that the phenomenon was attributable to chance"); Nat'l Abortion Fed. v. Ashcroft, 330 F. Supp. 2d 436 (S.D.N.Y. 2004), *aff'd in part*, 437 F.3d 278 (2d Cir. 2006), *vacated*, 224 Fed. App'x. 88 (2d Cir. 2007) ("According to Dr. Howell, . . . a 'P value' of 0.30 . . . indicates that there is a thirty percent probability that the results of the . . . [s]tudy were merely due to chance alone."). Such statements confuse the probability of the

To recapitulate the logic of significance testing: If p is small, the observed data are far from what is expected under the null hypothesis—too far to be readily explained by the operations of chance. That discredits the null hypothesis.

Computing p-values requires statistical expertise. Many methods are available, but only some will fit the occasion. Sometimes standard errors will be part of the analysis; other times they will not be. Sometimes a difference of two standard errors will imply a p-value of about 5%; other times it will not. In general, the p-value depends on the model, the size of the sample, and the sample statistics.

2. *Is a difference statistically significant?*

If an observed difference is in the middle of the distribution that would be expected under the null hypothesis, there is no surprise. The sample data are of the type that often would be seen when the null hypothesis is true. The difference is not significant, as statisticians say, and the null hypothesis cannot be rejected. On the other hand, if the sample difference is far from the expected value—according to the null hypothesis—then the sample is unusual. The difference is significant, and the null hypothesis is rejected. Statistical significance is determined by comparing p to a preset value, called the significance level.[100] The null hypothesis is rejected when p falls below this level.

In practice, statistical analysts typically use levels of 5% and 1%.[101] The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure. An unexplained reference to highly significant results probably means that p is less

kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome—the "transposition fallacy."

Instances of the transposition fallacy in criminal cases are collected in David H. Kaye et al., The New Wigmore: A Treatise on Evidence: Expert Evidence §§ 12.8.2(b) & 14.1.2 (2d ed. 2011). In *McDaniel v. Brown*, 130 S. Ct. 665 (2010), for example, a DNA analyst suggested that a random match probability of 1/3,000,000 implied a .000033 probability that the DNA was not the source of the DNA found on the victim's clothing. *See* David H. Kaye, *"False But Highly Persuasive": How Wrong Were the Probability Estimates in* McDaniel v. Brown? 108 Mich. L. Rev. First Impressions 1 (2009).

[100] Statisticians use the Greek letter alpha (α) to denote the significance level; α gives the chance of getting a significant result, assuming that the null hypothesis is true. Thus, α represents the chance of a false rejection of the null hypothesis (also called a false positive, a false alarm, or a Type I error). For example, suppose $\alpha = 5\%$. If investigators do many studies, and the null hypothesis happens to be true in each case, then about 5% of the time they would obtain significant results—and falsely reject the null hypothesis.

[101] The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as "suspect to a social scientist" when a statistic from "large samples" falls more than "two or three standard deviations" from its expected value under the null hypothesis. Although the Court did not say so, these differences produce p-values of about 5% and 0.3% when the statistic is normally distributed. The Court's standard deviation is our standard error.

information. Indeed, when confidence intervals and *p*-values can be computed, the interpretation is the same with small samples as with large ones.[108] The concern with small samples is not that they are beyond the ken of statistical theory, but that

1 The underlying assumptions are hard to validate.
2. Because approximations based on the normal curve generally cannot be used, confidence intervals may be difficult to compute for parameters of interest. Likewise, *p*-values may be difficult to compute for hypotheses of interest.[109]
3. Small samples may be unreliable, with large standard errors, broad confidence intervals, and tests having low power.

3. *One tail or two?*

In many cases, a statistical test can be done either one-tailed or two-tailed; the second method often produces a *p*-value twice as big as the first method. The methods are easily explained with a hypothetical example. Suppose we toss a coin 1000 times and get 532 heads. The null hypothesis to be tested asserts that the coin is fair. If the null is correct, the chance of getting 532 or more heads is 2.3%. That is a one-tailed test, whose *p*-value is 2.3%. To make a two-tailed test, the statistician computes the chance of getting 532 or more heads—or $500 - 32 = 468$ heads or fewer. This is 4.6%. In other words, the two-tailed *p*-value is 4.6%. Because small *p*-values are evidence against the null hypothesis, the one-tailed test seems to produce stronger evidence than its two-tailed counterpart. However, the advantage is largely illusory, as the example suggests. (The two-tailed test may seem artificial, but it offers some protection against possible artifacts resulting from multiple testing—the topic of the next section.)

Some courts and commentators have argued for one or the other type of test, but a rigid rule is not required if significance levels are used as guidelines rather than as mechanical rules for statistical proof.[110] One-tailed tests often make it

[108] Advocates sometimes contend that samples are "too small to allow for meaningful statistical analysis," United States v. New York City Bd. of Educ., 487 F. Supp. 2d 220, 229 (E.D.N.Y. 2007), and courts often look to the size of samples from earlier cases to determine whether the sample data before them are admissible or convincing. *Id.* at 230; Timmerman v. U.S. Bank, 483 F.3d 1106, 1116 n.4 (10th Cir. 2007). However, a meaningful statistical analysis yielding a significant result can be based on a small sample, and reliability does not depend on sample size alone (*see supra* Section IV.A.3, *infra* Section V.C.1). Well-known small-sample techniques include the sign test and Fisher's exact test. *E.g.*, Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers 154–56, 339–41 (2d ed. 2001); *see generally* E.L. Lehmann & H.J.M. d'Abrera, Nonparametrics (2d ed. 2006).

[109] With large samples, approximate inferences (e.g., based on the central limit theorem, *see infra* Appendix) may be quite adequate. These approximations will not be satisfactory for small samples.

[110] *See, e.g.*, United States v. State of Delaware, 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 WL 609331, *10 n.4 (D. Del. 2004). According to formal statistical theory, the choice between one

easier to reach a threshold such as 5%, at least in terms of appearance. However, if we recognize that 5% is not a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the *p*-value are made explicit.

4. *How many tests have been done?*

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield "significant" findings, even when there is no real effect. To illustrate the point, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce 10 heads when tossed 10 times is $(1/2)^{10} = 1/1024$. Observing 10 heads in the first 10 tosses, therefore, would be strong evidence that the coin is biased. Nonetheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. Ten heads in the first ten tosses means one thing; a run of ten heads somewhere along the way to a few thousand tosses of a coin means quite another. A test—looking for a run of ten heads—can be repeated too often.

Artifacts from multiple testing are commonplace. Because research that fails to uncover significance often is not published, reviews of the literature may produce an unduly large number of studies finding statistical significance.[111] Even a single researcher may examine so many different relationships that a few will achieve statistical significance by mere happenstance. Almost any large dataset—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by diligent search. Having detected the pattern, the analyst can perform a statistical test for it, blandly ignoring the search effort. Statistical significance is bound to follow.

There are statistical methods for dealing with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.[112] However, no general solution is available, and the existing methods would be of little help in the typical case where analysts have tested and rejected a variety of models before arriving at the one considered the most satisfactory (*see infra* Section V on regression models). In these situations, courts should not be overly impressed with

tail or two can sometimes be made by considering the exact form of the alternative hypothesis (*infra* Section IV.C.5). *But see* Freedman et al., *supra* note 12, at 547–50. One-tailed tests at the 5% level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. One-tailed tests are also called one-sided (with no pejorative intent); two-tailed tests are two-sided.

[111] *E.g.*, Philippa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 Lancet 867 (1991); John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 JAMA 281 (1998); Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 New Eng. J. Med. 426 (1987).

[112] *See, e.g.*, Sandrine Dudoit & Mark J. van der Laan, Multiple Testing Procedures with Applications to Genomics (2008).