

# CAUSAL AND ASSOCIATIVE HYPOTHESES IN PSYCHOLOGY

## Examples From Eyewitness Testimony Research

Daniel B. Wright  
University of Sussex

Two types of hypotheses interest psychologists: causal hypotheses and associative hypotheses. The conclusions that can be reached from studies examining these hypotheses and the methods that should be used to investigate them differ. Causal hypotheses examine how a manipulation affects future events, whereas associative hypotheses examine how often certain events co-occur. In general, experimental methods with random allocation are well suited for addressing causal hypotheses, whereas random sampling is an asset when examining associative hypotheses. These hypotheses are discussed primarily with reference to 4 topics within eyewitness testimony research: the own-race bias, emotion and memory, event duration estimation, and system variables in lineups. Some other examples in forensic psychology are provided to illustrate difference between causal and associative hypotheses.

*Keywords:* eyewitness testimony, own-race bias, emotion, causation, association

Imagine that you, as an expert witness on the psychology of memory, are asked the following question: Does the fact that an event is highly emotional affect the accuracy of an eyewitness (Q1)? Your answer might refer to the weapon focus effect (Loftus, Loftus, & Messo, 1987) and dozens of carefully controlled laboratory studies in which researchers attempted to create events identical except for their degree of emotionality. If your answer is based on a recent meta-analysis of many of these laboratory studies (Deffenbacher, Bornstein, Penrod, & McGorty, 2004), you will report that moderate levels of emotionality, those that can be studied in the laboratory, tend to negatively affect eyewitness accuracy. As a cautious expert witness, you might stress that your comments apply only to the types of events that are representative of those included in the meta-analysis.

Suppose instead that you are asked this question: Do emotional events tend to be remembered more accurately than nonemotional events (Q2)? To address this question, you would need to compare samples of emotional and nonemotional events or examine a sample of events and determine if emotionality is associated with accuracy. If you are interested in all events, you might cite Brewer's (1988) study in which people had to record an event every time a random beeper was activated. If you are interested only in more personally significant events, you might cite a study by Burt, Mitchell, Raggatt, Jones, and Cowan (1995) of memory for holiday photographs. For this sample of events, those that are emotional tend to be more memorable than nonemotional events, which also tend to be more mundane and less personally significant. If you are interested in

---

This research was funded by the British Academy.

Correspondence concerning this article should be addressed to Daniel B. Wright, Psychology Department, University of Sussex, Brighton BN1 9QH, United Kingdom. E-mail: danw@sussex.ac.uk

eyewitness memory, you might look at one of the archival studies of lineups and operationalize emotion according to whether a weapon was present. Using this approach, Valentine, Pickering, and Darling (2003) found weapon presence was not associated with suspect identifications. As a cautious expert, you would stress that your answer depends on the sample of events.

The lawyer asking these two questions (Q1 and Q2) may not realize the importance of the distinction between them; therefore, it is important to seek clarification about which question you are answering. The two types of questions imply distinctly different hypotheses and fundamentally different approaches in psychology as a science. The aims of this article are to describe the different forms that hypotheses can take in psychology, to explore how empirical approaches match these forms, and to describe how these different forms apply in legal settings. Although this article refers most extensively to research on eyewitness testimony, the approach holds for other topics in psychology and science more generally, examples of which are discussed throughout this article. Eyewitness testimony research has been chosen as a focus because lawyers, judges, and politicians are often interested in both causal and associative hypotheses about eyewitness memories. For example, a lawyer might be interested in the first question (Q1) because two witnesses viewed the same event but one had a gun pointed at him or her, whereas the other did not. Alternatively, to inform the jury about the probability of an accurate identification, a lawyer might pursue the second question (Q2) to establish whether a witness to a violent crime is more likely to be accurate compared with a witness to a nonviolent crime.

### Hypotheses in Psychology

William James (1890), in *The Principles of Psychology*, defined psychology as “the science of mental life, both of its phenomena and of their conditions” (p. 1). There are two central tenets for what constitutes a science. The first is that systematic measurement takes place (Hand, 2004). According to Galton (1879), “until the phenomena of any branch of knowledge have been subjected to measurement and number, it [the branch of knowledge] cannot assume the status and dignity of a science” (p. 149). The second central tenet is that the research must inform and be informed by a set of theories or hypotheses. These hypotheses are important in evaluating different approaches to psychology. There are two main empirical approaches to psychology and two different forms of hypotheses that can be examined. Cronbach (1957) described how these approaches, the experimental and correlational psychologies, have existed relatively independently. The approaches address different types of hypotheses: causal and associative hypotheses, respectively. Although the two can be examined together (Cronbach, 1975), their differences must also be acknowledged.

Some simple notation is helpful in describing the different types of hypotheses. Let  $X$  be some event at Time 1 and  $Y$  be some other event at Time 2. Hypotheses are about relationships between  $X$  and  $Y$  but vary in many ways. Some hypotheses are *absolute* in that they describe nature with certainty. The role of time in hypotheses is important, and there are three different ways that  $X$  and  $Y$  can be temporally related for absolute hypotheses. The first way is where  $X$  and  $Y$  always co-occur;  $X$  and  $Y$  can be thought of as the same extended event. An

example drawn from physics is Newton's third law (for every action, there is a reaction), which is properly described without causal terms: Action does not cause reaction, "action equals reaction" (Feynman, Leighton, & Sands, 1963, p. 10-2). The second way is where X always precedes Y, in which case Y can be thought of as a sufficient condition of X and X a necessary condition for Y. For example, if a diagnostic requirement of posttraumatic stress disorder (PTSD) is experiencing a traumatic event, then experiencing a traumatic event should not be viewed as causal of PTSD; rather, it is part of the definition (Grove & Barden, 1999). People who have PTSD can simply be viewed as a subset of those experiencing trauma. The third way is where Y always follows from X. All three of these (X if and only if Y, Y implies X, and X implies Y) are absolute in two ways. First, they are applicable for all situations: "Action equals reaction" is applicable in any situation. Second, the relationships can be described with certainty: Action always equals reaction. However, hypotheses in science are seldom absolute in both of these ways. In the remainder of this section, three different types of hypotheses are distinguished. The first is a version of the absolute hypotheses described above but is true with only a certain probability (i.e., a stochastic relationship). The second two, causal and associative, are the main hypotheses examined in psychology.

Some hypotheses are not defined with respect to specific situations, but neither are they fixed or certain. These involve adding a probability function to an absolute hypothesis. The probability function is the same across all situations. For example, a physicist might put forward the hypothesis that an atom of some radioactive substance at Time 1 has a 50% probability of decaying by Time 2 in any situation. This hypothesis assumes that "God plays with dice." The probability function should be simple, like the roll of a die. As noted above, there are three ways for events to be temporally related for absolute hypotheses ( $X \leftrightarrow Y$ ,  $X \rightarrow Y$ , and  $X \leftarrow Y$ ), to each of which probability functions can be applied. Let  $p$  mean a probability function. Then,  $X \rightarrow Y p$  can be read as X leads to Y with a probability of  $p$ . These statements are not causal. Outside of quantum physics, there are few hypotheses of this type in practice. It is important to introduce them for two reasons. First, it is necessary to introduce a random probability term for developing a taxonomy of hypotheses. Second, these types of hypotheses are implicitly assumed in many statistical tests. For example, in most regression analyses, it is assumed that the model is accurately specified and that all that is left is random error (Berry, 1993).

In psychology, hypotheses are more often about specific causes. In fact, Fodor (1991) stated that psychologists should be concerned only with what are called *ceteris paribus* (CP) hypotheses, sometimes called "all other things being equal" hypotheses, which allow causal attributions. CP hypotheses can occur for  $X \leftrightarrow Y$ ,  $X \rightarrow Y$ , and  $X \leftarrow Y$  hypotheses, but most often, they are  $X \rightarrow Y$  CP because of the relationship between time and cause in all psychological theories.  $X \rightarrow Y$  CP means X leads to Y, but it requires what is called a *completer* situation to be true. An important aspect of CP hypotheses is that they are described in relation to a set of completers (i.e., the situations under which the hypotheses hold). Requiring a completer is important for distinguishing CP from other hypotheses and is also critical for how they are used in science. One problem often described in relation to CP hypotheses is that if care is not taken in constraining what in practice can

be completers, the hypotheses can be circular. If the set of completers is defined as those that satisfy the hypothesis, then the hypothesis is a tautology, and although it may be an accurate description (in essence, this is the anthropic principle; see <http://www.anthropic-principle.com/>), it is of little predictive power. A forensic example is if therapists use lack of a memory of sexual abuse to confirm suspicions that their clients have been sexually abused (Ofshe & Watters, 1995, p. 89). They would be implying that the repression hypothesis is supported because of the lack of a memory.

CP hypotheses can also have probability functions applied to them. For example, wanting chocolate could lead to having chocolate for some set of completers (such as having money), but this still occurs only some random 50% of the time. The important aspect of this is that 50% of the time is not predicted by other variables. If the shop being open is necessary, then it would be an aspect of the completers. Research into lineups provides a detailed example of this later in the article.

The final type of hypothesis to consider is associative hypotheses. For some population of situations, for example, when I am in my office and the shop is open, there is some association between when I want chocolate and when I have chocolate. The interest is in estimating the probability, not the causal mechanisms. It is not necessary to assume that the probability is random; it can be a function of other variables. Whether this type of query is a scientific hypothesis is of some dispute because many believe that science is only about constructing theories of causal relationships. According to Fodor and Pylyshyn (1981), an associative hypothesis is scientifically inadequate: "the goal of psychological theory construction is not to predict most (or even all) of the variance; it is to explicate the underlying mechanisms upon whose operation the variance depends" (p. 154). One objective of this article is to encourage readers to question whether such (atheoretical) hypotheses should or should not be included in scientific psychology.

Within forensic psychology, associative hypotheses are deemed important in risk assessment, such as the risk of offending or reoffending (Rosenfeld & Lewis, 2005). In what is called the actuarial, or statistical, approach, researchers try to identify the set of variables that best predicts offending by using a statistical algorithm. The concern is not about causal relationships but simply about associations, trying to predict most of the variance. Not surprisingly, because this is what the statistical procedures are designed to optimize, actuarial approaches provide more accurate predictions than clinical judgments. What is surprising is how poor some clinical judgments are even when compared with simple algorithms (Grove & Meehl, 1996) and the fact that some people continue to make clinical judgments and others continue to accept them when more accurate predictive methods are available (Swets, Dawes, & Monahan, 2000).

As with the other types of hypotheses, associative hypotheses can be bidirectional (the correlation or odds ratio between X and Y), forward (the conditional probability of Y given X), and backward (the conditional probability of X given Y). Associative hypotheses tend to be more data driven and descriptive than other types of hypotheses. Their lack of causal terms can make them atheoretical. Probability terms can be added to associative hypotheses in the same way as with other hypotheses. Although this may seem counterintuitive, in fact it is a fairly

common assumption in many statistical procedures (i.e., that the residuals have a specific distribution and that they are independent and identically distributed). For example, suppose that there are some purely chance aspects of having chocolate that are not associated with any situational differences. In these cases, a probability function could be included. It would allow random and nonrandom error to be separated.

Associative hypotheses can also be coupled with CP conditionals. The CP aspects define a subset of the populations over which the hypotheses hold (or in which they hold with some probability). For example, a researcher may be interested both in the causal mechanisms for a set of completer situations and in how these mechanisms operate for different situations and individuals.

### Hypotheses and Science

In the previous section, different types of hypotheses have been classified. Each of these expressed a relationship between two events separated in time. The classification of these hypotheses helps one to understand how different studies can be used to evaluate different hypotheses. This section discusses the relationships among the different forms of hypothesis and several critical aspects of science: causation and association, effect size, generalization, and the progress of science. At the end of this section, the relationships between causal and associative hypotheses and the *Daubert* criteria (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993) are reviewed.

#### *Causation and Association*

Causal and associative inferences can be drawn from different types of hypotheses. Causal hypotheses have CP conditionals. Associative hypotheses include probability terms that may be functions of other variables. Thus, a hypothesis can be causal, associative, both, or neither. It is worth stressing that absolute hypotheses are not causal; action is reaction. Applying a probability term does not make the hypothesis causal. If a scientist discovers that 100 g of plutonium at Time 1 will have a mass of about 50 g in 24,000 years, the scientist is saying that 100 g of plutonium at Time 1 is a probability distribution centered on 50 g at a time 24,000 years later (unless it is observed beforehand). The statement is important scientifically, but it is not about cause.

Many philosophers have grappled with the meaning of causality (Cook & Campbell, 1979, chapter 1). Here, the focus is only on Rubin's model of causality (Rubin, 1974; Holland, 1986) because it is particularly well suited to psychology (Wilkinson & Task Force on Statistical Inference, 1999) and provides an account of causal hypotheses that can be adapted for associative hypotheses. To understand Rubin's model, consider a simple two-group experiment where one group (S1) is manipulated in some way ( $X = \text{treatment}$ ) and the other group (S2) is not ( $X = \text{control}$ ). At some later point, the groups are measured on some variable; call it  $Y = Y1$  for the first group and  $Y = Y2$  for the second group. The hypothesis is  $X \rightarrow Y$  CP. The usual approach is that if the difference between  $Y1$  and  $Y2$  is large enough, then one attributes causality to the treatment (see Figure 1). This requires several assumptions. The two most relevant here have to do with the CP conditional. The assumptions are that S1 and S2 are the same and that the

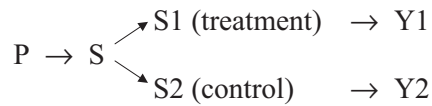


Figure 1. A schematic for Rubin's (1974) model of causation. P = population; S = sample; Y = response.

distinguishing feature between the treatment and the control conditions is what one wants to attribute causality to.

It is not possible to detect or to measure a cause directly. Instead, researchers measure the effect of a cause. Ideally, this would be by measuring Y for every person both with and without the treatment in identical situations. This would mean S1 and S2 were the same and would satisfy the CP conditional. Although this can be done in computer simulations, it cannot be done in the behavioral sciences; it is necessary, therefore, to make assumptions about how groups would have performed in the treatment conditions to which they were not assigned. To make assumptions about the two unobserved measurements, it is necessary to know something about the differences between S1 and S2. The two unobserved measurements are how the participants in the control group would have performed if they had been given the treatment and how the participants in the treatment group would have performed if they had not been given the treatment. S1 and S2 are part of a larger sample, which is a subset of a population. For causal hypotheses, the critical aspect is how the sample is allocated into S1 and S2. As Cook and Campbell (1979) put it, "random assignment is the great *ceteris paribus*—that is, other things being equal—of causal inference" (p. 5). This does not mean that S1 and S2 are the same, but as sampling theory predicts, the likelihood is that they will not be very different from each other. Because of random assignment, a researcher can treat the people in the two groups as roughly equivalent. This means that sampling error has to be taken into account in any empirical modeling of experiments. Statistical techniques are used to determine how large a difference between Y1 and Y2 could be expected if the treatment had no effect. The standard approach is if the observed difference exceeds this amount, then the treatment is assumed to have caused an (not the) effect.

Although random allocation is not necessary to make causal attributions, it makes causal attribution easier (Banaji & Crowder, 1989). The adage "correlation does not imply causation" is more properly written "X being correlated with Y does not imply that X causes Y," but the correlation does strongly suggest that somewhere in the vast network of relationships involving X and Y, there are causal relationships (Meehl & Waller, 2002). In quasi experiments, researchers use preexisting groups, such as people with and without some neurological disorder. Cattell (1988) argued that the distinction is unimportant: "it is basically quite immaterial whether an 'act of God' [not the researcher] or an act of man [the researcher] has arranged the series of events under observation" (p. 22). However, with quasi experiments, it is difficult to calculate likelihoods about how different the groups are likely to be on various characteristics.

Even when people are supposed to be randomly assigned, randomization is not always successful. Consider a well-known example in which random assign-

ment went wrong in a study to test whether giving milk to children affected their height and weight. Gossett (1931) described how 20,000 children were randomly allocated to either a milk or a control condition, but teachers were allowed to substitute well-fed or ill-nourished children if the milk and control groups in their classrooms did not appear equally nourished. Gossett stated,

it would seem probable that the teachers, swayed by the very human feeling that the poorer children needed the milk more than the comparatively well to do, must have unconsciously made too large a substitution of the ill-nourished among the “feeders” and too few among the “controls.” (Gossett, 1931, p. 399)

Because of this bias, the study “failed to produce a valid estimate of the advantage of giving milk to children” (Gossett, 1931, p. 406). Random assignment still often goes wrong. There are numerous cases where, in supposedly carefully controlled randomized medical trials, the people administering the trial have biased the allocation (Berger, 2005).

The second assumption is that there is only one difference between the control and treatment conditions and that this is what causal inference should be based on. In much behavioral research, there are no true placebos. Because any effect can have multiple causes, it is critical for researchers to take care in identifying what the possible causes are. If there are several differences between the conditions, then caution is urged if the researcher is trying to claim that any single one of them had an effect.

It is worth placing a restriction on what can be considered a cause. Following Holland (1986), “put bluntly and as contentiously as possible . . . causes are only those things that could, in principle, be treatments in experiments” (p. 954). In practice, this means that some hypotheses, for example, the own-race bias in face recognition, are not causal. It is not appropriate to say that being of the same race caused the witness to be more accurate with the identification of the culprit.

Although Rubin’s (1974) model of causation is based on a simple experimental design and is closely linked with experimentation, it is also a useful framework for examining associative hypotheses. Associative hypotheses state that two types of events, X and Y, co-occur with some regularity. For these hypotheses to make sense, they must be defined with respect to a population of events. These events can vary according to the materials used, the physical environment, and the people involved (Clark, 1973; Wells & Windschitl, 1999). In a typical study, a researcher might measure two variables and calculate their correlation or odds ratio. These sample measures would be used to estimate the association in some population. The critical aspect of this inference is that the sample characteristics are representative of the population characteristics. In Figure 1, this would mean S is representative of P. If random sampling is done, researchers can calculate how likely it is to have a sample that is unrepresentative of the population. An example where random sampling failed is described by Fienberg (1971) regarding the 1970 draft, which determined if men were sent to Vietnam. Three hundred and sixty-six pieces of paper, one for each date during a year, were individually placed into 366 capsules, which were put into the box from which they were chosen. The way the capsules were put into the box was biased: People who were born in later months were much more likely to have lower lottery numbers (Spearman’s correlation =  $-0.84$ ) and therefore more

likely to be drafted than those born in earlier months. The draft was carried out differently the following year.

Random sampling from the population does not mean that the sample is representative of the population, but it does mean that the sample's characteristics are usually not very different from those of the population. Sampling theory quantifies the likelihood that the characteristics of the sample will be very different from those of the population. As with random allocation, this adds sampling error into the estimation. For causal hypotheses, it is less important whether the sample is representative of a population because a single atypical situation can falsify that a particular hypothesis holds generally. There are no restrictions on which things can be associated.

### *Effect Sizes, Generalization, and the Progress of Science*

Estimating how large an effect or an association is and for what set of situations each occurs is critical for evaluating any hypothesis. In psychology, many journals urge, and others require, authors to report the magnitude of the effect/association (Thompson, 1996; Wilkinson & Task Force on Statistical Inference, 1999; Wright, 2003). Given the emphasis of this article, it is particularly unfortunate that the statistical term for the magnitude of both effects and associations is *effect size*, but given its widespread use, this phrase is used in this article. Knowing what population any hypothesis refers to is also clearly important, although this is often not specified in empirical studies and is related to the ecological validity of a study.

The typical study investigating causal hypotheses differs from the typical study investigating associative hypotheses in the role that effect sizes and generality of the results play. First, X is different for the two forms of hypothesis. For causal hypotheses, X is the manipulation or treatment, either real or hypothesized, of some particular variables. For associative hypotheses, X refers to the differences within the sample on some variables. Second, there tends to be less concern about how representative of the population the sample is for studies investigating causal hypotheses. Because effect sizes in the behavioral sciences usually vary for different subsets of a population (Engels, Schmidt, Terrin, Olkin, & Lau, 2000), the reports of effect size and the associated confidence intervals are often used to convey more about confidence that the effect size is in a particular direction than about the size of the effect in the whole population. Finally, because the sample is often not representative of the population, generalization applies differently for studies designed to examine causal hypotheses. Experiments rejecting the null hypothesis allow the researchers to falsify that the null hypothesis,  $X \rightarrow Y$  CP, is true for all situations. To make this statement, it is necessary to reject the hypothesis in only one situation. Therefore, how representative of the population the sample is, is not that important. However, if the hypothesis is not rejected, then (depending on the power of the test) it is valid to accept that the effect is negligible in only one particular subset of the population. This is often not a very valuable scientific contribution, which is why experimentalists are often more concerned with Type I errors (incorrectly saying that an effect is present) than Type II errors (incorrectly saying that an effect is absent). This difference is less important when evaluating associative hypotheses because the sample is usually

assumed to be representative of some population, and therefore, the researcher can accept that an effect is negligible in a population. This can be a valuable contribution.

A study by Culhane and Hosch (2005) illustrates how the sampling is important. For a videotaped burglary trial, Culhane and Hosch compared verdicts of law enforcement officers and those with friends or family members who were officers with verdicts of other people. Their sample was composed of people who had been summoned for jury duty and were waiting to be allocated to a jury, and therefore, they had a very good sample of potential jurors in that jurisdiction. They found no statistical significant differences. Because of the high-quality sampling of jurors, Culhane and Hosch were able to claim that for this trial, there were at most only small differences between how these two types of jurors would behave. However, because they used only a single trial, they should have generalized their results to only this type of trial (Wells & Windschitl, 1999). Because the effect size may differ for different types of trials, Culhane and Hosch could not argue that there would be no difference in verdicts between these two groups of people in trials more generally.

How do causal and associative hypotheses differ for the progression of science? One prominent philosophy of science is based on Popper's (1959) notions of falsification. In fact, in defining science, the U.S. Supreme Court relied heavily on Popper's notions (*Daubert*, 1993). Hypotheses are put forward, and studies are designed that, depending on how the data turn out, can be consistent or inconsistent with the hypotheses. If the data are inconsistent with a hypothesis, then the hypothesis is either rejected or modified or else the data are discounted. If the data are consistent with the hypothesis, then the hypothesis is supported. The degree of support depends on many factors, including the power of the study and whether other plausible hypotheses are also consistent with the data. In psychology, this is often done through null hypothesis significance testing (NHST). Meehl (1967) described two very different ways to use NHST. The first, which he argued is typical in psychology, involves testing a hypothesis that is not believed to be true. This hypothesis is often that there is no difference between groups on some measure. For CP hypotheses, a statistically significant result allows the hypothesis, which was not believed to be true anyway, to be rejected as always occurring. The researcher can either change the CP conditionals so that this situation is not covered by the hypothesis or reject the hypothesis. Usually, the hypothesis is rejected, and the researcher claims that there is some difference between the groups, but the researcher can claim only that there is a difference in the situation studied. Because the hypothesis has not been believed to be true anyway, rejecting it in just a single situation is not seen as a great scientific breakthrough.

For example, in the United States, jurors may be excluded from some murder cases if their views toward capital punishment make them unable to render a guilty verdict in a capital case. Several studies have been conducted to examine whether people who oppose the death penalty are less likely to render guilty verdicts. These studies have differed greatly in their samples and their measures of death penalty attitude and likelihood to convict, and therefore, it is not surprising that they have come to different conclusions. Thus, the authors of each individual study would need to be cautious about generalizing their findings.

Allen, Mabry, and McKelton (1998) conducted a meta-analysis of 14 studies and showed that overall there was a relationship between these variables but also that there were greater differences among the associations observed in the studies (from  $r = .06$  to  $r = .60$ ) than could be attributable to chance. This means that no single study could provide a valid estimate for this association but that a meta-analysis can provide valuable information.

Meehl (1967) compared this with the second approach in which researchers test the predictions from a substantive model believed to be true. Meehl argued that this is the norm in physics. If the data fit the prediction, the model is stronger. If the data do not fit, then the hypothesis is rejected, the CP conditionals are changed, or the data are discounted. Meehl claimed that as methods improve, more precise measurements make it easier to reject uninteresting null hypotheses if using the first approach. However, with improved measurement, tests of substantive hypotheses become more stringent. Because the first approach is dominant in psychology (often, psychology students are even told that they should want significant results), Meehl (1978) argued, “psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else” (p. 807). Where psychologists have made quantitative predictions, there have been considerable advances. For example, within forensic psychology, researchers can tease apart the contributions of nature and nurture for antisocial behaviors using methods from behavioral genetics (Moffitt, 2005). Researchers are able to compare sibling pairs that vary in how closely related their genotypes are from unrelated siblings (e.g., adoptees) to monozygotic twins and therefore are able to make specific predictions for associations in phenotypes based on genetic relatedness (DeFries & Fulker, 1985).

### *Different Types of Hypotheses and the Daubert Criteria*

Given that causal and associative hypotheses require different methods and are about fundamentally different aspects of science, it is worth addressing how each relates to the admissibility requirements for expert scientific testimony as specified in *Daubert* (1993). Initially there was some debate about whether the *Daubert* criteria apply just to scientific expert testimony, in which case, Fodor’s (1991) arguments cited above suggest they might apply only to causal hypotheses, or whether they apply more broadly to other types of expert testimony (Faigman, 1995). In *Kumho Tire Co. v. Carmichael* (1999), it was argued that the *Daubert* criteria apply broadly to other types of knowledge, including knowledge of associations.

In *Daubert* (1993; Sanders, Diamond, & Vidmar, 2002), the Supreme Court ruled that expert testimony must be relevant and reliable and that in determining reliability, the judge should consider (a) whether the theory is falsifiable and whether it has been tested in this way, (b) the error rate, (c) whether the theory has been published in a peer reviewed source, and (d) whether the theory is generally accepted. Reliable science should meet all four criteria, according to *Daubert*. Surveys have shown that judges have altered the way they admit expert witnesses since 1993 (Dixon & Gill, 2002; Groscup, Penrod, Studebaker, Huss, & O’Neil, 2002; Krafka, Dunn, Johnson, Cecil, & Miletich, 2002) but that they often fail to consider all four of the *Daubert* criteria (Dahir et al., 2005).

Over the past decade, many scholars have debated whether particular hypotheses referred to by psychologists called on as experts in court, such as recovered memories (Faigman, 1995; Gordon, 1998), Rorschach tests (Grove & Barden, 1999), dissociative identity disorder and PTSD diagnosis (Grove & Barden, 1999), and lie detection machines (Saxe & Ben-Shakhar, 1999) and methods (Vrij, 2005), meet the *Daubert* criteria. These commentators have often found testimony drawing on these psychological theories deficient on one or more of the four *Daubert* criteria. Faigman (1995) noted that some expert testimony on eyewitness memory meets these standards, but there are many different topics on which experts could testify (Kassin, Tubb, Hosch, & Memon, 2001), some of which do not meet these standards.

Here, the issue is not whether any particular hypothesis meets the necessary criteria but whether each type of hypothesis—causal and associative—and the methods usually used to evaluate each can address the criteria. Because many scientific journals regularly publish both studies examining causal hypotheses and studies examining associative hypotheses, both of these types of hypotheses meet the peer review criterion. Also, because surveys have shown that experts generally agree about the reliability of some causal and some associative hypotheses (Kassin et al., 2001), these types of hypotheses also can meet the general acceptance criterion. The falsifiability and error rate criteria are more difficult to meet.

*Daubert* (1993) explicitly referred to Popper's (1959) notions of falsifiability, and this fits well with evaluation of causal hypotheses. An expert could argue, for example, that showups are more suggestive to eyewitnesses than are lineups (Stebly, Dysart, Fulero, & Lindsay, 2003). However, the research conducted on this topic has generally tested the hypothesis that there is no difference in false identification rates for the two methods. Thus, the finding that showups are more suggestive has not been tested. Someone could argue that these studies were testing a directional hypothesis and that, therefore, Stebly et al.'s (2003) meta-analysis falsified that showups are less suggestive, but the problem identified by Meehl (1967) and discussed earlier, about psychologists testing hypotheses in which one is not interested, does create problems for *Daubert's* falsifiability criterion.

The problem here is due to the way that much social science research is conducted, not being compatible with *Daubert's* (1993) reliance on falsifiability of a hypothesis to establish a scientific method. For example, matching the showup versus lineup research with notions of trying to falsify the hypothesis of interest requires some hand waving and argument. Notwithstanding problems in the way many psychology studies have been conducted and reported (Wilkinson & Task Force on Statistical Inference, 1999), many psychologists believe that falsifiability is not the only criterion for scientific progress. Some of the most important studies in forensic psychology (e.g., Haney, Banks, & Zimbardo, 1973; Milgram, 1963) are studies designed to demonstrate an effect, not to falsify one. For example, the Stanford Prison Study showed "the extremely pathological reactions which emerged in both groups of subjects" (Haney et al., 1973, p. 10). Another important advance that does not rely on falsification is how meta-analysts aggregate effect sizes for several studies. This has led to a meta-analytic

framework for conceptualizing scientific progress, different from falsification, that can be used for all research (Thompson, 2002).

The *Daubert* (1993) ruling also listed error rate as important. This criterion can be met in some areas of forensic psychology by a measure of effect size. For example, in risk-assessment research, one would want to know the hit rate (sensitivity) and false-alarm rate (specificity) for any diagnostic test. Depending on the publication outlet, psychologists are urged or required to report effect sizes for their studies for both causal and associative hypotheses. In both cases, they report the effect size for their study using their sample. Given that effect sizes may vary according to the stimuli and sample, if the researcher wants to generalize that effect size to the entire population of interest, then it is important that the stimuli and sample are representative of the population of interest. As discussed previously, those researching associative hypotheses are more likely to ensure that their sample is representative.

### Examples From Eyewitness Testimony Research

During the past 10 years, there have been significant advances in applying psychological theories and findings to increase understanding of the reliability of eyewitness testimony (Connors, Lundregan, Miller, & McEwen, 1996; Technical Working Group for Eyewitness Evidence, 1999; Wells & Olson, 2003). This is an area of psychology where there is interest in both causal and associative hypotheses, so it provides a useful domain to further illustrate the different types of hypotheses. Other topics, both in legal and forensic psychology and in other fields, can be analyzed in similar ways.

One of the most useful distinctions in eyewitness research is between system and estimator variables (Wells, 1978). System variables, such as how interviews and lineups are conducted, are those under the control of the judicial system. As such, the justice system can, theoretically, manipulate these variables. This means hypotheses about system variables, for example, about the effect of different ways that interviews and lineups can be conducted, are particularly well suited for experimental methods and, thus, causal hypotheses (Seelau & Wells, 1995). Estimator variables are those outside of the control of the judicial system, such as how long the eyewitness sees the culprit. Here, police, judges, and juries may be interested in both causal and associative hypotheses. They may be interested to find that increasing the time of exposure, while keeping everything else constant, increases memory accuracy (a causal hypothesis), but they may also be interested to find that eyewitnesses who view the culprit for a longer time (for whatever reasons) tend to have more accurate memories (an associative hypothesis). These different hypotheses require different empirical designs and should be used differently by those in the justice system. Next, I discuss topics of eyewitness memory research to illustrate the different ways that causal and associative hypotheses are used: own-race bias, emotion and memory, event duration estimation, lineups, and a few others.

#### *Own-Race Bias (Only Associative Hypotheses Possible)*

People are more accurate at recognizing faces of people of their own race than of other races (Meissner & Brigham, 2001; Meissner, Brigham, & Butz, 2005).

This associative hypothesis states that the probability of correctly recognizing a previously seen face is higher if the face is that of a person of the same race as the eyewitness. This does not mean that being of a certain race causes memories to be more accurate because it is not possible to manipulate race in the way that race is usually conceived (Sankar & Cho, 2002). Causes can be attributed only to variables that can be manipulated (Holland, 1986). Even if it were possible to modify someone's genes so that he or she was genetically of another race, this would presumably not leave all other things equal, thus invalidating the CP conditional.

This limitation does not imply that the own-race bias can be studied only in natural settings. The causal versus associative distinction is different from the laboratory versus everyday/naturalistic distinction. Most own-race bias research takes place in artificial and laboratory settings where dozens of to-be-remembered photographs are presented on a computer to participants. More ecologically valid designs include field experiments where members of the public of different races are approached by confederates of different races who later have to be identified (Wright, Boyd, & Tredoux, 2001). Furthermore, some archival studies have examined own-race bias (Valentine et al., 2003). Despite the variability in control, all these studies evaluate an associative hypothesis.

Some important variables that relate to the explanations of the own-race bias can be manipulated. One of the main causal explanations for the bias is that increasing the amount of contact with other races increases people's memory accuracy for these races. This is a causal statement. With respect to the participants, a sample of people could be given more exposure to people from other races. The data suggest that this decreases the own-race bias (Chiroro & Valentine, 1995; Sangrigoli, Pallier, Argenti, Ventureyra, & de Schonen, 2005; Walker & Hewstone, in press). The faces can also be manipulated. MacLin and Malpass (2001) have altered facial stimuli so that the race of the target is not clear. This manipulates some characteristics of the face, so it does address important causal hypotheses but does not manipulate race per se. The hypothesis changes from an associative one about own-race memory to a causal one about memory for certain facial characteristics that can be manipulated and that happen to covary with race. Thus, this new causal hypothesis can help to account for the associative own-race bias, indicative of the way in which causal and associative hypotheses are complementary.

### *Emotion and Memory (Both Associative and Causal Hypotheses Relevant)*

Interest exists both in the theories about emotion and memory and in their application to eyewitness testimony (Christianson, Loftus, Hoffman, & Loftus, 1991; Deffenbacher et al., 2004), the recovered memory debate (Loftus, 1997), and other areas of applied psychology. Most of the interest with respect to eyewitness testimony concerns the highly anxious, stressful, and traumatic conditions that accompany witnessing certain crimes. Two questions are often asked: First, are memories for highly emotional events different from memories for other events? Second, does the high level of emotion affect the accuracy and completeness of the memory? The first question is associative; the second question is

causal. As discussed at the beginning of this article, these are very different questions. Emotional events tend to differ from other events in many ways other than emotionality, so it is often difficult to isolate emotion to assess causality.

A study involving an associative hypothesis might involve looking at some set of events to see whether those that are more emotional produce more accurate memories than events that are less emotional. For example, Tollestrup, Turtle, and Yuille (1994) compared the descriptions of culprits from eyewitnesses of robbery and fraud. Part of their reason for choosing these cases was because of the different levels of arousal generally associated with these crimes. Although they found that the robbery witnesses recalled many more details, Tollestrup et al. were careful not to claim that the heightened level of emotion during these crimes caused the greater number of details being recalled. They hinted that the more likely reason was that the robbery witnesses involved in their study were generally questioned soon after the event, whereas most of the fraud descriptions were taken several months after the crime. Such caution in making causal attributions is laudable.

One of the main points of the Tollestrup et al. (1994) study was a plea for others to examine real-world cases and to be satisfied with associative hypotheses. The associative hypotheses that Tollestrup and colleagues could make were limited to the two types of crimes that they examined. Although these hypotheses are useful, if the interest is in differences among events more generally, then a more diverse set of events, which vary in emotionality, is needed. Several methods have been used to sample autobiographical memories (Brewer, 1988; Skowronski, 2005). Within the eyewitness testimony field, the norm for archival studies is to examine all lineups available (e.g., Valentine et al., 2003). This means that the associative findings are applicable to that population.

It is also possible to conduct experiments by varying how emotional the stimuli are. Several such studies have been conducted. Some of the early studies focusing on the weapon focus effect are summarized in Steblay (1992). Consider Christianson et al. (1991, Experiment 2), where participants watched either a neutral slide sequence or an emotional slide sequence. Both sequences showed a mother and her son leave a house and walk down the street. In the neutral sequence, the pair walked past a car, and the mother dropped the boy off at school before returning home. In the emotional sequence, the boy was shown on the hood of the car bleeding profusely, with "one of his eyeballs . . . hanging out" (Christianson et al., 1991, p. 696), and then being treated in a hospital emergency room. After a filler task, participants were asked the color of the car. Participants in the emotional condition had better memory of this central detail. The conclusion was that the manipulation caused the difference. However, did emotion cause the difference? Changing that aspect of the slide sequence certainly changed the emotion of the sequence (as manipulation checks showed), but presumably, it also changed other aspects. The researchers should have asked if emotion caused the difference or if one or more of these other aspects caused the difference.

The manipulation in Christianson et al. (1991) also made the event more distinct and unexpected than it otherwise would have been. Several researchers feel that these characteristics could have produced the observed effects. Mitchell, Livosky, and Mather (1998, Experiment 1; Pickel, 1999) found that somebody holding a stalk of celery, which was unusual in their scenario, produced an effect

similar to holding a handgun in a threatening manner. To conclude that emotion caused a difference, the researchers would have had to assume that any additional aspects that may also have been changed by the manipulation (e.g., how unusual the event was, the seriousness of the event, etc.) did not affect the memory.

### *Estimating Event Duration (Making Assumptions for Causal Inference)*

Police officers are often interested in the duration of an event. Not only can this be of importance in itself but also jurors are sometimes told that they should consider the duration over which an eyewitness viewed an event when judging the reliability of the eyewitness's account. Burt and colleagues (Burt, 1993; Burt & Popple, 1996) have conducted several studies investigating people's duration estimates. In one study, Burt (1999) was interested in the relationship between using words that implied that the action during the scene was rapid (e.g., the thief ran through the store, as opposed to the thief walked through the store) and the estimated duration. In particular, the interest was in whether the words used to describe an event caused a certain duration estimate. Two approaches for investigating this hypothesis have been described, one by Burt (1999) and one by Pedersen and Wright (2002). Each requires assumptions for the causal inference to be valid.

Burt (1999) showed participants a video of a bank robbery, then later asked them to describe the event and estimate the duration. He found that the more action words used to describe the event, the shorter the estimated duration, concluding that "causation is from the construction of a narrative to describe the event to an estimate of the event's duration" (Burt, 1999, p. 353). Thus, he implicitly argued that although there may be several things that guide construction of a narrative (e.g., the memory of the event), it is the language used that affects event duration estimates: memory → narrative → duration. To assume this causal order means the narrative has a direct effect on the duration estimate; it is not spurious, with some other variable affecting both the construction of the narrative and the duration estimate (Simon, 1954). This assumption explains why it is often difficult to argue for a specific causal hypothesis on the basis of correlational data. An alternative conclusion from these data is that memory of the event dictates both how the event is described (memory → narrative) and the duration estimate (memory → duration) and that narrative and duration are not causally related.

Pedersen and Wright (2002) took a different approach. They manipulated the way people described an event by telling them to describe it either as if they were a tabloid newspaper journalist or as if they were an eyewitness speaking to a police officer. The instructions stressed that the tabloid journalist should use expressive high-impact language and that the eyewitness should describe just the facts. The journalists did use more high-impact language than the eyewitnesses. Therefore, if language use had directly affected duration estimates, then the manipulation also should have altered duration estimates. However, to make this causal attribution, it would be necessary to assume that the manipulation did not change anything other than the language that might affect duration estimation.

Assumptions are necessary for making inferences from any study. The main question for Burt (1999) is not whether using high-impact words varied with other characteristics and for Pedersen and Wright (2002) is not whether asking people

to write in a certain style changed more than just the use of high-impact words. Clearly, assuming either of these would be wrong. The question is whether the other characteristics that are likely to vary with language use and the style manipulation are also likely to affect duration estimates. To make causal attribution, it is necessary to consider, in theory at least, the variables in isolation even if it is not possible to isolate them in practice. This is why causal attributions can be made only about variables that one can at least contemplate being manipulated (Holland, 1986), allowing scientists to make causal attributions in areas where manipulation is both somewhat unethical (neuropsychology) or impossible (astronomy).

### *Quasi Experiment in Lineups (Statistical Example Disentangling Covariates and Causes)*

Conducting a lineup properly is a difficult task. In the United Kingdom, several suites dedicated to conducting lineups have been established. It is now the norm to use these suites instead of police stations (Valentine et al., 2003). When the first two suites were introduced in London, the police service wanted to know if conducting the lineups in the suites produced different results than if they had been conducted in the police stations (Wright & McDaid, 1996). This is a causal hypothesis. It was not practical to randomly assign cases to the lineup suites versus the police stations, and there were systematic differences between the suites and the stations (see Table 2 of Wright & McDaid, 1996) that are associated with (and probably causally related to) different outcomes (see Table 3 of Wright & McDaid, 1996). For example, there tended to be a longer period of elapsed time between the crime and the lineup at the suites, and as elapsed time increased, so did the likelihood of choosing innocent fillers.

This was a quasi experiment where naturally occurring groups were compared (Cook & Campbell, 1979). To make a causal inference, it would be necessary either to assume that there were no systematic prelineup differences between the groups that could affect the outcome or to make some statistical adjustments so that groups were somehow matched on prelineup differences that could affect the outcome. There is discussion (and some disagreement) about how best to make these statistical adjustments, and no single right way exists (Wainer, 1991). Because there were differences on key estimator variables between the locations, Wright and McDaid (1996) statistically controlled for differences on some of these estimator variables. They first entered into the model some of the estimator variables that they felt could be related to the outcome (elapsed time, whether the crime was violent, and race of the suspect) and then tested whether the location of the lineup had any additional predictive value. Adding the variable for location did improve the model, suggesting that where the lineup took place affected the outcome.

If it is assumed that once the estimator variables have been taken into account, all other things are equal or, at least, nothing that differs will have an impact on the response outcomes, then any statistically significant effects of where the lineups took place can be used to infer causality. This assumption is clearly important and merits further testing. In particular, researchers should test for overdispersion, which means that the variability of response proportions is larger

among conditions than would be expected by chance. This can be due to several causes (e.g., Browne, Subramanian, Jones, & Goldstein, 2005), but it is often assumed to be because the model has been misspecified (Hinde & Demétrio, 1998).

This final example is of a complex hypothesis that involves both associative and causal hypotheses. There are associative hypotheses stating that the estimator variables are associated with both where the lineups take place and the identification outcomes, and there are causal hypotheses stating that some system variables influence outcomes. Controlling for estimator variables is assumed to make the CP assumption more plausible. With these assumptions in place, causal hypotheses about the system variables can be evaluated. The response is *Y*. The hypothesis states that the probability for each outcome of *Y* is a function of the estimator variables and system variables, and there is an associated probabilistic error term. Thus, this design evaluated a hypothesis that was both causal and associative and included a probability term.

### *Other Eyewitness Topics: The Surveys by Kassin and Colleagues*

The four topics described above illustrate the different types of hypotheses of interest to eyewitness researchers, but other eyewitness topics can also be considered. The most definitive list of eyewitness topics was created by Kassin and colleagues (Kassin, Ellsworth, & Smith, 1989; Kassin et al., 2001) for use in their surveys of eyewitness experts. In their most recent survey, respondents were asked their opinions about the reliability of evidence for 30 different topics (see Table 1, Kassin et al., 2001, p. 408). Of these, 11 were causal hypotheses, and most of those remaining were associative hypotheses, although some may have been interpreted as causal by many of the respondents. On the basis of this list, it appears that about half of the topics of interest to eyewitness testimony researchers involve causal hypotheses, and about half involve associative hypotheses.

Examples of the causal hypotheses are (italics have been added to the words suggesting that the hypothesis is causal; the numbers are those assigned to topic–statement pairings in Table 1 of Kassin et al., 2001, p. 408) Statement 2 (“The presence of a weapon *impairs* an eyewitness’s ability to accurately identify the perpetrator’s face”), Statement 14 (“Hypnosis *increases* the accuracy of an eyewitness’s reported memory”), and Statement 20 (“Alcohol intoxication *impairs* an eyewitness’s later ability to recall persons and events”). Each of these should be interpreted as including the CP clause. For example, the hypnosis statement (Statement 14) should read, “If two identical people are in the exact same situation and if one is given hypnosis, that person will have more accurate memory than the other person.” This hypothesis is addressed by the typical laboratory study. An alternative associative hypothesis is whether the typical eyewitness who has had hypnosis to help retrieve a memory will be more accurate than one who has not. In the real world, eyewitnesses are not randomly assigned to hypnosis and control conditions. It is likely hypnosis would be used only as a last resort (or not at all, depending on the laws in the jurisdiction), and therefore, it is likely that even without hypnosis having any causal effects, there would be differences in accuracy between these two naturally occurring groups.

Examples of associative statements (probably correctly interpreted by most

survey respondents as associative) include (numbers are those assigned to topic–statement pairings in Table 1 of Kassir et al., 2001, p. 408) Statement 8 (“An eyewitness’s confidence is not a good predictor of his or her identification accuracy”), Statement 18 (“Eyewitnesses are more accurate when identifying members of their own race than members of other races”), and Statement 30 (“The more quickly a witness makes an identification upon seeing the lineup, the more accurate he or she is likely to be”). A critical aspect of associative hypotheses is deciding for which types of events the hypothesis should hold. For example, the association between confidence and accuracy could be increased by including some very easy and some very difficult trials. The final survey item, Statement 30, is interesting because the associative hypothesis, referring to the finding that accurate identifications tend to happen more quickly than inaccurate identifications (Weber, Brewer, Wells, Semmler, & Keast, 2004), is clearly associative. If it were changed into a causal hypothesis, where identification speed was purported to affect accuracy, it is likely that the effect would be in the opposite direction as forcing people to rush responses can impair performance on many cognitive tasks (Wickelgren, 1977).

Many of the survey questions were presented as associative hypotheses but may have been incorrectly interpreted by some respondents as causal, for example (numbers are those assigned to topic–statement pairings in Table 1 of Kassir et al., 2001, p. 408), Statement 6 (“The less time an eyewitness has to observe an event, the less well he or she will remember it”), Statement 10 (“Judgments of color made under monochromatic light [e.g., an orange streetlight] are highly unreliable”), and Statement 17 (“Eyewitnesses have more difficulty remembering violent than nonviolent events”). Because eyewitness researchers tend to use experimental methods and to couch their conclusions in causal terms (even when inappropriate), it would not be surprising if some respondents interpreted these survey statements as saying that lessening the exposure time worsens memory (Statement 6), removing much of the color information in light worsens judgments of color (Statement 10), and controlling for all other characteristics, violence negatively affects memory (Statement 17).

Taking Statement 10 as an example (because the causal hypothesis is indisputable, based on what is known about color perception), how would a respondent answer this question if he or she read it as an associative hypothesis? The respondent would probably assume that the relevant situations were all those events that police investigators might be interested in (i.e., crimes) and consider which of these situations occur in monochrome light. Because the situations in monochrome light (under streetlights at night, in clubs, etc.) tend to be those also associated with witnesses being intoxicated by alcohol and other drugs, respondents might be able to answer this question without reference to the psychophysics of color perception.

The surveys by Kassir and colleagues (Kassir et al., 1989, 2001) were designed so that the respondents did not have to spend a long time answering the questions. Therefore, they did not spend a long time explaining the meaning of each statement. The researchers sought to differentiate causal and associative hypotheses. Presumably, this is why both Statement 2 and Statement 17 were included: Statement 2 is about the causal hypothesis; Statement 17 is about the related associative hypothesis. However, the importance of this distinction may

not have been clear to many respondents or to many in the eyewitness field. Therefore, caution is advised before using these results to argue that each of their survey statements is generally accepted unless there is certainty that the respondents interpreted that particular survey statement appropriately. Eyewitness testimony is an area of research that has had many successes bridging the scientist–practitioner divide, resulting in improvements in different justice systems (Wells et al., 2000). For this to continue, it is important that both the scientists and the practitioners are clear about whether they are referring to causal hypotheses or to associative hypotheses.

## Conclusions

There are different types of hypotheses that scientists consider. Some are about classification. These apply across all situations and are best viewed as defining certain phenomena in terms of others. Thus, action is reaction, and experiencing trauma is one of the criteria for diagnosing PTSD. The two types of hypotheses that psychologists are most concerned with are causal and associative.

Causal and associative hypotheses differ in how they should be explored. When making causal inference, it is useful to have an experimental design with random assignment. The sample, in terms of people and stimuli, is often not representative of any larger populations because the researcher is usually trying to falsify the hypothesis that some particular hypothesis holds in all situations. Effect sizes are often not reported, perhaps because the field has developed implicit knowledge that effect sizes are likely to vary across samples. Thus, causal hypotheses and experimentation are related to Popper's (1959) notion of falsifiability, but they are different.

In contrast, associative hypotheses make sense only if described in relation to some population. Because of this, how people and materials are sampled from the population is of vital interest. This allows generalization to the population to be made about the size of the effect. Thus, for associative hypotheses, the effect size is more important than it is with causal hypotheses. Perhaps this is why the correlation coefficient,  $r$ , is one of the few effect size measures almost always reported in psychology journals and is a statistic often used for associative hypotheses rather than causal ones.

In the past, exploring causal and associative hypotheses has defined distinct areas of psychology. In fact, a major thread of Spearman's (1904, p. 205) classic article on there being a single general intelligence advocated a correlational psychology in juxtaposition to experimental psychology. Fifty years later, Cronbach (1957) described how these disciplines developed separately. Another 50 years on, people often talk about mixing methods and how people from different research perspectives can work together. Although statistical approaches permit associative and causal hypotheses to be considered within the same design, it is important that the fundamental differences continue to be stressed, so that people are aware that causal and associative hypotheses are very different.

## References

- Allen, M., Mabry, E., & McKelton, D.-M. (1998). Impact of juror attitudes about the death penalty on juror evaluations of guilt and punishment: A meta-analysis. *Law and Human Behavior, 22*, 715–731.
- Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist, 44*, 1185–1193.
- Berger, V. W. (2005). *Selection bias and covariate imbalances in randomized clinical trials*. Chichester, England: Wiley.
- Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.
- Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 21–90). Cambridge, England: Cambridge University Press.
- Browne, W. J., Subramanian, S. V., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit over-dispersion. *Journal of the Royal Statistical Society: Series A, 168*, 599–613.
- Burt, C. D. B. (1993). The effect of actual event duration and event memory on the reconstruction of duration information. *Applied Cognitive Psychology, 7*, 63–73.
- Burt, C. D. B. (1999). Categorisation of action speed and estimated event duration. *Memory, 7*, 345–355.
- Burt, C. D. B., Mitchell, D. A., Raggatt, P. T. F., Jones, C. A., & Cowan, T. M. (1995). A snapshot of autobiographical memory retrieval characteristics. *Applied Cognitive Psychology, 9*, 61–74.
- Burt, C. D. B., & Popple, J. S. (1996). Effects of implied action speed on estimation of event duration. *Applied Cognitive Psychology, 10*, 53–63.
- Cattell, R. B. (1988). The principles of experimental design and analysis in relation to theory building. In J. R. Nesselrode & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 21–67). London: Plenum Press.
- Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own race bias in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48(A)*, 879–894.
- Christianson, S.-Å., Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 693–701.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.
- Connors, E., Lundregan, T., Miller, N., & McEwen, T. (1996). *Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial*. Alexandria, VA: National Institute of Justice.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.
- Culhane, S. E., & Hosch, H. M. (2005). Law enforcement officers serving as jurors: Guilty because charged? *Psychology, Crime, and Law, 11*, 305–313.
- Dahir, V. B., Richardson, J. T., Ginsburg, G. P., Gatowski, S. I., Dobbin, S. A., & Merlino, M. L. (2005). Judicial application of *Daubert* to psychological syndrome and profile evidence: A research note. *Psychology, Public Policy, and Law, 11*, 62–82.

- Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 113 S. Ct. 2786 (1993).
- Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior, 28*, 687–706.
- DeFries, J. C., & Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior Genetics, 15*, 467–473.
- Dixon, L., & Gill, B. (2002). Changes in the standards for admitting expert evidence in federal civil cases since the *Daubert* decision. *Psychology, Public Policy, and Law, 8*, 251–308.
- Engels, E. A., Schmidt, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine, 19*, 1707–1728.
- Faigman, D. L. (1995). The evidentiary status of social science under *Daubert*: Is it “scientific,” “technical,” or “other” knowledge? *Psychology, Public Policy, and Law, 1*, 960–979.
- Feynman, R. P., Leighton, R. B., & Sands, M. (1963). *The Feynman lectures on physics: Mainly mechanics, radiation, and heat*. Palo Alto, CA: Addison-Wesley.
- Fienberg, S. E. (1971, January 22). Randomization and social affairs: The 1970 draft lottery. *Science, 171*, 255–261.
- Fodor, J. A. (1991). You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanations. *Mind, 100*, 19–34.
- Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson’s “ecological approach.” *Cognition, 9*, 139–196.
- Galton, F. (1879). Psychometric experiments. *Brain, 2*, 149–162.
- Gordon, J. D. (1998). Admissibility of repressed memory evidence by therapists in sexual abuse cases. *Psychology, Public Policy, and Law, 4*, 1198–1225.
- Gossett, W. [writing as “Student”]. (1931). The Lanarkshire milk experiment. *Biometrika, 23*, 398–406.
- Groscup, J. L., Penrod, S. D., Studebaker, C. A., Huss, M. T., & O’Neil, K. M. (2002). The effects of *Daubert* on the admissibility of expert testimony in state and federal criminal cases. *Psychology, Public Policy, and Law, 8*, 339–372.
- Grove, W. M., & Barden, R. C. (1999). Protecting the integrity of the legal system: The admissibility of testimony from mental health experts under *Daubert/Kumho* analysis. *Psychology, Public Policy, and Law, 5*, 224–242.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.
- Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. London: Edward Arnold.
- Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). Study of prisoners and guards in a simulated prison. *Naval Research Reviews, 9*, 1–17.
- Hinde, J., & Demétrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis, 27*, 151–170.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.
- James, W. (1890). *The principles of psychology*. Cambridge, MA: Harvard University Press.
- Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The “general acceptance” of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist, 44*, 1089–1098.
- Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the “general acceptance” of eyewitness testimony research. *American Psychologist, 56*, 405–416.

- Krafka, C., Dunn, M. A., Johnson, M. T., Cecil, J. S., & Miletich, D. (2002). Judge and attorney experiences, practices, and concerns regarding expert testimony in federal civil trials. *Psychology, Public Policy, and Law*, 8, 309–332.
- Kumho Tire Co. v. Carmichael, 526 U.S. 137, 119 S. Ct. 1167 (1999).
- Loftus, E. F. (1997, September). Creating false memories. *Scientific American*, 277(3), 70–75.
- Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about “weapon focus.” *Law and Human Behavior*, 11, 55–62.
- MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law*, 7, 98–118.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103–115.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Meehl, P. E., & Waller, N. G. (2002). The path analysis controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods*, 7, 283–300.
- Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7, 3–35.
- Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, 19, 545–567.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Mitchell, K. J., Livosky, M., & Mather, M. (1998). The weapon focus effect revisited: The role of novelty. *Legal and Criminological Psychology*, 3, 287–303.
- Moffitt, T. E. (2005). The new look of behavioral genetics in developmental psychopathology: Gene-environment interplay in antisocial behaviors. *Psychological Bulletin*, 131, 533–554.
- Ofshe, R., & Watters, E. (1995). *Making monsters: False memories, psychotherapy, and sexual hysteria*. London: André Deutsch.
- Pedersen, A. I. C., & Wright, D. B. (2002). Do differences in event descriptions cause differences in duration estimates? *Applied Cognitive Psychology*, 16, 769–783.
- Pickel, K. L. (1999). The influence of context on the “weapon focus” effect. *Law and Human Behavior*, 23, 299–311.
- Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.
- Rosenfeld, B., & Lewis, C. (2005). Assessing violence risk in stalking cases: A regression tree approach. *Law and Human Behavior*, 29, 343–357.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Sanders, J., Diamond, S. S., & Vidmar, N. (2002). Legal perceptions of science and expert knowledge. *Psychology, Public Policy, and Law*, 8, 139–153.
- Sangrigoli, S., Pallier, C., Argenti, A.-M., Ventureyra, V. A. G., & de Schonen, S. (2005). Reversibility of the other-race effect in face recognition during childhood. *Psychological Science*, 16, 440–444.
- Sankar, P., & Cho, M. K. (2002, November 15). Genetics: Toward a new vocabulary of human genetic variation. *Science*, 298, 1337–1338.
- Saxe, L., & Ben-Shakhar, G. (1999). Admissibility of polygraph tests: The application of scientific standards post-Daubert. *Psychology, Public Policy, and Law*, 5, 203–223.
- Seelau, S. M., & Wells, G. L. (1995). Applied eyewitness research: The other mission. *Law and Human Behavior*, 19, 319–324.

- Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, *49*, 467–479.
- Skowronski, J. J. (2005). In diversity there is strength: An autobiographical memory research sampler. *Social Cognition*, *23*, 1–10.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.
- Stebly, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior*, *16*, 413–424.
- Stebly, N. M., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*, 523–540.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000, October). Better decisions through science. *Scientific American*, *283*(4), 82–87.
- Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: U.S. Department of Justice, Office of Justice Programs.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 24–31.
- Tollestrup, P. A., Turtle, J. W., & Yuille, J. C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 144–160). Cambridge, England: Cambridge University Press.
- Valentine, T., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology*, *17*, 969–993.
- Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, *11*, 3–41.
- Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147–151.
- Walker, P. M., & Hewstone, M. (in press). Intergroup experience and the own-race face effect: A case study. *Applied Cognitive Psychology*.
- Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10–12-second rule. *Journal of Experimental Psychology: Applied*, *10*, 139–147.
- Wells, G. L. (1978). Applied eyewitness testimony research: System variables versus estimator variables. *Journal of Personality and Social Psychology*, *36*, 1546–1557.
- Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist*, *55*, 581–598.
- Wells, G. L., & Olson, E. A. (2003). Eyewitness identification. *Annual Review of Psychology*, *54*, 277–295.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125.
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information-processing dynamics. *Acta Psychologica*, *41*, 67–85.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
- Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136.

- Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2001). A field study of own-race bias in South Africa and England. *Psychology, Public Policy, and Law*, 7, 119–133.
- Wright, D. B., & McDaid, A. T. (1996). Comparing system and estimator variables using data from real line-ups. *Applied Cognitive Psychology*, 10, 75–84.

Received June 20, 2005

Revision received January 25, 2006

Accepted January 25, 2006 ■

### Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.