CHAPTER 1

From Empirical Findings to Custody Evaluations

ROBERT M. GALATZER-LEVY, JONATHAN GOULD, and DAVID MARTINDALE

EVERYONE WORKING IN the area of child custody agrees that these decisions should be based on credible information. In this chapter, we describe a framework for evaluating the significance of these findings. Expert custody evaluators generally, though often informally, proceed through a three-stage process in forming opinions in custody matters:

1. Based on their knowledge of the kinds of facts that are likely to be relevant to the decision, they collect data about the particular situation.
2. Using previous knowledge and further research, they refer to studies that may be relevant to the facts they find, assess the pertinence of the studies to the situation, and apply their findings to the facts at hand.
3. Weighing the relative importance of these facts, they attempt to integrate the resulting conclusions into an overall scientifically reliable recommendation.

Often the evaluator will not carry out these steps in sequence, or the process may loop back on itself. For example, on reviewing the pertinent scientific literature, the evaluator may see the need to gather additional facts before making a recommendation. In developing an opinion, evaluators commonly go through several iterations of this process. In practice, the care with which data about custody matters is gathered varies greatly. Many evaluators use their training and an unsystematic reading of the scientific literature as their sole source of information about the state of research in the matters about which they give opinions; their process of reaching an opinion is incompletely thought through. Legal professionals are
often uncertain about how to assess the statements of behavioral science experts in these matters. We describe several considerations that can lead to more credible opinions and more meaningful assessment of experts’ statements.

Currently, the legal criteria for using scientific expert opinion are in flux. Roughly speaking, the courts appear to be moving from a sociologically based view of truth (Truth is what experts in the field believe) toward a method-based view (Truth is what is discovered through the appropriate application of scientific research methods). This shift reflects an increasing unwillingness in our society to accept statements as true simply on the authority of credentialed experts. Most people familiar with the field would heartily endorse this move. Despite their limitation, scientifically based opinions are more likely to provide pertinent information to decision makers and information that is more easily evaluated than opinions with other bases, such as personal experience. Where possible, we want to use scientifically based opinions to inform decisions. At the same time it is important to recognize that the complexity of human psychological life entails that custody decisions often must be made in situations that do not exactly fit with available research. In these cases it is particularly important that evaluators neither leave the court without assistance in understanding that complexity nor claim levels of certainty that they do not have.

Evaluating the merit of scientific evidence places tremendous demands on the courts. The boundaries of adequate scientific method generally, but especially in disciplines that study human behavior, are the subject of intense dispute among behavioral scientists and philosophers of science. Even when there is reasonable agreement about these boundaries, technically challenging questions arise in assessing research studies. These questions are often difficult for people with years of training in methodology to resolve. Attempts to resolve them in a legal context demand much of the courts.

The main difference between scientific and other forms of knowledge is that scientific knowledge is always accompanied by an assessment of its own credibility. Each step in a scientifically conducted study is at least potentially subject to open scrutiny and evaluations. The question of the truth of the statement always remains open and subject to test. A typical scientific research report includes a section on materials and methods that describes how the reported data were collected and the means used for analysis. It does so in such a fashion that a reader could, at least potentially, replicate the research. The report also includes specific reasoning that shows how these data are related to any conclusion drawn and an indication of the extent to which alternative possibilities have been assessed and the likelihood of their being true. Thus, by their very nature scientific opinions are never certain, and their limitations and defects should be made clear.
Although scientists have elaborate methods for assessing the credibility of scientifically based statements, no particular method is inherently necessary for an investigation to be scientific. In studying human behavior, many investigators find statistical methods extremely informative. However, when treated as a sort of magical ritual to make an investigation appear to be scientific, statistics can provide a false impression of the status of the investigator’s findings (Cohen, 1990; Salsburg, 1985). Nonquantitative studies have yielded some of the most useful information about human behavior and psychological function, whereas some quantitative studies, despite elaborate statistical trappings, are of little use. In the study of people, a trade-off commonly occurs between “extensive” and “intensive” study (Chas-san, 1979). To reach statistically valid conclusions, many subjects are usually necessary. The richer the matter studied (in effect, the more variables studied), the more subjects are needed. However, studying many subjects in depth is difficult and time consuming. The following example displays this contrast.

In studying the long-term effects of divorce Wallerstein (e.g., Wallerstein & Johnston, 1990) and McLanahan and Sandefur (1994) used distinctly different methods that provide distinctly different results. Wallerstein studied a small number of subjects in depth over many years using an open-ended interview technique. Her studies provide clear, rich pictures of some of the common psychological configurations seen in children of divorce, with particularly good insight into their subjective experience. They do not, however, tell us how common these configurations are, to what extent they result from interacting aspects of the subjects’ lives, including the historical and demographic peculiarities of the sample studied, or the extent to which the subjects differed from other people like themselves. McLanahan and Sandefur based their studies on survey data collected by others. These surveys largely explored easily quantified aspects of the child’s situation such as family income and years of school completed. They involved very many subjects and systematically compared the subjects with individuals who had not been involved in divorce or who had lost parents in other ways. At the price of a much less rich picture of the children’s situation these authors provide reliable, quantitative information about the impact of divorce. Both types of research are scientific in the sense that they are clear about the means by which their data was collected and the logic that ties their conclusions to those data. Each has advantages and disadvantages: The Wallerstein studies provide the kind of richness but not the level of certainty or generalizability we would like; the McLanahan and Sandefur study is clear and convincing but does not reflect the psychological depth that most of us would want in reaching conclusions about children’s well-being. In this instance, the studies complement one another in the sense of looking at matters from different views and coming to compatible conclusions.
In the following material, we explore how custody evaluators collect information about specific cases, how they can assess literature on the subject and its applicability to the particular case, how they form a framework for thinking about the relative importance of the information obtained through these processes, and how they deal with problems such as potential bias.

FACTS

In making custody assessments, mental health professionals make observations and collect reports from numerous sources. These provide a set of facts that can be further processed. Decisions about which facts to collect and what credence to give them are of great importance.

The information an evaluator needs will depend on the questions to be answered. Too frequently, custody evaluators fail to focus on the specific issues pertinent to the situation and develop conclusions that go beyond or are irrelevant to the questions asked. They often collect both too much and too little information. Failure to gather pertinent information may leave the examiner ignorant of significant aspects of the situation. Irrelevant information can sway both the evaluator and the trier of fact in ways inconsistent with the goals of the evaluation. When working in a legal context, evaluators should develop a clear picture of the governing law to address the issues in the case. Many jurisdictions include specific issues that the court must consider in making decisions. For evaluators to be helpful in the decision process they must collect information pertinent to these questions. Many states' statutes specify that the judge must consider the residential custodial parent’s capacity to support a relationship with the noncustodial parent in deciding with whom the youngster should stay. Some experts in child custody do not agree that this should be a central consideration (Goldstein, Freud, & Solnit, 1975). However, whatever opinions examiners may have about the law, they impede the judicial process when they fail to address the issues the court must address. Child custody experts have much to contribute to public policy, but the child custody evaluation is not the place to try to transform the law.

Most custody decisions do not depend per se on the general psychological health of the parents but do depend, in part, on the impact of the parents' psychological functioning on the child. Many mental health professionals, however, approach custody evaluations as they would the diagnostic assessment of a patient. The implications of the resulting diagnosis is often obscure in its meaning for the parent’s interactions with the child. Nonetheless, the presence of a serious-sounding diagnosis or test finding may sway the trier of fact or derail the focus of the evaluation. This is particularly problematic with regard to psychological test reports. Although the reports themselves are
usually carefully worded to indicate that the subject shows a pattern of responses consistent with some condition or commonly observed among individuals with a particular condition, the inexperienced reader is likely to finish reading the report with a sense that the subject suffers from many and severe psychological disturbances.

The way facts are collected often profoundly affects their content and significance. The information collected in custody evaluations can be profoundly shaped by the collection process. For example, an interviewer who is perceived as sympathetic may be told many things that an apparently unfriendly interviewer is not told. Differentiating between the factors that substantially influence the data collection and those that do not can be difficult. This is partly because the subject’s response to the interviewer may be so idiosyncratic that the interviewer is unaware of it (e.g., the interviewer may remind the subject of a pleasant or unpleasant person the subject has known). These problems are particularly marked in interviewing children, where such factors as the child’s desire to please the interviewer or fears of parental disapproval may greatly shift the information provided (Ceci & Bruck, 1995). The greatest danger occurs when the evaluator is unaware of the factors in the interview or testing situation that shape the information received. Although this hazard has been particularly well studied with regard to interviews about abuse allegations, interviewers are at constant risk of suggesting “correct” answers to the questions through responses such as approving remarks, gestures, and interested further questions when the interviewee responds in a certain fashion.

Because the attitudes of the interviewer and the intentions of the interviewee are so different depending on whether the interviewer is performing as an evaluator or a therapist, these roles should not overlap (Greenberg & Daniel, 1997). In talking to a therapist, it is in the client’s best interest to provide as full and accurate a picture as possible of the situation; in talking to an evaluator it is in the client’s best interest to provide a picture that will lead the evaluator to concur with the client’s opinions. The way an interviewer listens to the interviewee depends on an assessment of the interviewee’s intentions. This distinction is also important because most custody evaluators were trained primarily as clinicians. It is difficult enough for them to keep in mind that the underlying intentions of litigants are different from those of patients without the further confusion that results from trying to fill more than one role.

In a forensic situation most subjects are highly motivated to have their point of view prevail. They are likely therefore to use whatever means they think will be effective to convince the evaluator of the merits of their own position. More or less subtle forms of dishonesty are common in custody evaluations. These range from overt lies to significant omissions to enhance
presentation of the subject’s point of view. Attempts to form alliances with the subject in the interest of discovering what is best for the child sometimes help minimize the consequences of such tendencies. However, the fact that custody is being litigated is in itself an indication that the parents have been unable to lay out facts dispassionately and agree on conclusions. Especially in so emotionally fraught a matter as the custody of one’s children, it is rare that a litigant is willing or able to forgo efforts to achieve their desired result in favor of providing the evaluator with objective information. Evaluators should proceed on the basis that the information being provided is intended to influence them rather than to provide as accurate a picture as possible of the situation. When interviewers take on more than one role, they are likely to become confused about which role they occupy at a given time. The therapist’s goal is often to communicate support and understanding to the client, which will aid the client in overcoming some difficulty. Evaluators have no such intention and are, in fact, appropriately skeptical that the information being given has any purpose other than persuasion. The vigorous questioning that may be appropriate in an evaluation and the use of collateral sources of information used by evaluators are rarely appropriate for the therapeutic situation. It is generally not possible to be both a good therapist and a good evaluator. For this reason, all guidelines for custody evaluators recommend avoiding a dual role (Ackerman, 1995). (For a discussion of the numerous other reasons to avoid being both therapist and evaluator at the same time, see Strasburger, Gutheil, & Brodsky, 1997.)

As with other matters, a scientific attitude in collecting facts is not characterized by the search for absolute certainty but rather in careful attention to possible sources of error and their correction. To this end, evaluators note the apparent credibility of sources of information, including the motives an interviewee may have for distorting information; the manner in which the information is presented; knowledge of common forms of distortion, including inaccuracies of memory; psychological test results that may pertain to credibility; and how consistent the information is both internally and with other sources of information. Some studies show that mental health clinicians are not particularly good judges of the veracity of the reports they receive (Ekman, 1992). Some of these studies are flawed in that they either placed the clinician in a trick position where he or she had little reason to anticipate that anyone would try to be less than honest or the clinician was given information of a type different from that ordinarily used by clinicians in making judgments of this kind. Methods are available to increase the clinician’s detection of deceit (Ekman, 1992). Still, the clinical assessment of credibility remains significantly imperfect, and clinicians cannot claim to be better at assessing truthfulness than other individuals based on research. Where significant facts in a case are controversial, as when
one of the litigants is accused of physical violence, the court, with its capacity to gather sworn testimony subjected to cross-examination and to assess evidence, is often in a far better position than the evaluator to determine the truth of various claims. In some jurisdictions the evaluator may ask the court to make a determination of the pertinent facts. When this is not possible evaluators may state how their opinions would be affected by various fact scenarios.

Human memory and perception are more fallible than most of us would like to think. When collecting information from many sources, it is essential to remember that distortions are common and not necessarily the result of intentional deception (Schachter, 2001). Once people frame a view of a matter they are likely to selectively perceive and recall events in a fashion that supports their preconceptions. For this reason, evaluators should, where practical, utilize collateral sources of information (Dietz, 1996) and weigh conflicting descriptions of events. Evaluators themselves may forget or misperceive aspects of a situation. The more carefully, systematically, and contemporaneously events are recorded, the less likely examiners are to introduce significant distortions into situations. In addition to using ordinary thoroughness, many examiners follow comprehensive checklists when making observations during interviews and examinations to ensure that pertinent matters are recorded. Structured interviews serve a similar function. These guides have the advantage that the examiner is likely to observe and record items systematically. They have the disadvantage that they may discourage pertinent open-ended exploration and the noting of observations that do not fit within the preassigned format.

When collecting information and making observations that are highly complex or available only once, recording the information may be useful. Many examiners have discovered the value of videotaping or audio taping at least portions of the evaluation. Sometimes significant elements of interpersonal interactions between parent and child during an evaluation become clear only through the retrospective evaluation of recorded interactions.

An experienced examiner observed the interaction between a mother and her 6-year-old child, who had been removed from her care because of neglect associated with the mother’s depression and substance abuse. Although the examiner felt ill at ease observing the interaction, the mother appeared animated, interactive, and concerned for the child’s interests, as manifested by her encouraging play with several toys that she knew he enjoyed. When the examiner reviewed a videotape of the session, the source of discomfort became clear. Over the 45-minute session, the mother never once responded to something the child initiated. In fact, when the child introduced a new topic, the mother continued speaking about the matter she had been focused on before the youngster interrupted, giving no indication that she had heard
what he had said. Even had the examiner been astute enough to pick up this pattern during the interview, without the videotape it would have been difficult to thoroughly and rigorously document it.

Studies of the suggestibility of children show that youngsters’ reports may be profoundly influenced by subtle suggestion in previous and current interviews (Ceci & Bruck, 1995). Although this problem has been particularly noted in regard to sexual abuse allegations, it is important in any situation in which youngsters’ historical reports are important. Simply having been asked about an event may constitute the basis for the later “recollection” of the occurrence. For this reason, it is extremely useful to retain videotapes of sessions in which children report historical events. Not only do such tapes make it more possible to assess the extent to which the examiner unwittingly employed various means that are known to produce distorted reports, but they also make it possible to document situations in which the subject incorporated a question as a memory.

Selective recording may be challenged on cross-examination, where the question will naturally arise as to why the entire session was not recorded or with an implication of bias as to what was recorded. The evaluator should also expect that recordings will be subpoenaed, with the possibility of data being taken out of context for possible use in cross-examination. For these reasons we believe that where possible, all interviews should be recorded.

THE EVALUATION OF STUDIES

The evaluator’s opinions should be based in the knowledge accumulated in his or her field. In the behavioral sciences, as in any scientific discipline, the rate of growth of an evaluator’s discipline is always too great for any one person to maintain a comprehensive knowledge of recent developments. Evaluators do have a responsibility, however, to remain aware of major recent trends in their field so that they can place specific knowledge in their discipline in context and reasonably evaluate whether older studies are consistent with current disciplinary thinking.

Following World War II researchers found substantial evidence that certain styles of communication from parents to their children induced Schizophrenia in the child (Lidz, Cornelison, & Fleck, 1958; Lidz, Fleck, & Cornelison, 1965). On the basis of these studies an examiner might reasonably conclude that a youngster with early signs of Schizophrenia was the victim of what was termed a “schizophrenogenic” environment and would have appropriately considered this in making custody and visitation recommendations. More recent studies strongly suggest that the psychological environment in which the future schizophrenic grows up has a significantly smaller impact than biological factors in determining the course of the disease. Further
exploration of the data on which the schizophrenogenic parent hypothesis was based demonstrated the flaws in the original analysis of the data (Lidz, 1984). Rather than the child’s Schizophrenia being the result of abnormalities in parental behavior, the already disturbed child appears to produce the parental abnormalities.

Scientific knowledge characteristically changes across time. New research changes the best information available on a topic. This is a strength, not a weakness, of scientific knowledge. As noted, an expert giving the best information available in the early 1960s would have stated that the parents of a schizophrenic youngster brought about or at least contributed significantly to their child’s pathology through their disturbed form of communication. This is not the view supported today by available evidence, and the expert would state that. Part of the expertise expected of evaluators is a sufficient knowledge of their discipline and a sufficient capacity for critical thought so that they are able to assess the scientific status of published information they encounter.

It is a reasonable expectation that evaluators will not only stay current with their field through reading and conference attendance but that they will also use the extensive resources available on the Internet to research topics in which there is ongoing conceptual and empirical development or that involve areas of knowledge with which they are not fully conversant. Although consultation, whether in the traditional form of discussing a case with a trusted colleague or by way of Internet discussions, is one of the ways that mental health professionals commonly deal with problems they cannot solve on their own, a consultant’s ideas should be used as a guide to developing an opinion of one’s own, not as the basis of that opinion. This is primarily because a scientific argument should rest on its own merits and not rely on an appeal to authority. Secondarily, if the opinion is really that of the consultant rather than the evaluator, the consultant should be made available for cross-examination.

In assessing particular studies, several factors are ordinarily considered, including the source of the study, the status of the publication, the rigor of the reported study, and comparison of the study to related research. Each of these factors will contribute to the evaluator’s degree of confidence in a particular study.

Researchers’ training should reflect the practical appreciation of sources of potential error, learning how to avoid elementary mistakes, and becoming familiar with the range of problems that commonly beset a research effort. Various forms of direct and implicit certification identify the level of the researcher’s training and experience. The researcher’s highest academic degree and its source may indicate the extent of his or her research training. Many mental health professional degrees, including the MSW, some PhDs,
and MD, include little or no formal training in research methods. A PhD in psychology from a major university is very likely to have a strong background in research methodology. Similarly, the researcher’s career, as indicated by academic rank, record of publication, and the degree to which those publications are cited by other researchers, suggests the level of regard with which the researcher is held within the professional community and the extent to which he or she has a track record of producing research that has been found to be of high quality by individuals qualified to review it. A further indication that the work has been thoroughly reviewed is found in mention of sources of financial support. Major foundations and government agencies generally impose high standards of rigor for research proposals that they fund, so that funding from such sources suggests the work is of high quality.

Scientific publications are of four types: publication in refereed or peer-reviewed journals, publication in nonrefereed journals, chapters in edited volumes, and monographs. Publications may report new research findings or summarize and comment on older findings. Publication in a refereed journal indicates that two or more readers who are recognized authorities in the field concur that the article is of sufficient rigor and significance to merit publication. The quality of peer review itself varies substantially among journals, and failings of this system have been noted on numerous occasions (Burnham, 1990; Lock, 1990). Journals with many submissions are likely to impose higher standards on their authors than those that must work to fill their pages. Factors such as whether the peer reviewer can identify the author despite a blind reviewing process (McNutt, Evans, Fletcher, & Fletcher, 1990) and the peer reviewer’s conceptualization of his or her role (Horrobin, 1990) can strongly influence the peer review process. Nonetheless, it remains an effective method for ensuring the rigor of published materials.

The quality of non-peer-reviewed publications is more problematic. Lack of peer review may lead to lack of quality, but sometimes these articles are of particularly high quality because they are written by leading researchers in the field at an editor’s invitation. The editorial review of chapters in edited books also varies markedly, resulting in uneven quality. In the hard sciences (chemistry, physics, etc.), researchers rarely publish significant new results in edited volumes or as monographs. The situation is different in the behavioral sciences, where significant research commonly appears in these formats. This means that in the behavioral sciences the consumer of research must often independently assess the quality of the research presented. An additional factor external to the study itself that should activate the reader’s skepticism is the likelihood of the authors’ financial gain from particular outcomes. For example, if a study demonstrates the validity of a commercially available test
or the effectiveness of a drug in which the researcher has a financial interest, this should alert the reader to the greater likelihood of the report’s being shaped by those interests.

Scientific literature suffers from a common problem with regard to negative results. A finding of the presence of some effect is more likely than the absence of an effect to stir interest both in those who decide whether an article should be published and in those who write the article. As a result, negative findings tend to be underrepresented in the published literature (Dickersin, 1990). Negative results are of two kinds: those that show the absence of an effect and those that do not reproduce the findings of previous investigators. Though they are underrepresented in the literature of most sciences, in assessing the credibility of research findings both of these results are of great importance. Despite researchers’ best efforts, factors of which they are unaware may influence outcomes. The failure to reproduce a finding in another study is a strong indication of such factors.

ASSESSING THE CONTENT OF SCIENTIFIC PUBLICATIONS

The assessment of research on the basis of its content and process is a complex matter. We simply outline here some major themes and particularly important areas. The reader should refer to detailed treatises to learn about the means of assessing research findings (Campbell & Stanley, 1963).

Is It Science?

The U.S. Supreme Court’s decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993) attempted to address the question of what research and opinion should be characterized as “scientific.” In doing so, the Court attempted to answer a problem that has concerned those interested in the philosophy of science since at least the sixteenth century. Prior to this time, knowledge was regarded as embedded in authoritative texts from which correct information could be mined by those with the authority to do so. Starting with the work of Francis Bacon, a new way of knowing emerged, having as its chief reference observable facts about the external world. The interpretations of these facts was not a matter of authority but was based on the possibility that any adequately trained person could make the same observations and that the chain of reasoning from facts to interpretation could be made explicit. Over the past four centuries, but especially in the twentieth century, scientists and philosophers of science refined and debated these core questions of what methods are scientific and when an investigation is properly labeled as such. The resulting ideas are both more technically demanding and more controversial than the Supreme Court’s decision suggests.
The *Daubert* decision actually goes in the opposite direction from much of the most recent work in the philosophy of science. *Daubert* holds that science is characterized by its methods and that its claims to truth are based on those methods. This was the dominant view among methodologists until approximately 20 years ago. More recently, students of the history and sociology of science have argued that the actual way science is conducted is not related so much to method as to the social structure of scientific communities (Feyerabend, 1975; Kuhn, 1970; Latour & Woolgar, 1986). The newer picture of the nature of scientific knowledge is, in fact, more compatible with the Frye rule (*Frye v. U.S.*, 1923), which held that expert opinions must be compatible with widely accepted ideas within the expert’s field.

However, if we accept the view that there are characteristic, if not utterly precise methods of inquiry that set scientific knowledge apart from other kinds of knowledge, then the centerpiece of such knowledge is that ideas are tested against reliable observations about the world. If I say X is a scientific fact, I must be able to demonstrate this through observations about the world. Moreover, terms must have commonly accepted, agreed-on definitions. Observations must be replicable both across raters and through time. The question of which data will determine whether X is a scientific fact is not simple. The things referred to in X must be clear enough so that researchers can agree whether what is observed is an example of this particular sort of thing. If someone says, “All heffalumps are green,” there is no way this proposition can be scientifically tested until we know and agree what a heffalump is and how to define “green.” This becomes a significant problem in much psychological research because apparently meaningful terms may in fact have very unclear empirical referents. For example, mental health professionals are often asked to comment on the “dangerousness” of an individual. But until this term is defined in a way that refers to agreed-on, specific, empirically observable factors, the behavioral scientist has no way to collect data pertinent to the issue. Phrases used in custody evaluations such as “the best interest of the child” have little meaning in a scientific context until they are conceptualized in terms of reliable, observable phenomena.

Scientific statements are subject to tests against observations. Consider this process conceptually. For many years, the proof of a hypothesis was regarded as the collection of observations that confirmed it. If a scientist claimed that all cells have nuclei, he or she could test and confirm this assertion by observing many cells under the microscope. But this method of testing statements can prove faulty in two ways. First, it can involve a “sampling error”: For some reason the scientist may not look at those cells that lack nuclei. Second, although the statement appears to be about the external world, it may really be in a form in which no observation could show it to be wrong. For example, if a theory stated, “Every cell contains a nucleus, but some of these nuclei are
invisible,’’ no microscope observation could prove the statement false, so any
positive evidence about this matter would be essentially meaningless. Popper
(1959) argued that only propositions for which one could imagine an experi-
mental outcome contrary to the proposition should be regarded as scientific.
This is the concept of falsifiability. For a hypothesis to be falsifiable, one must
be able to design an experiment that would lead to the conclusion that the
hypothesis was false. In the study of human behavior this can be a difficult
requirement.

Consider the psychoanalytic hypothesis that all young boys have anxiety-
laden sexual desire for their mother that they manage either by failing to
become socialized to the taboo against incest or by forcing the idea from
awareness, minimizing its emotional significance. Stated this way, no empiri-
cal data could possibly lead to the conclusion that the original statement was
false because both the presence of erotic interest in the mother and also the
absence of such interest would be taken as confirming the hypothesis. Thus
the proposition is not falsifiable.\(^3\)

Not every investigation must involve a crucial experiment that tests a hypo-
thesis. Most investigations simply provide additional reason to believe or
disbelieve a statement. However, to be scientifically meaningful a hypothesis
must be capable of some form of critical test. As mentioned earlier, scientific
knowledge rests on publicly available observation and reasoning. The argu-
ment that someone simply knows more or has more direct access to truth is
 unacceptable in scientific discourse. In some ways, this is a theoretical rather
than a practical distinction. Most people cannot follow the reasoning involved
in the theory of general relativity, nor can they understand, much less perform,
the experiments that confirm that theory. At no point, however, did Einstein
say or imply, “I know more about the universe than you do and this theory is
simply right.” His reasoning was based on logical steps and empirical evi-
dence. (We are talking here about how the theory was substantiated, not how
Einstein discovered it.) When experts claim to be offering a scientific opinion
about human psychological functioning, they must be held to this same
standard. The empirical and logical steps leading to a statement must be
capable of being laid out and not be dependent on authoritative assertions.
This can be a real problem in behavioral science studies, in which a convincing
rhetorical ability may be confused with scientific methodology (Spence, 1994).
Vivid case histories are likely to impress readers and remain with them even
when clear data contradict conclusions the researcher derives from them.

In an experiment still often quoted in psychology textbooks, John B.
Watson claimed to demonstrate his theory that phobias result from Pavlovian
conditioned reflexes. In a conditioned reflex a stimulus that regularly elicits a
response (the unconditioned stimulus) is paired with one that ordinarily does
not elicit that response (the conditioned stimulus). This often results in the
conditioned stimulus eliciting the response. The most famous example is Pavlov’s experiment; after a ringing bell and the presentation of food had been paired, Pavlov’s dogs learned to salivate when the bell was rung. Watson proposed that phobia resulted in a similar way when a noxious unconditioned stimulus was paired with an otherwise neutral situation. He vividly described an experiment in which “Little Albert,” an 11-month-old child who had previously shown no fear of rats, cried and avoided the rat after it was presented to him paired with the stimulus of a loud, clanging noise (Watson & Rayner, 1920). The experiment became the backbone of a theory of phobias that to this day has many adherents (Wolpe, 1990). Watson’s original description still makes compelling reading. However, attempts to replicate the experiment repeatedly failed (Seligman, 1971), a fact that received little notice as the rhetorical power of Watson’s writing was so great.

It is worth keeping in mind that there are other kinds of valuable knowledge besides scientific knowledge. We argue only that if knowledge is presented as scientific it should be scientific, and that courts are entitled to know the status of the opinions presented to them. Because of the enormous prestige of science in our society, many people are tempted to describe their opinions and beliefs as scientific when they are not. Personal values and common sense may be presented as though they had scientific backing when they do not; this is misleading to decision makers who are relying on the information provided to have a scientific basis.

In child custody matters, experts often testify based on their experience or clinical experience. It is important for all those concerned to treat these opinions for what they are worth. They may be of considerable value, reflecting accumulated knowledge based on real situations. However, opinions are subject to a variety of problems and rarely reflect the systematic collection of data characteristic of scientific investigation. When the expert does not recognize these limitations it causes problems. The opinion may reflect the expert’s experience with a special subpopulation; for example, a mental health professional who works with emotionally ill adults who were abused as children is likely to believe that abuse usually has devastating psychological effects because the abused individuals he or she sees are all psychologically disturbed. Also, when information is not systematically collected or recorded, vivid experiences are likely to stand out. Before the era of effective antibiotic therapy tonsillectomy was a routine procedure in the United States because most physicians found that some very ill children got dramatically better following the procedure; the children who were not affected did not stand out in the same way. Similarly, experiences that are congruent with the expert’s views are likely to stand out. Every therapist of some experience has treated individuals who, after an unrewarding time
spent in some other form of therapy, did well in treatment with them. Therapists are commonly convinced that such experiences indicate the general superiority of their own therapeutic approach, at least for this type of patient, and fail to appreciate that similar experiences probably occur involving their own former patients and that other therapists probably enjoy success rates similar to their own. Finally, clinical experience is usually limited to the period during which the subject is in treatment. This can easily lead to a misunderstanding of the long-term course of a situation. The mental health professional sees the outcome in terms of when services were completed, which may not reflect important changes in the course of the condition over years.

**DO THE DATA SUPPORT THE CONCLUSIONS?**

In attempting to answer a question scientifically, data are collected that are believed to address the issue. In assessing whether data support a conclusion, it is important to ask how well this is accomplished. Doing research, especially research about people, is often extremely difficult and time consuming; at the same time, researchers want to produce meaningful findings. As a result, there is often some slippage between data and conclusions, usually in the direction of drawing stronger conclusions than the data warrant. Much of the methodology of scientific investigation is directed at attempting to get as much information from data as possible while avoiding reaching conclusions that are not supported by the data.

Many of the issues discussed earlier with regard to the information collected in assessments apply to data collected for research. Accurate descriptions of the setting in which the data are collected, attention to the accurate recording of those data, and the minimizing of reliance on memory and complex perception are even more important in the collection of research data. Most research publications carefully specify these matters. In assessing a research publication, the methods section should be reviewed with these considerations in mind. Surprisingly often, methods sections will indicate limitations in the work that otherwise would not be recognized. For example, for many years it was common in biomedical research to use only male subjects, at least ostensibly because the use of female subjects complicated the analysis of data. This information, which certainly opens up the question of the applicability of these studies to women, is invariably present in the methods section and can be used to correct the impression that the results are universally applicable. When a methods section is missing from a paper, this should raise a red flag about the soundness of the reported observations.

Much psychosocial research involves various instruments and test batteries designed to measure some aspect of psychological function. Several
thousand such instruments are sufficiently well developed to have been described in multiple publications and reviews. Each claims to measure some aspect of human function. The validity of a measure is the extent to which it, in fact, measures what it claims. When the matter to be measured is clear and there is some sort of gold standard against which to compare the result, determining validity is relatively easy. Thus, if we wanted to decide whether a scale validly measures weight we would compare its results to widely accepted standards, such as those maintained by the National Bureau of Standards. But for many of the matters we try to address in studying people, no such simple absolute standards exist. We must first clarify what we mean by the terms we use and then see how well a measure reflects these concepts. For example, if we have a test that purports to measure whether a person is a “good parent,” we would need to clarify what was meant by this term and the extent to which the test measured the resulting qualities.

Validity always refers to a purpose for which the test is being used. A test may be valid for one purpose but not another (e.g., grades in college mathematics courses may be valid predictors of grades in physical science courses but not grades in English courses). The concept of validity has been the subject of considerable refinement and clarification. Face validity refers to the observation that a test, on the face of it, is relevant to the quality measured. For example, performance on a test in which the subject is asked to do arithmetic problems on the face of it shows that subject’s abilities to do that sort of problem. Content validity refers to a more thorough analysis in which we explore the extent to which a test measures matters pertinent to the question we are trying to answer. For example, a test of parenting ability might be shown to measure factors known or believed to contribute to parental function. However, measures need not have content validity to be useful. Certain important measures may accurately reflect an aspect of psychological function even though they do not necessarily directly address it. For example, although an individual claiming to have unusual experiences is not necessarily delusional, reports of such experiences may be highly correlated with delusions. Correlation validity refers to the extent to which test results correlate with significant aspects of the person’s current or future performance. The correlation need not reflect any other meaningful relationship between the item measured and its correlate. Construct validity refers to the extent to which a measure makes conceptual or scientific sense. In the absence of a gold standard, the validity of a measure is assessed by how well it fits within a network of measures that are used to study related concepts, known as the nomothetic net. In assessing the validity of a measure, researchers commonly combine several points of view about validity. In assessing the usefulness of a study, the various points of view regarding validity may serve as guideposts in mapping the investigator’s thinking.
Looked at from a different angle, the reader of published studies should always be skeptical about the way the author uses terms. Simply because a test is said to measure a quality that goes by the same name as one with which the reader is familiar does not mean that the test measures that familiar quality. One must look at the way the test was put together and demonstrated to be related to the underlying construct to know whether, in fact, the test measures that construct well.

All the issues discussed in Chapter 4 on testing apply to tests used in research studies. Especially in the case of projective tests and tests that have not been well studied, the reader should not assume they are as solidly based as one might hope. For many years, the standard text of psychological testing, which focused primarily on projective tests, was a volume by Rapaport, Gill, and Schafer (1945) based on the diagnostic test battery developed at the Menninger Clinic. This book formed the basis of many research and clinical studies. Though the studies claimed to include normal controls, these turned out to be Kansas State Troopers. With no offense meant to these men, it is reasonable to doubt that they reflect the ordinary members of the community. Major criticism can be raised of the research supporting the validity of the most commonly used scoring system for the Rorschach, a widely used projective test (Groth-Marnat, 1997). Much of the supporting research is correlational, with little information given about the actual false-positive or false-negative rates of inferences based on the use of this test.

An important aspect of validity is sensitivity. If the instruments used in a study cannot pick up significant differences between subjects, important aspects of the situation will not be adequately studied. The more sensitive a measure, the less reliable it is likely to be (i.e., the less likely it is that repeated measures of the same quantity or measures of the same quantity by different individuals will be the same). Obviously for a measure to be meaningful it must also be reasonably reliable. Researchers must demonstrate or refer to demonstrations that the measures they use are both appropriately sensitive and reliable if their claims are to be credible.

It is rare that findings of research studying people will show the clear delineation seen in the physical sciences. If we do a study of whether weights fall at the same rate when dropped in a vacuum, we expect that all the weights will so fall and would regard the experiment as incomplete if a small proportion of the weights fell at a different rate without explanation. In studies of people, things are rarely so straightforward; among the many factors that are likely to affect any significant aspect of human living, we rarely find one factor so powerful that it almost always causes a given effect. Because studies of people seldom have the sharp clarity of studies in the physical sciences, it is necessary to develop tools for assessing whether an observed relationship results from an actual relationship or occurs by chance.
Over the past century and a half, scientists who have had to deal with the messy data typical of the study of people have developed a discipline that addresses this problem: statistics.

The discipline of statistics is devoted to the question of how to reliably extract information from observations and how to estimate the degree of certainty with which that information is asserted (Stigler, 1986). It is usually difficult to extract meaningful patterns from the welter of information with which an investigator is confronted. For example, we might observe declining school performance in youngsters following divorce. We would want to determine the relationship of this finding to possible specific changes in the children’s behavior (inattention in class, decreased attendance, indifference to good performance, unavailability of parental support in doing academic work) and the changes in the children’s environment (the fact of divorce, change of residence, parental reactions to divorce, community reaction to divorce). Finding a systematic approach to answering such questions is one job of statistics. A second job is estimating to what extent we should rely on the findings. In this example, if we study 10 youngsters involved in divorce and discover that 8 of them are less attentive in class, 7 have parents who have become less engaged, and 6 of these cases overlap, are we justified in drawing any conclusions about the relationship of these factors? To what extent is the overlap the result of chance? What is the likelihood that we would have a similar finding with another group of youngsters? How well do the 10 youngsters we have studied represent children with school difficulties following divorce? As anyone who has observed gambling or estimated probabilities knows, human intuition about the likelihood of probabilistic events is extremely poor. Statistics often provides quantitative estimates of how likely it is that a finding results from chance or reflects a real difference; more generally it produces estimates of how likely various experimental outcomes are given an underlying causal model. Even an outline of the statistical methods commonly used in studying the behavioral science issues pertinent to divorce would require far more space than we have available (see, e.g., Hopkins, Hopkins, & Glass, 1996, for a standard introduction to these issues; see Keren & Lewis, 1994, for a comprehensive review of them). However, we will briefly outline some common sources of confusion that arise from statistical studies in the behavioral sciences and some common methodological problems of which readers should be aware in thinking about the usefulness of the study.

As discussed earlier, the mere presence of a statistical or quantitative treatment in a research paper may suggest a greater level of authority and credibility than the paper deserves. In assessing the quality of a study, remember that statistics and quantity are only tools for finding information. If used wisely to gather important information, they can be exceedingly
useful. However, because of their prestige, they can also hide substantial limitations in the study.

The concept of statistical significance is often misunderstood. Often given as a probability, statistical significance states the likelihood that a finding would arise by chance in a given situation. In the behavioral sciences, findings are generally regarded as significant if the likelihood of their being chance results is less than 1 in 20 ($p < .05$). Failure to demonstrate statistical significance indicates only that the study fails to demonstrate the hypothesis. It says nothing about the truth or falsity of propositions not being specifically studied. Providing a large enough sample to achieve statistical significance, even when a phenomenon is real, may be difficult; then the distinction between truth and statistical significance becomes particularly important. In long-term outcome studies it is far easier to collect information about such matters as economic well-being than psychological states in depth. As a result, we do not have good statistical studies showing whether children’s views of themselves are negatively altered by parental divorce. Some authors have concluded that the absence of such findings means that such effects do not occur or occur only as a result of other, incidental factors. The data support no such conclusion. They instead point to the absence of adequate studies to address the question.

The presence of statistical significance may be equally confusing. A study may indicate a probability of less than 1 in 20 for a relationship between two variables simply because so many variables were studied. If we study 15 relationships, the probability that one of them will have a probability of less than 1 in 20 becomes substantial. Statistical significance does not necessarily point to a meaningful relationship. In particular, such matters as causation cannot be properly inferred. A relationship between two variables may exist, but this may result from a common cause. Statistical significance does not necessarily point to meaningful differences; it simply points to the extent of certainty with which one can say there is some difference. Some differences, though definitely present, are nonetheless practically meaningless. Statistical significance tells us little about the size of the effect being observed. The numbers 3.0000 and 3.0001 are different from one another with absolute certainty, yet there are very few contexts in which this difference matters.

Although many journals now require that all articles involving statistical reasoning be reviewed by a competent statistician, publications that have not been subjected to such review surprisingly often have simple statistical mistakes. A common error is to use statistical tests that are not appropriate to the data. Many statistical tests assume that a variable is distributed in a certain way in a population (e.g., that a given variable, such as height, is distributed in a bell-shaped curve). When this is not the case, the reasoning
that supports the computation of various statistical measures may no longer apply and these measures may produce misleading results.

An important group of methods commonly used in statistics involves correlations and regressions. In almost all situations involving people, many things are happening at once and it is the job of the researcher to discover how these things are interrelated. A correlation study explores the extent to which two variables change together. Thus the extent to which children’s behavior becomes disruptive in the period immediately after a divorce varies with the age of the child. The description of this relationship is a correlation. In many circumstances, correlations can be described as a single number, the correlation coefficient, which measures how much one variable changes with the other. In regression analysis, the situation is conceptualized somewhat differently. If many factors contribute to an observed change, one can think of that change as being the result of summing those factors. Studying the change in many situations may allow the researcher to describe it as the sum of various factors, each weighted by some constant. This is called a regression analysis. Both forms of analysis aim at showing the relationship between important variables. In the case of each analysis, we can discover how much of the change is not accounted for by the factors we have studied. Properly used correlation and regression analysis are powerful statistical tools for unraveling complex situations.

A common error that particularly affects decision processes is equating correlation with causation. If two situations are causally linked it makes good sense that altering one will change the other. If the link is a mere correlation, however, such a change does not make sense. In the example discussed earlier concerning communications between parents and their schizophrenic children, we saw how correlation and causation can be confused. In the initial studies, a solid correlation was observed between the quality of parent-child interaction and the child’s diagnosis. However, the researchers erroneously concluded that this correlation represented a causal relationship, in which the parent’s communication style caused the offspring’s illness.

As mentioned earlier, statistical methods have obtained a status in the behavioral sciences that sometimes goes beyond their actual usefulness. In particular, it is extremely difficult to study complex situations over time using the standard methods of sampling statistics. In the attempt to gather sufficient quantifiable data for statistical analysis, it often becomes necessary to simplify the situation to such an extent that factors that are commonly believed to be important cannot be addressed in a study. Additionally, for all their limitations, in-depth studies of individuals over time have contributed much of what we know about human psychological function in complex, emotion-laden situations, such as the ones we are trying to address in
custody evaluations. In recent years, many methodologists have recognized that the old, conceptually weak case history method can be improved in several ways to produce very informative case studies (Yin, 1994). By paying close attention to the way conclusions are drawn and generalized, researchers have found case studies to be an increasingly effective means for studying complex psychological function. In addition to care in collecting and interpreting data, contemporary case study methodology often focuses on closely monitoring the time course of events (e.g., whether an improvement occurs in a subject at some regular interval following an intervention). When assessing reports of case studies or groups of case studies, readers should not simply discard them as methodologically unsound but rather examine the extent to which they use contemporary methods to ensure quality.

**IS THE STUDY CONSISTENT WITH OTHER RESEARCH IN THE SUBJECT?**

Although a study may be carefully performed and its careful analysis may show no glaring failings, it is still appropriate to ask how well it coincides with related studies. In its most formal version, this process takes the form of replication, usually in a different context. Replication is often particularly important when studying behavioral and psychological issues because factors of which the investigator may be unaware can influence outcomes. When a study appears significantly inconsistent with other research findings, a high level of skepticism about that study is appropriate and a search for the causes of the differences should be made. Because isolated studies are subject to substantial errors, behavior scientists rely on them only when they must.

Studies of the same subject matter are often difficult to compare. Differences in the population studied and the study procedure may make it unclear how much weight should be given to each study. In the past four decades, a means of comparing and consolidating groups of studies has emerged. Called meta-analysis, the method is continuing to evolve and has at times been controversial (Garfield, 1991; Hunt, 1997). Yet meta-analysis provides a systematic way to bring together, compare, and use comparable data on the same topic in a manner far superior to the traditional review article.

Findings that are inconsistent with other research are not necessarily in error. However, the differences in the findings must be explained for the work to be taken seriously. A particular group of problems arises with regard to studies of therapeutic and similar interventions. Initial evaluations are commonly performed by enthusiasts for the intervention, who, in addition to other difficulties, are often naive about methodology. The very enthusiasm of the investigators working on a new form of treatment may produce nonreplicable positive results unrelated to the purported mode of action of the treatment. Particularly when an intervention is reported as producing
striking successes when other methods have not, the likelihood that the research can form a reliable basis for recommending that therapy is small. Custody decisions often require that a parent receive treatment for a condition that impacts the child negatively on the basis of the treatment’s enthusiastic endorsement by a mental health professional. These recommendations may be contrary to the broad consensus among students of the same conditions that these conditions are not significantly responsive to treatment.

APPLYING EMPIRICAL STUDIES TO PARTICULAR CASES

In the best of all possible worlds, experts would have a firm database of findings about the best interests of children in various circumstances, match the child’s situation to the database, and give a clear, empirically based opinion about how the particular youngster’s best interests could be served. Rarely is this possible. Several factors make the situation more complex and difficult. We have already discussed the challenges of assessing the individual situation and reaching meaningful research conclusions. When we try to bring the two together additional questions arise. First, we need to ask whether the particular child’s situation is actually an instance of the type studied in the research. Second, we need to ask how important the finding is to the child’s situation. Third, we need to know whether the finding can be meaningfully translated into something that can actually be done in the real situation. Finally, we need to think through, given the imperfection of our knowledge, what standards should be applied to the quality of expert knowledge in custody decisions.

DOES THE RESEARCH APPLY TO THIS CHILD?

Research is always done on a sample drawn from a population. Thus a researcher might study the impact of visits with a noncustodial parent by taking his subjects from 8- to 10-year-old urban African American boys of low socioeconomic status whose parents have been divorced for at least two years. Assume that the study was of high quality and produced clear recommendations about the frequency of such visits that would most benefit the child. Is this study useful in making decisions about a middle-class White girl whose parents are in the midst of a divorce? Is it useful in making decisions about middle-class African American boys who are otherwise similar to the boys originally studied? Stated generally, if we find a situation where the subject of an evaluation belongs to the population from which the study sample is taken, we can apply the findings of the study directly. However, in other circumstances we need to ask how well or poorly the findings apply.
One approach to this problem begins with asking why we might expect the findings not to apply: In what way does the subject of evaluation differ from the research sample? In our example, many investigators might believe that African American boys of low socioeconomic status tend to have limited relationships with paternal figures and would particularly benefit from those relationships compared with other youngsters. Insofar as this belief is correct, the importance of supportive contact with a non-custodial father is likely to be greater than for other populations of youngsters. Having identified likely perceived differences, one can then ask whether there are empirical studies that suggest these differences are meaningful. If (as is the case) most studies of the effects of father absence on the well-being of youngsters showed that there is little difference in its impact as a function of race (Mclanahan & Sandefur, 1994), then the evaluator may conclude that the difference between a study population of African Americans with low social status and the population from which the subject comes is unlikely to be significant for the purposes at hand. It is important to remember that scientific findings are characterized by the degree to which they are self-critical, not by their definitive nature. When applying studies to individual children, it is the evaluator’s job to analyze the reasoning and findings that lead to his or her conclusions and the extent to which they are likely to be true—not to achieve absolute certainty.

HOW IMPORTANT IS THE FINDING?

All real custody decisions involve factors that weigh on each side. One parent may be able to spend more time with the child, whereas the other may provide the child with access to superior schools. One parent may have a better sense of the child’s emotional life, whereas the other may provide a better model of socially acceptable behavior. In many instances, significant empirical evidence of the value or disadvantage of various parental qualities will be available from the research literature. However, the comparative importance of these various factors has rarely been systematically studied. Some methods of evaluating custody implicitly weight these factors in relation to one another (Bricklin, Elliott, & Halbert, 1995). Still, implicitly or explicitly, in coming to a recommendation the evaluator must weigh the relative importance of findings.

As discussed previously, studies should not only tell us that some factor makes a difference but also give some picture of the magnitude of that difference; this is called the effect size. When comparing parents’ abilities to care for the child, effect sizes from pertinent studies should be a central concern. Sometimes effect size is addressed in terms of common sense: If one parent engages in behavior that threatens the child’s life and the other
provides limited opportunity to play with peers, we do not need a systematic study to conclude that the latter parent should have custody. However, more subtle situations may require precisely such a comparative analysis; for example, with certain forms of parental psychiatric illness, empirical findings about the impact on children run contrary to most people’s intuition (see Jenuwine & Cohler, Chapter 13).

In many instances, the best that evaluators can do is make their analysis as explicit as possible. This will allow evaluators and others who use their findings to explore the extent to which implicit formulation, including the evaluator’s personal values and biases, have found their way into the recommendations.

CAN THE RECOMMENDATIONS BE IMPLEMENTED?

In the course of custody studies, evaluators commonly observe situations that they believe could be improved if the parties acted in some particular fashion. The evaluator may conclude that a particular visitation schedule would be optimal for the child or that one or more of the parties would benefit from therapy; the evaluator may conclude that peace could be maintained between the parents by the introduction of a third party who would supplement the custodial parent’s caretaking function, thereby making up for observed limitations in that parent. Whatever the intrinsic merits of these recommendations, they are of no value unless they can be implemented.

Parents may voluntarily adopt the recommendation for the child’s benefit, although the process of convincing them to do so may involve the evaluator in problems of playing dual roles in the divorce process. It benefits the child little for parents to agree to terms in the midst of litigation unless the parents are likely to abide by those terms once the litigation is complete. Evaluators commonly confuse the court’s role in custody matters with its role in child abuse and neglect cases. In the latter situation, the court or those acting with its implicit backing freely exercise considerable power over parents who have been shown to be inadequate in their parental functioning. The court can and does require parents to participate in arrangements for the child’s benefit and through its agents often polices the parents’ compliance. In contrast, in custody disputes in divorce there is no finding of parental incompetence, and within very broad limits the court is likely to endorse a settlement reached by the parents whether or not it is optimal for the child (Weyrauch, Katz, & Olsen, 1994). Furthermore, whatever the court may order or whatever may be recommended to the parents, the court will not police compliance with its order unless the other parent actively complains of failure to comply. In making custody recommendations, evaluators should not accept parents’ reports of their intention to behave differently than they currently are
behaving because such promises are generally not enforceable. Although sometimes the outcome is positive, parents desiring custody often will describe plans that they are unlikely to carry out once an order is in place.

An arrangement workable for both parents is much more likely to be followed than a plan that is complex, crafted to absolutely optimize some aspect of the child’s situation, or is scrupulously fair to each parent. Part of the custody evaluator’s work is to try to discover workable arrangements that minimize conflict between the parents and meet the child’s needs.

**BIAS AND ITS SOURCES**

Even though legal and mental health professionals have long been aware that examiner bias can result in problematic evaluation, an appreciation of the extent to which bias, usually bias that is outside of awareness, can impact evaluations is only now becoming clear. Over the past two decades studies of human decision processes, including the decisions of professionals, have revealed a range of cognitive factors that lead to “predictable irrationality” (Arieli, 2008) and shown that cognitive and social bias can profoundly impact critical decisions (Groopman, 2007).

Social psychologists distinguish prejudice (an attitude) from discrimination (a pattern of behaviors that is often motivated by prejudice but may not be). One of the most important forms of bias in custody evaluations is the tendency to stick to a conclusion irrationally. Whether conscious or unconscious, confirmatory bias (like prejudice) can be difficult to identify and, so, difficult to address. Martindale (2005) distinguishes the internal processes referred to as confirmatory bias and a more readily detectable pattern of professional behavior that can effectively be addressed in the cross-examination of an expert.

*Confirmatory distortion* refers to the process by which an evaluator, motivated to bolster a favored hypothesis, intentionally selects or skews interpretation of data. The way evaluators are selected may contribute to such distortion. Whether evaluators are selected by judges, litigants’ attorneys, or a panel of evaluators, they know that offering opinions that run counter to their selectors’ views may greatly reduce the likelihood of their being retained again, and so they feel pressured to provide evaluations that are consistent with the views of those who retain them. The resulting distortions of the evaluation process may range from truncating the evaluation meet a judge’s demand that litigants not be forced to pay too much to outright dishonesty about what a litigant said in an effort to help the litigant prevail. The recognition that retention in future cases is contingent on how pleased judges and attorneys are with the evaluations and that the degree to which the legal professionals are pleased with the work often depends on how
well it conforms to their own view of the case puts enormous pressure on evaluators. Attempts to avoid evaluator bias by choosing a presumably neutral court-appointed expert in many cases simply shifts the situation from one in which the evaluator is pressed to please the court rather than the litigants’ attorneys. This is particularly problematic in the area of child custody evaluation, in which judges are often unaware of the bias they introduce through personal convictions and commonsense psychology that runs counter to research findings.

Bias can also come from conflation of personal values with scientific knowledge. Especially among professionals who have decided to dedicate themselves to these issues, feelings tend to run high about how children should be raised, particularly in relationship with parents. Becoming aware of these biases and differentiating personal attitudes from scientifically based opinions is a well understood part of the work of mental health experts, and the extent to which the expert has done this work is an appropriate area of the court’s scrutiny. Emotional attitudes ranging from visions of the ideal family to personal religious convictions to racism, homophobia, and gender prejudices are extremely common in our society. Few custody evaluators or legal professionals are entirely free of attitudes that potentially threaten the objectivity of their work. However, although it is not a reasonable expectation that evaluators will be entirely free of such feelings, it is reasonable to expect that they are aware of them, understand how they might inappropriately interfere with the objectivity of evaluations, and take appropriate steps to avoid such interference.

In the field of child custody, where knowledge is changing as the result of new research, an additional form of bias may lie in an inappropriate commitment to the ideas of revered teachers. When an expert relies too strongly on what was learned in graduate school, sometimes decades ago, the information provided to the court may be out of date but not recognized as such. For example, some of the most coherent, persuasive, and authoritative writing about child custody during the 1970s focused on the idea of protecting the child’s relationship with the “primary psychological parent” at the expense of almost all other concerns (Goldstein et al., 1975), an idea that has little empirical support. Yet a few evaluators hold on to it because of the enormous prestige it once enjoyed among their teachers.

Similar problems may arise because a particular theory has become popular. Over the past three decades we have seen the pendulum swing in divorce courts from inattention to issues of sexual abuse, to a view that any allegation of abuse must be true, to the notion that such allegations frequently arise as the result of “parental alienation” and more recently to the belief that the child must be protected from the supposedly alienating parent. Vigorous presentations of various points of view on such matters
by experts in the field can lead evaluators and legal professionals to enthusiastically endorse ideas independent of their scientific merit.

Dramatic and particularly memorable cases are another source of bias for both mental health and legal professionals. This is particularly true if they have bad outcomes. For example, in 1993 a 3-year-old child was returned to his psychotic mother by a Chicago judge. The mother subsequently killed the child. There was extensive newspaper coverage of the case. The tragedy resulted in the Chicago courts being much more cautious about placing children with psychiatrically ill parents, even though the actual risk to children in this situation had not changed. Many mental health and legal professionals bring memories of particularly vivid cases to ongoing evaluations, and these cases are likely to have an undue impact on their consideration of a current situation.

A related effect is that recent experiences and information tend to be given undue weight in decisions. The most recent article an evaluator has read on a topic or the most recent argument a judge has heard is likely to produce bias in the direction of the new information that is disproportionate to its intrinsic merit.

Bias that is based on cognitive distortions is more likely to be problematic in custody evaluations than conscious prejudices because the latter are better known, relatively easily addressed by the court, and understood to be a source of potential difficulty by most custody evaluators. Courts and evaluators are less familiar with cognitive distortions. There are no empirical studies on the frequency of various forms cognitive distortion or their impact in custody evaluations or decisions. Anecdotal information from work product reviewers suggests that confirmatory bias, an inclination of evaluators to seek information that will confirm an initial hypothesis and a disinclination to seek disconfirming information, is common. Several well-documented psychological dynamics function either to elicit or to support confirmatory bias (Beattie & Baron, 1988; Martindale, 2004; Skov & Sherman, 1986; Snyder & Swann, 1978).

**Cognitive Dissonance**

Information that supports one’s current impressions is gratifying. Information that calls them into question generates discomfort (Brehm & Cohen, 1962; Elliot & Devine, 1994; Festinger, 1957; Festinger & Carlsmith, 1959; Festinger, Riecken, & Schachter, 1956; Steele, Spencer, & Lynch, 1993; Wicklund & Brehm, 1976). The extent to which individuals will distort their perception of reality and make further investments in positions that have been disproved to avoid admitting to themselves that they have made a significant mistake has been shown to be truly astounding. Festinger’s initial study, for example,
involved a cult whose members had been urged to give up their worldly goods in anticipation of the predicted end of the world on a certain date. The uneventful passing of the date not only failed to disabuse cult members of their beliefs (a new date was set) but led to further commitment to the cult roughly in proportion to the amount they had already given away. Evaluators are, of course, subject to these psychological forces and like the cult members are likely to pour resources into maintaining a belief in proportion to the resources they have already committed to that belief. Evaluators who find themselves building a case, putting effort into proving a point, or simply putting disproportionate time into investigating a particular aspect of a situation should wonder if cognitive dissonance is in action.

An important source of potential cognitive dissonance arises in custody evaluations when the examiner is asked to give a preliminary report on some aspect of the case. This sets up a situation in which the examiner will be pressed by cognitive dissonance to conform the ultimate report to the preliminary report and so reduces the examiner’s objectivity.

**Suggestibility**

Another reason that evaluators tend to confirm their own hypotheses is that once an idea takes hold people tend to better remember data that support it than data that disconfirm it. They also, without realizing they are doing so, tend to collect confirming data.

In a study by Strohmer, Shivy, and Chiodo (1990), counselors were asked to consider a particular clinical hypothesis and, in doing so, to reflect on a narrative report they had read a week earlier. The counselors recalled more confirmatory information than disconfirmatory information even though the report contained more disconfirmatory information. In a second study by the same investigators, counselors were asked to consider a particular clinical hypothesis and, using a narrative report available to them as they contemplated the hypothesis, to list “confirmatory and disconfirmatory pieces of information” (p. 469). More confirmatory than disconfirmatory pieces of information were listed even though the narrative report contained more pieces of disconfirmatory information. It should be emphasized that the hypothesis was offered by the researchers (not a peer, supervisor, or more experienced colleague), so the subjects were not merely complying with authority. Havercamp (1993, p. 313) found that when hypotheses are self-generated there is a significant risk, regardless of the professional experience of the subjects, that “information relevant to alternative hypotheses will not be elicited.” Her findings differed from Strohmer et al.’s in that she did not find evidence of confirmatory bias when hypotheses were client-initiated.
In their discussion of confirmatory bias, Klayman and Ha (1987) have hypothesized a “positive test strategy”—a tendency to pose questions the answers to which are more likely to yield confirmatory than disconfirmatory responses. An example: Parent A alleges that Parent B is an “angry person.” With this information in hand the custody evaluator asks Parent B about anger. Hypothetical situations are presented in which anger would be more likely to occur than not. The questions are posed in a way that does not invite a distinction between the internal emotional anger and the overt expression of anger. Furthermore, the questions are about situations whose connection to parenting is tenuous. In other words, the questions are framed in a manner that is likely to support the conclusion that Parent B is indeed an “angry person.”

As our example suggests, follow-up questions designed to elicit more information about matters that have come up in the evaluation are points of particular vulnerability in which suggestion is likely to come into play. Close attention to how follow-up questions are framed often provides information about this form of bias. It is a good practice, except in instances where it is patently inappropriate, to ask both parties identical follow-up questions even though the question may have initially arisen with regard to only one party. For example, if Parent C alleges that Parent D is a heavy drinker, though it is certainly wise to get a thorough history of alcohol intake by Parent D, it is also wise to get the same information about Parent C using identical language unless it is patently clear that such questions will provide no new information.

The notion that adults are susceptible to suggestion is often met with skepticism. The decision-making power that we give to others is best understood by examining the dynamic in its extreme form: obedience. Milgram’s (1963) classic, and terrifying, study of obedience showed that many students at an elite university would comply with researchers’ orders to apply what they believed to be a potentially lethal electric shock to their fellow study subjects. Even when Milgram moved his study from the prestigious setting of the Yale campus to a dilapidated warehouse in the industrial section of Bridgeport, Connecticut, and when the authority figure firmly reciting verbal prods was an assistant whose only symbol of authority was a white lab coat, obedience remained high. Milgram’s work on the extreme power of obedience to authority has been more than confirmed by almost 50 years of additional investigation (Zimbardo, 2007). Custody evaluations can easily devolve into something close to police interrogations, and evaluators should be aware that if they too vigorously press for desired information (i.e., information that supports their hypotheses) they are at risk of eliciting false confessions (Gudjonsson, 2003).
The suggestion to which an evaluator responds need not come from an authority figure. Likeable litigants, credible-seeming litigants, and even the litigant who has most recently been seen by the evaluator may influence the evaluator through suggestion. An example: Mrs. A. is the first to be seen by the evaluator, and during the initial session she expresses concern that her wealthy husband “may have been coached for this.” Early in the initial session with the husband, the evaluator poses several hypothetical questions to which the husband provided well-reasoned responses. A notation appears in the margin of the evaluator’s contemporaneously taken notes: “Coached!!!” In all likelihood, whatever additional information the husband provided was assessed in light of the supposed coaching. Responses that would ordinarily be seen as desirable were interpreted as evidence against the husband’s suitability as a parent because they were taken as evidence that he was coached.

Anchoring refers to a perceptual and cognitive dynamic in which information that may be neither pertinent nor true is treated as salient. Information that is perceptually and cognitively prominent functions as a reference point (an anchor) and is used in processing other information. A common form of anchoring is to place an expensive item on a sales floor, thus creating an anchor price, and to offer other lower priced items near it so that the lower priced items appear to be inexpensive. The shift in attention caused by an anchor is called “priming” (Arkes, 1981). In a study conducted by Chapman and Johnson (1999), subjects considering the features of various apartments were more attentive to positive features when they had first been told that the rental fee was high and were more attentive to negative features when they had been informed that the rental fee was low. This finding is consistent with cognitive dissonance theory. We like things to make sense. It makes sense that an expensive apartment has many positive features and few negative features. It makes sense that an inexpensive apartment has many negative features and few positive features. If two pieces of a puzzle fail to fit but one of those pieces is malleable, it is that piece that we will endeavor to manipulate. The price of each apartment is a given; perceptions of the positive and negative features of each apartment, however, can be shaped to fit what is given. Through the operation of selective attention, perceptions of the apartments are constrained by the anchor: the rental fee. Northcraft and Neale (1987) conducted a similar study; they included real estate appraisers among their subjects and asked that a house with a predetermined listing price (the anchor) be appraised. Even the professional appraisers were affected in their judgments by the stated listing price. In some jurisdictions custody matters involving never-married couples are heard in different courts from those involving divorce. The very fact of appearing in one or the other court has an anchoring effect on many evaluators and judges in that it is easy to assume
that litigants in the never-married court are all less responsible than those in the divorce court.

Judges are not immune to anchoring. A group of judges was randomly assigned to a control group or a group that was given an anchor. Both groups were given a vignette of a civil case in which a person was severely injured by a negligent driver and told they had presided over an unsuccessful settlement conference. The control group was told that the plaintiff "was intent upon collecting a significant monetary payment." The anchor group was told that the plaintiff's lawyer had demanded $10 million. Judges in both groups were asked what compensatory damages they would award the plaintiff. Judges in the control group awarded a mean of $808,000, judges in the anchor group $2,210,000. The study's authors concluded that anchoring is an important factor in judicial decision making and that who goes first may affect how judges process testimony (Guthrie, Rachlinski, & Wistrich, 2007).

Though the impact of confirmatory bias on custody evaluations has not been studied directly, even methodologically sophisticated research is pervaded by confirmatory bias (Greenwald, Pratkanis, Leippe, & Baumgartner, 1986). Greenwald and Pratkanis (1988, p. 575) conclude that when people become ego-involved, they "persevere with data-collection variations until support for the theory is obtained."

**Confidence Levels and Credibility** Evaluators whose work has been contaminated by confirmatory bias may express a high level of confidence in their own opinions and so be quite persuasive. The *Specialty Guidelines for Forensic Psychologists* states that evaluators should examine "the issue at hand from all reasonable perspectives, actively seeking information that will differentially test rival hypotheses" (Committee on Ethical Guidelines for Forensic Psychologists, 1991, p. 661). When faced with competing hypotheses, each of which is supported to some extent by the data that have been collected, the critical thinker is unable to achieve complete certainty (Koriat, Lichtenstein, & Fischhoff, 1980). When called on to support an opinion that has been offered, the expert who has never seriously contemplated more than one hypothesis engages in a confirmatory mental search process. Data supportive of the focal hypothesis readily come to mind; nonsupportive data are not encountered (Koehler, Brenner, & Griffin, 2002, p. 693).

**Primacy and Information Integration** When an opinion is formulated from conflicting information, information received earlier has more impact than information received later (Crano, 1977). Applied to custody evaluations this means that there is a natural (albeit unintentional) tendency for the first person interviewed to have more influence on the evaluator's opinions than the person interviewed second.
This suggests that an initial joint interview, where feasible, is likely to create less bias than interviewing parents in sequence. Bow and Quinnell (2001) found that among psychologist custody evaluators 31% conducted the initial session with both parents. If an evaluator meets with one parent before the other, the evaluator begins to construct a mental framework within which information is placed based on statements made by the first parent. Though first impressions are not cast in stone, they are far stronger than people are aware. Belsky and Gilovich (1999) found that if potential stock buyers are provided with a mix of information about a company, the decision to purchase or not to purchase the company’s stock is significantly influenced by the sequence in which information is presented.

**Premature Hypothesis Generation**  At what point in the evaluative process do evaluators formulate their initial hypotheses? No current survey data shed light on this issue. Three decades ago, Sandifer, Hordern, and Green (1970) reported that psychiatrists formulated their initial diagnostic hypotheses only minutes into their intake interviews. Reviews of custody evaluators’ contemporaneously taken notes suggest that evaluators occasionally record hypotheses in the early pages of their first session notes and, at times, record what can only be described as premature conclusions. An example: Examining contemporaneously taken notes, a reviewer notices “LLPF” in the margin next to some statements made by the mother during the initial joint interview. At trial, the evaluator explained that “LLPF” is an abbreviation for “Liar, liar, pants on fire.” That is, early in her first session, the evaluator concluded that the mother is a liar. (The notation is particularly striking because the father, who was present during the interview, neither refuted nor modified the statement.)

Evaluators’ early conclusions about credibility based on their supposed ability to differentiate forthright from deceitful litigants is particularly interesting in light of research showing that mental health professionals are no better at lie detecting than anyone else and that, in general, 20 years of research has found that people are simply not able to detect when someone is lying—there are no reliable verbal or nonverbal cues to deception. Individuals’ confidence in their ability to detect deception is unrelated to their actual accuracy in doing so (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997). Experience does not necessarily heighten one’s ability to detect deception (Ekman & O’Sullivan, 1991).

**The Investigative Component**  Information provided by third parties (collateral sources) and obtained through the review of records has become an integral part of custody evaluations (Austin, 2002). Ackerman and Ackerman (1997) found that approximately 28% of the evaluator’s time was spent on obtaining
information from documents, disinterested collateral sources, and other nonparties. However, standards and practices in collecting these data have yet to be well developed (Kirkland & Kirkland, 2001).

Indiscriminate investigations (“fishing expeditions”) are ill advised and possibly unethical (because they create undue risk of harm with little likelihood that pertinent information will be unearthed). An evaluator cannot, for practical reasons, assure litigants that every document described by them as being of potential importance will be reviewed and that every collateral source identified by them as having useful information will be interviewed. Though lists of documents to be reviewed and lists of collaterals to be interviewed must sometimes be trimmed, the evaluator must take great care to avoid conducting a focused search for data supportive of a favored hypothesis. Similarly, evaluators should be aware that many collateral sources are highly biased or inaccurate even if they do not have a direct interest in the case. Teachers, physicians, and therapists are like everyone else in having their observations shaped by factors that may be irrelevant to the issues at hand. The quality of information obtained from third parties can be improved by asking for specific factual information. For example, a teacher’s impression that “Mr. Jones seems to have an firm emotional relationship to his child” is far less useful than her observation that “Mr. Jones has come to each available parent-teacher conference. At the last conference he asked about three specific grades the child had received and wanted to know what he could do to help the child do better.”

Addressing Bias  Bias and the threat of bias are inevitable aspects of custody evaluations, just as they are inevitable aspects of any research or decision-making process. The presence of bias and threats of bias do not in themselves negate the value of custody evaluations. However, it is the responsibility of the custody evaluator, as it is the responsibility of the researcher or decision maker, to be as aware as possible of the impact of bias on opinions and to work to address the impact of bias on evaluations. Several means of accessing bias are described here.

Audio and Video Recording  Among other advantages, recordings of interviews provide excellent opportunities to detect and address bias. Even the best notes condense interactions. An evaluator cannot write down aspects of the interaction of which the evaluator is unaware. Not only the actual words used to phrase a question (which are usually lost unless the interview is recorded) but the evaluator’s inflection and tone carry markers of bias. The evaluator’s facial expressions and body language are often out of the examiner’s awareness and virtually never recorded in the notes. Consider the question “You never used illicit drugs, right?” Depending on the examiner’s tone, inflection, facial expression, and body movement, these words could
indicate a friendly inquiry whose anticipated answer is affirmative or a hostile, skeptical interrogation designed to elicit a negative response. Not only would these two questions have different impacts on the litigant, but they would also indicate very different biases in the evaluator. These differences would be discernible only through adequate recording.

**Contemporaneous Notes** Martindale (2001, p. 488) observes, “The search for indications of bias is most efficiently begun by comparing the contents of an evaluator’s contemporaneously taken notes with the evaluator’s description of factors supporting the opinion(s) offered.” Confirmatory bias or distortion is most easily demonstrated where there is a pattern of discrepancies between the information in contemporaneous notes and the information in the advisory report. Where parenting strengths in the nonfavored parent are described in the contemporaneous notes but not mentioned in the advisory report, either bias or outright distortion is at work. The same is true when deficiencies in the favored parent are noted but not alluded to in the report.

**Document Inspection** Many evaluators retain documents they have relied on in forming opinions. Some evaluators, however, feel no obligation to retain items that have been considered but not relied on. In law, considering something involves examining it and deliberating about it (Nolan & Nolan-Haley, 1990, p. 306). The question of whether a document was not relied on because of bias can be answered only if the documents remain available.

**Assessment Data Review** Evaluators whose neutrality has not been contaminated by bias “describe the strengths and limitations of test results and interpretations” (American Psychological Association, 2002, p. 1071), even if their profession is not psychology. Similarly, when factors come into play that might affect their “judgments or reduce the accuracy of their interpretations . . . , they indicate any significant limitations of their interpretations” (p. 1072). It is difficult to imagine a situation in which there would not be any limitations worthy of mention in a custody evaluation. The absence of such a statement suggests either that the evaluator is not knowledgeable enough about bias to address it or that an attempt is being made to persuade rather than inform. In either case it points to a problem in the management of bias by the evaluator.

**THE GOALS AND STANDARDS OF EVALUATION**

In their analysis of the best interest standard, Goldstein et al. (1975) observe that it is unrealistic to attempt to make recommendations that are optimal for the child and recommend instead that the evaluator attempt to find “the least
detrimental alternative.’’ Although some of the recommendations based on this formulation now appear mistaken and the concept has not found wide acceptance, a central problem to which they pointed remains of concern. It is generally not possible to frame a scientific opinion regarding a particular child in such a way that definitively solves all the problems associated with a custody decision. Almost always, any custody decision will have adverse as well as desirable affects.

Very often, the scientific evaluator will not be able to achieve the degree of certainty that all involved wish we had in making decisions that impact children as profoundly as custody decisions do. Far from implying that the opinions of scientific custody evaluators are useless, this means that their opinions can be relied on to be accurate and sound as far as they go. Courts and others who rely on these opinions may be less comfortable than they would be with more definitive statements, but they will at least have the advantage of accurate information on which to reach conclusions. Although we know of no studies on the matter, it is our impression that custody evaluators are rarely hired guns in the sense of deliberately making recommendations favorable to the party that has retained them. However, many experts believe it is useful to advocate for the opinion they have come to; sometimes they believe doing so is to advocate for the child. Such experts misunderstand their role in the legal process: providing the best information possible to those who must make custody decisions. The expert’s job is to communicate accurately and informatively in a way that is useful to the court and other decision makers. To obtain reliable testimony courts must tolerate the distress of decision making and the ambiguity that is part of the reality of complex human psychology. When courts displace their judicial responsibility for legal decision making onto enthusiastic experts, they create a situation in which the experts may fill this role at the expense of providing accurate information.

Increasingly in the behavioral sciences, researchers have recognized the value of framing findings in terms of specific risks and probabilities. In so-called second-generation studies of the risk of violence, investigators have tried to make quantitative predictions of the likelihood of an individual’s being violent in various circumstances, rather than attempting the essentially impossible task of saying that the person will, in fact, be violent. Following such a process, the expert’s goal should be to reach the most reliable and valid opinion possible under the circumstances and to be able to describe how reliable and valid that opinion is. Rarely can the expert accurately forecast exactly what will or will not happen. (When an expert gives an opinion in this form, it almost always means that the opinion is inaccurate, has been poorly expressed, or responds to a situation where the outcome is so obvious that an expert opinion is not needed.) Experts should, however, be able to say whether the information changes the likelihood of some outcome and in some
instances may be able to quantify that conclusion. In all instances, experts should be prepared to describe the reliability and validity of their own opinions.

RELEVANCE AND HELPFULNESS

Custody evaluations are valuable only insofar as they are admissible in court. Evaluators should be aware of the criteria used by the courts to ensure that their reports meet these criteria. Courts apply two specific tests of relevance and helpfulness in determining the admissibility of expert testimony. Most state evidence codes adopt or closely resemble the Federal Rules of Evidence (2006). Some states continue to use the Frye rule or variations on it. The Federal Rules of Evidence will form the basis of our discussion, but readers should be aware that rules may differ substantially from jurisdiction to jurisdiction and even from court to court.

Rules 401, 402, 403, and 702 of the Federal Rules of Evidence are the most important for custody evaluators. Rule 401, Definition of ‘Relevant Evidence,’ states, ‘‘Relevant evidence’ means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.’’ With special exceptions, relevant evidence is generally admissible, and irrelevant evidence is not admissible. Even when evidence is relevant, Rule 403 states, ‘‘evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.’’

Although courts seldom exclude reports for providing too much information, some evaluators include information that is much more prejudicial than relevant. For example, although sexual behavior is relevant to some psychiatric diagnoses, evaluators have been known to include lurid details of sexual conduct of litigants in their reports. Such details can be prejudicial and are sometimes given far more weight than their relevance would support. Test results, as discussed earlier, when extensively quoted may be prejudicial if they seem to attribute many psychiatric problems to a litigant even though they in fact add very little to the actual evaluation.

Rule 702 addresses testimony by experts:

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

The issue of expert testimony has been hotly debated by both the legal and psychological communities (Goodman-Delahunty, 1997; Kraus & Sales, 1999, Shuman & Sales, 1999). One issue is who can best decide what are and are not reliable scientific procedures and therefore potentially admissible evidence. Before 1923 courts simply reviewed the proposed experts’ credentials and the likelihood that their testimony would be helpful. From 1923 through 1992, they based decisions about admissibility on the General Acceptance Test (often called the Frye test) described in Frye v. U.S. (1923). In that case, a federal appellate court wrote:

Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. (p. 1014)

The phrase “the particular field in which it belongs” has, in the case of the sciences, generally been understood to be the relevant scientific community. In 1962 the court tackled an important issue for psychologists. Rather than limiting the definition of expert based on credentials such as a medical degree, the court determined that what was more important was whether an opinion based on a reliable scientific procedure would be likely to help the judge or jury. There was less emphasis on the status of the witness, that is, physician versus social worker, and more emphasis on the helpfulness of the information. The Jenkins court (Jenkins v. U.S., 1962, p. 643) held, “The test, then, is whether the opinion offered will be likely to aid the trier in the search for the truth.”

In 1993 the U.S. Supreme Court created a storm of change with its ruling in Daubert v. Merrell Dow Pharmaceuticals, Inc. In an attempt to rein in the use of what it called “junk science” in litigation, the Court provided a set of flexible criteria to use in determining the admissibility of scientific evidence. It also assigned to the trial court judge the responsibility of gatekeeping. Rather than relying on the relevant scientific community to determine what a reliable scientific procedure is, judges were to make decisions about the scientific value of a technique. These flexible guidelines have come to be known as the Daubert standard. Daubert creates a set of flexible guidelines, not a standard of specific factors to be assessed by the judge.
The Daubert Court defined scientific knowledge as follows: “The adjective ‘scientific’ implies a grounding in the methods and procedures of science. Similarly, the word ‘knowledge’ connotes more than subjective belief or unsupported speculation” (1993, p. 590). But to qualify as scientific knowledge, an inference or assertion must be derived by the scientific method. Proposed testimony must be supported by appropriate validation, that is, “good grounds,” based on what is known. In short, the requirement that an expert’s testimony pertain to scientific knowledge establishes a standard of evidentiary reliability (p. 590).

Although Daubert placed courts in the position of gatekeepers of expert testimony, it provided only a flexible set of guidelines the trial that courts could (as opposed to should) use:

Faced with a proffer of expert scientific testimony, then, the trial judge must determine at the outset, pursuant to Rule 104(a), whether the expert is proposing to testify to (1) scientific knowledge that (2) will assist the trier of fact to understand or determine a fact in issue. This entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue. We are confident that federal judges possess the capacity to undertake this review. Many factors will bear on the inquiry, and we do not presume to set out a definitive checklist or test. But some general observations are appropriate. (1993, pp. 592–593)

The following guidelines were offered by the Court to be used in a trial judge's gatekeeping efforts.

Testability or Falsifiability

Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact is whether it can be (and has been) tested. “Scientific methodology today is based on generating hypotheses and testing them to see if they can be falsified; indeed, this methodology is what distinguishes science from other fields of human inquiry” (Daubert, 1993, p. 593).

Peer Review

Another pertinent consideration is whether the theory or technique has been subjected to peer review and publication. Publication (which is but one element of peer review) is not a sine qua non of admissibility; it does not necessarily correlate with reliability, and in some instances well-grounded
but innovative theories will not have been published. Some propositions, moreover, are too particular, too new, or of too limited interest to be published. But submission to the scrutiny of the scientific community is a component of “good science,” in part because it increases the likelihood that substantive flaws in methodology will be detected (Daubert, 1993, p. 593). Publication (or lack of such publication) in a peer-reviewed journal thus will be a relevant, though not dispositive, consideration in assessing the scientific validity of a particular technique or methodology on which an opinion is premised (p. 594).

**Error Rate and Standards of Control**

In addition, in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error and the existence and maintenance of standards controlling the technique’s operation (Daubert, 1993, p. 594).

**General Acceptance**

Finally, “general acceptance” still has a bearing on the inquiry. A “reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community” (Daubert, 1993, p. 594). Widespread acceptance can be an important factor in ruling particular evidence admissible, and “a known technique which has been able to attract only minimal support within the community” may properly be viewed with skepticism (Daubert, 1993, p. 594).

The Court wrote:

> The inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission. The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate. (Daubert, 1993, pp. 594–595)

In 1997 the U.S. Supreme Court heard General Electric Co. v. Joiner and expanded the Daubert decision, noting the following:

Conclusions and methodology are not entirely distinct from one another. Trained experts commonly extrapolate from existing data. But nothing in either Daubert or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the
expert. A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered. (p. 146)

In other words, there must be something more than just the expert’s word tying the data to the opinion.

The focus of the Daubert Court was on scientific knowledge because that was the nature of the testimony offered into evidence. This is clarified in a footnote to the opinion and sets the stage for the next case in what has been called the Daubert trilogy: “Rule 702 also applies to ‘technical, or other specialized knowledge.’ Our discussion is limited to the scientific context because that is the nature of the expertise offered here” (Daubert, 1993, p. 590).

In 1999 the U.S. Supreme Court further expanded the Daubert decision to include all expert testimony, not just testimony based on scientific knowledge. In Kumho Tire Co. v. Carmichael (1999) the Court noted the following:

The Daubert “gatekeeping” obligation applies not only to “scientific” testimony, but to all expert testimony. Rule 702 does not distinguish between “scientific” knowledge and “technical” or “other specialized” knowledge, but makes clear that any such knowledge might become the subject of expert testimony. It is the Rule’s word “knowledge,” not the words (like “scientific”) that modify that word, that establishes a standard of evidentiary reliability. Daubert referred only to “scientific” knowledge because that was the nature of the expertise there at issue. (p. 138)

We conclude that Daubert's general holding—setting forth the trial judge’s general “gatekeeping” obligation—applies not only to testimony based on “scientific” knowledge, but also to testimony based on “technical” and “other specialized” knowledge. We also conclude that a trial court may consider one or more of the more specific factors that Daubert mentioned when doing so will help determine that testimony’s reliability. But, as the court stated in Daubert, the test of reliability is “flexible,” and Daubert’s list of specific factors neither necessarily nor exclusively applies to all experts or in every case. Rather, the law grants a district court the same broad latitude when it decides how to determine reliability as it enjoys in respect to its ultimate reliability determination. (p. 141)

In other words, where the evidence code says “if scientific, technical, or other specialized knowledge will assist the trier of fact,” the key word is “knowledge,” and not “scientific” (Shuman & Sales, 1999). Ultimately, regardless of the type of testimony, it is the methodology underlying an expert’s opinion that must be sound, reliable, and generally accepted, not the opinion itself. The Daubert, Joiner, and Kumho cases were clarifications of the
Federal Rules of Evidence, and therefore did not apply directly to the states. Many states, however, have adopted the \textit{Daubert} standard. Several states have continued their reliance on the \textit{Frye} test, or an expanded version of the \textit{Frye} test. California, for example, has the \textit{Kelly-Frye} test (\textit{People v. Kelly}, 1976) and Florida has the \textit{Ramirez-Frye} test (\textit{Ramirez v. State}, 1995). Both of these states have expanded the \textit{Frye} test to determine if the methodology underlying the expert’s opinion is reliable and valid and helpful to the court.

\textbf{CONCLUSION}

Custody evaluation can incorporate empirical information to reach credible conclusions. This involves three overlapping steps: the collection of information about the particular situation, the evaluation of pertinent published reports, and the informed use of published reports in reaching conclusions. The outcome, if the process is conducted properly, may not necessarily be clear or definitive, but under the circumstances it is the best information to aid those making decisions.

\textbf{NOTES}

1. The pursuit of accurate and adequately recorded information is time consuming and expensive. It also often produces more complex findings than more superficial investigations. When the evaluation becomes part of a litigation process, this means that the scientific evaluator may appear to be at a disadvantage. Such an evaluator will be more explicit about the limitations of the evaluation and will acknowledge that certain investigations were not performed (or recorded). He or she will acknowledge lack of funds or time, will engage in more extensive and therefore more expensive evaluations whose cost may aid in suggesting that the expert is a hired gun, and at the end may come to conclusions that are less black and white than a less systematic investigator. Depending on the sophistication of the trier of fact, the scientifically oriented evaluator may thus appear at a distinct disadvantage as an expert witness compared to the less adequate examiner. As the courts move more in the direction of excluding junk science and develop increasingly sophisticated appreciation of the nature of scientific thinking, these apparent weaknesses are likely to turn into strengths. Scientifically oriented experts must be prepared to explain why their methods may appear less impressive than the false certainty of less careful evaluators.

2. Much research and scientific writing results from the collaboration of a seasoned researcher with a more junior colleague. The senior investigator, in essence, certifies the quality of the junior’s research and is regarded as responsible for the result.

3. This matter is introduced here only as an example. Contemporary psychoanalysts would answer this criticism in two ways. Some would point out that the
actual theory predicts the presence of certain observable residues of the Oedipal situation so that the theory is actually falsifiable (Edelson, 1984). Others would say that psychoanalysis is not a science in the same sense as the physical sciences and makes a different kind of claim to truth based on an enriched understanding of the psychological world rather than testable propositions about it (Cohler & Galatzer-Levy, 2007).

4. Even when those who construct tests carefully describe what they do and do not measure, it is easy to misconstrue their significance. Although intelligence tests were largely constructed to predict academic performance and although they have been shown to usefully measure a range of psychological functions, they are not adequate measures of overall psychological capacity or ability to function. Yet they are commonly treated as such not only by laypeople but also by professionals, often with negative consequences (see, e.g., S. Gould, 1981).

5. The applications of psychological tests to individuals are discussed at greater length in the chapter on psychological testing (Chapter 4). It is worth noting here that the application of psychological tests that lack content validity may lead to particular problems in assessing psychological function. One of the authors of this chapter recently had the experience, in two consecutive evaluations of Minnesota Multiphasic Personality Inventory II test results suggesting that the subjects were psychotic and delusional. Their scores in these areas had been elevated because both answered affirmatively that their pictures had appeared on magazine covers. In fact, both individuals had been depicted on the covers of national magazines. A certain test-taking set or stance is more likely to distort results when subjects strive so hard to be honest that they overinterpret self-report test items, for example, endorsing, “I have difficulty falling asleep” because they occasionally—not usually or often—have difficulty doing so.

6. We use the ‘best interest’ criterion here as an example both because it is widely accepted in the United States and because it points to the complexity of behavioral science opinions. The same logic would apply to virtually any question that might be posed to the behavioral scientist in this context, even though the issue at hand might not be the child’s best interest.

7. The impression that so-called hired guns are common arises from several sources. Most evaluators who do substantial work in this area develop reputations regarding the kind of opinions they are likely to render, so that attorneys are likely to try to select evaluators who agree with their client’s interests. When an evaluator comes to a conclusion contrary to the retaining attorney’s position, the case often settles without trial or the expert is not called to testify; thus at trial it usually appears that experts always agree with those who retain them. Finally, there are unethical individuals who, for a fee, will testify to almost anything. They stick out in memory because their behavior is often so egregious.

8. In his dissent from this opinion (92–102, William Daubert et ux., etc., et al., Petitioners v. Merrell Dow Pharmaceutical Inc.) William Rehnquist observed, “I defer to no one in my confidence in federal judges; but I am at a loss to know what is meant when it is said that the scientific status of a theory depends on its ‘falsifiability,’ and I suspect some of them will be, too.” Indeed the very
language of the Supreme Court’s decision suggests that he was correct because “falsifiability” as used in studies of the philosophy of science is not something that can itself be tested, but is the property that a proposition can, in theory, be shown to be untrue by empirical means.

9. The Latin phrase ipse dixit means “he himself said it; an assertion by one whose sole authority for it is the fact that he himself said it” (Gifis, 1991, p. 252).

REFERENCES


American Psychological Association 2002.


