Assessing the accuracy of young children’s reports: Lessons from the investigation of child sexual abuse

DEBRA A. POOLE
Central Michigan University

D. STEPHEN LINDSAY
University of Wales—Bangor, Bangor, United Kingdom

Abstract

Procedures for investigating allegations of child sexual abuse have come under intense scrutiny by social critics, researchers, and the courts. Concerns about under- and overidentification have fueled two approaches to evaluation: the indicator approach, which seeks to specify symptoms that can be used to identify sexually abused children, and the assessments approach, which analyzes conditions associated with accurate versus inaccurate event reports. A review of research from these approaches reveals a number of gaps between empirical results and commonly cited aphorisms about how to discriminate between true and false reports. Four principles for designing studies and communicating findings are suggested to improve the interface between research and practice.

Key words: Sexual abuse, Children’s reports, Assessment, Evaluations, Eyewitness testimony

During the past decade, the topic of child sexual abuse became a marquee issue for debates about the gap between science and social policy. Procedures for identifying abused children garnered supporters and critics from an interdisciplinary audience that included child advocates, memory researchers, social psychologists, decision scientists, and legal, medical, and mental health professionals. In the midst of ongoing and often bitter discussions, one recurring conclusion cut across areas of specialization and political agendas: There is no scientifically validated procedure for reliably identifying children who have been sexually abused. As expressed by Geyer (1995), “The clinical assessment processes which have been used in assessing allegations of sexual abuse have regularly and ambitiously raced ahead of the experimental/empirical foundations which would warrant their use as valid, and reliable methods” (p. 15). Similarly, Goodman, Emery, and Haugaard (1998) joined numerous other researchers and practitioners when they lamented, “There is as yet no generally accepted, scientifically valid way to determine whether an abuse report is true or false” (p. 821).

This article summarizes central issues regarding the assessment of children’s event reports, with particular emphasis on children’s reports of sexual abuse. In recent years, laboratory studies of children’s event reports have co-evolved with studies of the dynamics of sexual-abuse disclosures. When an overwhelmed child-protection system renewed interest in children’s accuracy and suggestibility in the 1980s, memory researchers responded with study procedures that mimicked at least some aspects of the conditions described in transcripts of sexual-abuse investigations. In turn, concerns generated by these studies shifted the focus of research on child sexual abuse from an emphasis on qualitative descriptions of abused children and their families to a comparative approach that looked for indicators to distinguish between abused and nonabused groups.

Because concerns about child sexual abuse span a large number of interrelated issues, we need to clarify the topics selected for this review. First, our review focuses on the assessment of children’s reports of victimization and does not address adults’ delayed reports of childhood sexual abuse (see Read & Lindsay, in press, for reviews pertaining to the later issue). Second, regarding the evaluation of children’s reports, the notions of “credibility,” “reliability,” and “validity” assessment have sometimes been used narrowly, to refer only to analyses of children’s narratives about abuse, and other times broadly, to refer to the entire context of the allegation.\(^1\) We summarize data relevant to the broad

\(^1\)The term “credibility” has been used in various ways throughout the literature on event memory. Traditionally, “accuracy” and “credibility”...
meaning of assessment, which includes consideration of reasons for initial suspicions of abuse, family and social contexts surrounding the allegation, as well as behavioral characteristics of the child and the abuse report itself. This approach is consistent with research that shows that the social context (e.g., the number and type of previous conversations about possible abuse) is crucial to assessing the likelihood of an accurate account. Third, child-sexual-abuse investigations include two types of cases: Those in which a child reportedly made a verbal allegation of abuse prior to involvement in investigations, and those in which a child came to make a verbal report only after a series of interactions with adults who evaluated suspicions of abuse (e.g., therapists or investigative interviewers). Children in the latter category often are referred for evaluation based on contact with an alleged perpetrator, reported changes in their behavior, sexual knowledge, or some other presumed indicator of abuse, and only some of these children eventually offer verbal reports of abuse (Keary & Fitzpatrick, 1994). Our review is relevant to these types of cases as well as those in which children reportedly made verbal allegations prior to any investigative interviews. Finally, we do not discuss medical evidence, because such evidence is relevant for only a small proportion of abuse cases and is generally not available to parents, neighbors, or mandated reporters who make initial decisions to file a sexual-abuse referral.

The first section of this review briefly discusses why there is an increased sense of urgency to educate helping professionals about potential sources of inaccuracy in children's event reports. The second section describes two approaches to assessing accuracy—the indicator and the assessments approaches—highlighting limitations in the existing database.

The concluding section lists suggestions for designing and reporting studies that would help remedy misunderstandings about the characteristics of valid and invalid reports.

Rethinking Validity Assessment:
New Challenges to Old Assumptions

Despite the lack of scientifically validated procedures, there is no doubt that judgments about child sexual abuse will continue to be made by individuals with a wide variety of skills, educational backgrounds, purposes, and sources of information. Mandatory reporting legislation currently requires teachers, physicians, and other helping professionals to decide when evidence crosses an unspecified threshold to become a "reasonable suspicion" of abuse (Kalichman, 1993; Levine & Doueck, 1995). Reports to child-protection agencies trigger investigations that may involve social workers, police officers, and prosecuting attorneys, initiating a process that often involves numerous professional opinions before final decisions are made about child protection, custody, or criminal matters. As Berliner and Conte (1993) argued, the role of professional judgment in sexual-abuse evaluations is not fundamentally different from the role of judgment in virtually all professional activity, and it is unlikely that legal or scientific advances in the near future will eliminate dependence on such judgments.

Four factors have generated concern about how professionals make judgments about abuse status: revised estimates of false allegation rates, laboratory data on the mechanisms and characteristics of false event reports, studies of professional decision making about child-sexual-abuse allegations, and judicial decisions regarding corroborative evidence. In concert, these factors have shined a spotlight on common practices and assumptions regarding the reporting and substantiation of child sexual abuse.

Estimates of False-Allegation Rates

For many years, it was frequently claimed that about 5%–8% of sexual abuse allegations were false. This estimate appeared in published articles throughout the 1980s (Sink, 1988), was underscored by the American Professional Society on the Abuse of Children (APSAC, 1993), and continues to be a benchmark in more recent articles (e.g., Penfold, 1993). Moreover, figures around 6% have been repeatedly mentioned in court by expert witnesses (Ceci & Bruck, 1995).

Several authors have commented that this estimate of the false allegation rate is misleadingly low because of the practice of counting only intentionally false allegations (Ceci & Bruck, 1995; Robin, 1991; Wakefield & Underwager, 1991). Ceci and Bruck (1995), for example, compared descriptions of coding categories in studies of false allegation rates with the studies' final conclusions. In some cases, conclusions about false allegations reflected only cases in which there was intentional lying, with authors dismissing cases in which false allegations were made for other rea-
sons. When these other mechanisms were included (e.g., honest misunderstandings of children’s statements or problems caused by reliance on presumed indicators of abuse followed by suggestive interviewing), rates of false allegations rose considerably (e.g., from 6% to 23% in Jones & McGraw, 1987; from 8.8% to 35% in Faller, 1991). Although these higher numbers might better be termed “false suspicions” of abuse (Poole & Lamb, in press), a broader definition of false allegations is being accepted by other authors, leading to higher estimates of the proportion of invalid cases. For example, in their study of 399 children between the ages of 8 and 15 years who were evaluated at a sexual abuse crisis center, D. M. Elliott and Briere (1994) categorized 17.8% as nonabused. Although methodological challenges ensure continued controversy about false allegation rates, three trends are clear: (a) scholars are adopting a broader view of false allegations that encompasses multiple mechanisms for inaccurate suspicions of abuse; (b) rates of invalid suspicions are probably higher than original estimates of about 6%, perhaps markedly so; and (c) these revised estimates have substantially increased interest in children’s suggestibility and the characteristics of accurate versus inaccurate reports.

**Inducing False Reports of Meaningful Events**

Why did earlier estimates of false allegations dismiss reports stemming from error or misinterpretation? One contributing factor might have been the fact that there was little evidence at the time that children could easily be induced to make false reports of meaningful events. In other words, even if adults did sometimes misconstrue children’s behavior or jump prematurely to conclusions about sexual abuse, interviewing children may have been viewed as a simple check against such errors. Nonmalicious false reports cause alarm only if they are not readily distinguishable from true reports. In recent years, however, there have been a number of compelling demonstrations that children do sometimes falsely report events that involve bodily touch or harm, that such reports are alarmingly easy to produce—at least at young ages—and that these reports often contain many of the characteristics considered to be “diagnostic” of true reports. This section does not offer a comprehensive review of research on the effects of misleading influences on children’s reports, but rather highlights a few recent studies that have contributed to the sense of urgency regarding the need for improved methods of discriminating between accurate and inaccurate allegations.

The most widely cited studies of children’s suggestibility are those conducted by Stephen Ceci, Maggie Bruck, and their colleagues (for reviews, see Bruck, Hembrooke, & Ceci, in press; Ceci, 1995; Ceci & Bruck, 1995). In one study, 96 preschool children were interviewed 7–10 times about events in their lives, with interviews spaced approximately 7–10 days apart (Ceci, Huffman, Smith, & Loftus, 1994). In each interview, the experimenter asked the child about some events that had really happened (determined by interviewing parents) and some events that had not. (This study came to be called the “mousetrap” study, because one of the fictitious events was that the child got a finger caught in a mousetrap and had to go to the hospital.) Interviewers told the children the following:

I am going to read some things that may have happened to you, and I want you to think real hard about each one of them that I am going to read. Try to remember if it really happened. We made this list up by talking to your mother and father to get them to tell us about things that really happened to you when you were younger, but not all of the things that I am going to read to you really happened.

Ceci, Huffman et al. predicted that few children would assent to the fictitious events in the first session, but that some would come to report them over repeated interviews. This was not the case. Instead, the children assented to 34% of the fictitious events in the first and seventh sessions, with some children assenting or denying consistently and a few flip-flopping between assenting and denying. (Stability coefficients for adjacent sessions ranged between .62 and .87.) A second study used a similar methodology with one exception: The children were told that the fictitious events had actually happened (Ceci, Loftus, Leichman, & Bruck, 1994). With this modification, erroneous reports increased over a 12-week period from an initial 34% assent rate to a final 45% assent rate among 3- and 4-year-olds, and from 25% to 40% among 5- and 6-year-olds. In both studies, the children not only indicated that the fictitious events occurred, they also described them with details, sometimes with emotional reactions and spontaneous elaborations. Other studies have also shown that children sometimes misreport bodily harm in convincing ways. For example, Bruck, Ceci, and their colleagues have documented that some preschoolers falsely report genital touch when interviewed suggestively with anatomical dolls (Bruck, Ceci, Francoeur, & Renick, 1995), and that some 6-year-olds attribute a painful medical procedure (an injection) to the wrong individual after exposure to delayed interviews with suggestive questions (Bruck, Ceci, Francoeur, & Barr, 1995). Of course, in these studies many of the children’s reports were accurate; nonetheless, the rate of false reports was sufficiently high to warrant concern.

Although the studies by Ceci and Bruck are disturbing, child-abuse experts sometimes dismiss them as irrelevant to

---

2There is disagreement about how to refer to the process of judging whether abuse did or did not occur. Myers (1993) pointed out that some professionals object to the term “diagnosis” because abuse is an event, not a disease or disorder. However, we agree with Myers et al. (1989) that “It would be a mistake to make too much of disagreement over applications of diagnostic terminolgy to child abuse. The concept of diagnosis is sufficiently broad to embrace clinical interpretations regarding sexual abuse. By way of analogy, battered child syndrome is an accepted medical diagnosis despite the fact that an abuse assault is an event rather than a disorder” (pp. 72–73).
the majority of sexual-abuse evaluations. One reason is that most of these studies emphasized heightened suggestibility in preschool children, but the majority of sex-abuse investigations involve school-aged children. Furthermore, errors by older children, such as the 6-year-olds in Bruck, Ceci, Francoeur, & Barr, et al. (1995), were elicited with highly suggestive questioning. During presentations of these studies, we have heard members of the audience express annoyance that suggestibility researchers “stack the deck” by choosing very young children and/or by using highly suggestive interview procedures. False event reports are not limited either to preschool children or to coercive interviewing practices, however, as a pair of our experiments recently demonstrated.

One aim of our studies was to increase parallels between research procedures and real-world forensic interviews. In most studies of children’s susceptibility to misleading suggestions, exposure to misinformation occurs in the context of the interview situation (e.g., in the form of suggestions given shortly before the interview or misleading questions during the interview itself). Children often resist suggestions in such studies, especially when they concern bodily touch and when performance is measured in terms of recall as opposed to yes/no responses, and such findings have been taken as evidence that it is difficult to induce children to give free reports of nonexperienced events. In actual cases, however, investigative interviews are typically conducted only after the child has had numerous opportunities to be exposed to leading and potentially misleading information about the suspected events. When allegations arise during the context of divorce and custody issues, it is particularly likely that children have been exposed to prior questioning or overheard conversations by at least one adult who suspects that abuse occurred. In such cases, the issue shifts from whether children will acquiesce to leading questions about nonexperienced events to whether children will discriminate in their reports between events they actually experienced versus events they learned about through conversations. Similarly, when suspicions are based on unusual sexual knowledge, the issue might be whether children can distinguish between events they actually experienced versus ones they merely saw on television.

Children need two different skills to respond accurately to questioning about events they have heard discussed: They must have the requisite source-monitoring ability (i.e., the ability to identify the sources of their knowledge about a sexual act) and they must realize that the interviewer wants them to report only from personal experience. Children could report nonexperienced events because of deficiencies in either or both of these skills. To study the development of these abilities, we expanded the usual suggestibility paradigm by asking how children would react to various types of interview questions when they had previously been exposed to narratives that described complete mini-events that were never experienced. We attempted to reproduce some of the contexts that might be involved when there are claims that a parent influenced a child to report abuse, including exposure to experienced events that were interactive (thereby providing a context into which to embed nonexperienced events), misleading suggestions by parents in the children’s homes, and long delays between events and interviews.

In our initial pilot study (Poole & Lindsay, 1995), 17 3- to 4-year-olds individually participated in a session that included four science demonstrations conducted by a man we called “Mr. Science.” Immediately afterward, each child was interviewed by a female experimenter who used only open-ended, nonsuggestive questions (e.g., “Tell me everything you can about what happened in the science room . . . ”). Children performed very well in these initial interviews, providing a good deal of accurate information and relatively little inaccurate information. Approximately 3 months later, we mailed a storybook to parents that described two demonstrations their child had experienced and two nonexperienced demonstrations. Additionally, the story described an instance of nonexperienced touching in which Mr. Science put something yucky in the child’s mouth; in fact, nothing like the described touching had occurred. After parents read the story three times, each child was again interviewed by a new female experimenter who first asked open-ended questions, followed by leading questions and, finally, source-monitoring questions that asked the children to identify whether each event was actually experienced or was only mentioned in the story. Note that there was no overt deception in this study: The children were never explicitly told that the fictitious events had actually occurred and, although each child’s name was used in the storybook, parents were instructed to make no effort to persuade their children that the events described in the story had happened.

Our results were surprising. Contrary to reports that free recall is highly accurate, 41% of the children reported during free recall that Mr. Science had done things that were only mentioned in the story. Leading questions dramatically increased rates of false reporting. For example, 53% of the children said “yes” to a question about whether Mr. Science had put something yucky in their mouths, and 59% described this event when asked to do so. (The latter percentage is higher because one child who initially responded “no” nonetheless described the suggested event.) The children’s accuracy did not improve when interviewers reminded them of the story and asked them to indicate which events had actually happened and which were only in the story. In fact, continuing questioning seemed to exacerbate the tendency to acquiesce. For example, 71% of the children erroneously answered “yes” during the source-monitoring procedure when asked if Mr. Science really put something yucky in their mouths.

Expecting that the high rate of error in this pilot study was a phenomenon confined to preschoolers, we conducted a second “Mr. Science” study (Poole & Lindsay, 1996) to
address three primary questions: (a) What are the developmental trends between 3 and 8 years of age in children’s tendency to report events they only heard described? (b) What is the stability of true and false reports across a 1-month delay with no further exposure to the misleading story? (c) Would children’s accuracy be improved by a source-monitoring procedure that explicitly asked them to report the source or sources of their knowledge? Because we expected that the frequency of false reports would drop off dramatically after 5 years of age, we selected age 8 as an upper-bound limit on false reports from our simple, non-coercive misinformation procedure. We were wrong. During the second interview (1 day after final exposure to the story), the 114 children related a total of 58 nonexperienced events in free recall, including 17 reports of nonexperienced, suggested touching. Furthermore, there was no linear age trend in the frequency of false reports during free recall: The older children reported as many nonexperienced events as the younger children, perhaps because they were more verbal and eager to demonstrate their knowledge. As in the pilot study, false reports increased dramatically during the leading-question phase of the interview (e.g., in Session 2, between 33%–42% of the children in each age group falsely responded “yes” to questions about whether Mr. Science had put something yucky in their mouths or hurt their tummies), although the older children resisted the misleading questions more often than the younger children. False reports were less stable over time than accurate reports, but they did show some stability across a 1-month delay (e.g., of children who falsely assented in Session 2 to science demonstrations that were only mentioned in the story, 66% also assented in Session 3). Finally, the source-monitoring procedure enabled the older children, but not the younger children, to reduce the rate at which they reported having experienced suggested events.

The Mr. Science studies demonstrate that children’s reports sometimes include material from events they personally experienced intermixed with material they gained from other sources. This finding is consistent with studies that have addressed the two processes we believe contribute to these types of reporting errors: studies of developmental trends in memory source monitoring, and studies of children’s developing understanding of the language and logic of interview questions. Regarding source monitoring, there is substantial evidence that young children often perform poorly on tasks that require them to differentiate between memories from different sources, even when performance on old/new recognition is invariant across age (Foley, Harris, & Hermann, 1994; Gopnik & Graf, 1988; Lindsay, Gonzales, & Eso, 1995). Schacter, Kagan, and Leichtman (1995) argued that the special difficulties experienced by 3- and 4-year-olds may result from immature development of the frontal lobe.

Problems encoding or remembering the source is only part of the story, however. In our study, for example, the older children sometimes reported events they merely heard described, only to indicate during the source-monitoring procedure that the events had not really happened. Other investigators have also found that children sometimes report erroneous suggestions even when they do not suffer from any memory impairment (e.g., Cassel, Roebers, & Bjorklund, 1996; Marche & Howe, 1995). One explanation for this behavior is that children often construe the interview situation as calling for them to demonstrate knowledge, without attention to where that knowledge originated. For example, studies by Newcombe and Siegal (1996) showed that young children more often incorporate misleading information into event narratives when interviewers ask general questions (e.g., “Do you remember how Karen was sick when you heard the story?”) in contrast to questions that mark a specific event (e.g., “Do you remember how Karen was sick when you heard her story for the first time?”). Based on these data, Newcombe and Siegal (1996) argued that one component of suggestibility is children’s lack of understanding about the intent of questioning.

To sum up, studies such as those described previously indicate that a variety of factors can compromise the accuracy of children’s reports. In some cases, merely asking young children about nonexperienced events is sufficient to lead some of them to provide narrative accounts. Furthermore, some children report nonexperienced incidents of bodily touch even in response to nonleading, free-recall probes when they have previously heard false events described by a significant adult. The frequency of false reports can be increased in a variety of ways—by exposing children to negative stereotypes about an individual and then asking them stereotype-consistent specific questions (Leichtman & Ceci, 1995), by establishing an atmosphere of accusation (Lepore & Sesco, 1994), by asking them to think repeatedly about nonexperienced events (Ceci, Huffman, et al., 1994), or by asking them explicitly leading questions (Poole & Lindsay, 1996). The limited amount of data currently available indicate that false reports induced with these procedures sometimes drop out of children’s narratives over time (without further exposure to misleading influences) at a faster rate than reports of experienced events, but false reports do show some stability even after delays as long as 1 year (see Brainerd & Poole, 1997, for a review).

Caution is required in generalizing from such studies to forensic situations, particularly with regard to the absolute rate of false reports. For example, in our Mr. Science studies substantial percentages of children volunteered false reports consistent with a suggestion that Mr. Science had touched them in a mildly and inadvertently unpleasant way; it is likely that, all else being equal, the rate of false reports would have been substantially lower if the suggestions had described more invasive and/or script-violating forms of touching (Pezdek, 1997). Conversely, some forensic cases involve much longer delays and much more powerful suggestive influences than the Mr. Science studies. Thus our
point is not that studies like those previously reviewed provide an index of the rate of false reports in forensic cases, but rather that these studies challenge the belief that children cannot be led to make false reports of personally significant or distinctive events.

Limitations of Professional Judgment in Sexual-Abuse Evaluations

There is no evidence that fictitious reports encouraged by various suggestibility manipulations are easily distinguished from true reports. For example, when Ceci, Loftus et al. (1994) showed 10 videos of children who participated in their “mousetrap” procedure to 12 clinical and research psychologists who specialize in interviewing children, the professionals were no better than chance at distinguishing between the true and false descriptions. Of course, these results may not generalize to abuse evaluations. In actual investigations, professionals rely on a large body of evidence when judging the validity of a report, including various behavioral indicators of abuse and the social context of the allegation. Studies of professional judgment about abuse cases show some degree of agreement about which indicators are useful, but little overall agreement in final decisions. These studies have concluded that professionals are widely variable in their judgments about individual children’s abuse status, that judgments about the probability of abuse may not correlate highly with final recommendations, and that professional experience is not strongly related to performance on tasks that require determinations of children’s abuse status.

In a study by Horner, Guyer, and Kalter (1993a), 48 mental health specialists heard a case presentation lasting approximately 2 hours that included parent interviews, child interviews, child—parent observation sessions, and opportunities to request additional details from the presenter. Participants estimated the probability that the child was abused and made one of six possible recommendations about court-ordered future child—father contact. The participants then engaged in a group discussion of the case for about 1½ hours and individually submitted second estimated probabilities of abuse and recommendations for future contact. Replicating another study (Horner, Guyer, & Kalter, 1993b), the range of probability estimates was extreme, and estimates were not always correlated with recommendations. For example, the social workers who submitted father—child contact recommendations ranged between .10 and .90 in their first estimate of the probability of abuse, yet all recommended that the child’s contacts with the father be supervised.

Variability in professional opinion is characteristic of these types of studies. When doctoral-level pediatric psychologists rated the probability of abuse in vignettes, respondents suspected abuse more often for vignettes that described stronger correlates of abuse (as determined by existing research), but judgments were variable, particularly when abuse evidence was weak (Finlayson & Koocher, 1991). One vignette, for example, described a child who made no specific statement about abuse but who exhibited a decline in academic performance, had nocturnal enuresis and separation anxiety, and who made statements about a bad man. The percentage of respondents who placed the probability of abuse in each of four categories was as follows: 0–24% probability of abuse (35%), 25–49% probability of abuse (27%), 50–75% probability of abuse (29%), and 75–100% probability of abuse (9%). This is a worrisome range of opinions. Similarly, clinician’s ratings for each of 16 vignettes in a study by Jackson and Nuttall (1993) ranged from 1 (very certain that sexual abuse occurred) to 5 (very certain that it had not occurred).

Other approaches to measuring professional consensus involve looking at the indicators professionals believe to be correlated with abuse status, or the confidence professionals have in various evaluation methods. The typical finding is relatively high agreement for a small number of abuse indicators or methods of evaluation, with enormous variability in responses for the majority of items rated. For example, Conte, Sorenson, Fogarty, and Rosa (1991) obtained responses from 212 professionals who spent more than half of their practice time on average in efforts to assess or substantiate sexual abuse. Among other tasks, participants rated the extent to which each of 41 physical, behavioral, and emotional indicators was important in substantiating an abuse report. Seven criteria were rated as important by over 90% of participants (physical indicators, age-inappropriate sexual knowledge, child’s abuse report consistent over time, sexualized play during interview, child’s abuse description relates pressure or coercion, precocious or seductive behavior, excessive masturbation). The percentage of endorsements, however, ranged between 30%–90% for the remaining indicators (e.g., family history of sexual abuse—82%, perpetrator angry at child report—30%). This lack of agreement mirrors results by Kendall-Tackett (1992a) on professional standards for normal behavior with anatomical dolls. Agreement in this study was good for highly sexualized behavior, but poor for behaviors such as touching the dolls’ anal or genital areas. A surprising range of opinions regarding methods for assessing credibility was also identified by Oberlander (1995) in a survey of 31 sexual-abuse evaluators who averaged 15 years postgraduate clinical experience. Respondents disagreed about whether standardized assessment techniques contribute to judgments of credibility (28% said they were useful, 48% said they were not useful), the value of projective assessments (55% said they were useful, 19% said they were not useful), and the use of anatomical dolls (46% said they were useful, 25% said they were not useful).

Studies of the sort just described, which found substantial disagreement among professionals regarding the techniques and indicators that should or should not be used, help explain why judgments about individual cases are so variable. Although professional consensus would not have proven
that decisions were correct (as Coulborn-Faller & Corwin, [1995, p. 74], succinctly stated, “Everyone can be in agreement, and everyone can be wrong”), lack of consensus clearly indicates that, at minimum, a significant minority of decisions must be wrong. Furthermore, variability in judgments does not seem to be an artifact of analyzing the judgments of professionals with insufficient experience. J. M. Wood and Wright (1995) found that professionals were more accurate than students at estimating the probability of abuse from information about sexual behavior, but years of experience in the field of child abuse and number of sexual-abuse cases did not correlate with performance within the professional group. Similarly, Horner et al. (1993a) found no correlation between estimates of the probability of abuse and clinical experience.

Judicial Decisions Regarding the Admissibility of Expert Testimony

Mental health professionals often provide corroborative evidence in child-protection hearings, custody disputes, and criminal trials. As reviewed by Levine and Battistoni (1991), there have been two types of expert testimony:

First, the mental health professional may testify about what is called the “sexually abused child syndrome.” The expert witness may discuss behaviors identified as indicators of this syndrome, that he or she has observed in the child, or which have been reported to the expert in the course of taking a history. The implication of the testimony is that the child was indeed sexually abused. Second, the expert may testify on aspects of the “sexually abused child syndrome” without making specific reference to the particular child in question. The thrust of the testimony is to explain aspects of the child’s behavior which may appear puzzling or inconsistent, such as delay in reporting the alleged abuse. (p. 9)

Legal challenges to the first type of testimony—use of behavioral indicators or syndromes as “validation” evidence—began almost as soon as the indicators were published. The most frequently debated example is Summit’s Child Sexual Abuse Accommodation Syndrome (CSAAS), a clinical description of sexually abused children that appeared in 1983. As Summit (1992) explained in an article entitled “The Abuse of the Child Sexual Abuse Accommodation Syndrome,” the CSAAS was compiled to address the fact that attorneys tended to “discredit delayed and inconsistent reports, insisting that any legitimate victim would have made an immediate and convincing complaint” (p. 154). The description contained five “categories,” including secrecy; helplessness; entrapment and accommodation; delayed, conflicted, and unconvincing disclosure; and retraction.

Stewart and Young (1992) presented an overview of Kentucky law with regard to the CSAAS. The first use of the CSAAS that was reviewed by an appellate court in Kentucky was Bussey v. Commonwealth (1984), for a case tried in 1983 in which a psychiatrist explained the “symptoms” of the syndrome. The Supreme Court of Kentucky ruled that the trial court erred in allowing this testimony into evidence. Stewart and Young commented, “So begins Kentucky’s legal history of the CSAAS in a series of convictions which were overturned by the Supreme Court of Kentucky, highlighting the failed attempts to use the CSAAS as a medically accepted scientific concept” (1992, p. 133).

Although many states did not mirror Kentucky’s policies regarding the admissibility of evidence concerning the CSAAS (Levine & Battistoni, 1991; Reichard, 1992), highly publicized trials nonetheless focused attention on the frequent misuse of syndrome evidence. Elias (1992) described the rapid infiltration of the CSAAS into legal battles as follows:

When the legal profession is faced with a difficult or perplexing problem, it frequently tries to find a clear (if not simple) solution to the problem. Such was the case with the Child Sexual Abuse Accommodation Syndrome. Too often, lawyers and judges who had not carefully read the article, misinterpreted its purpose and presumed that the CSAAS was a diagnostic tool which would help solve the difficult question of whether or not a child had been molested. In such instances, all involved wanted the CSAAS to be diagnostic. (pp. 169–170)

If lawyers and attorneys misjudged the purpose of the CSAAS, they did so at the shittails of mental health professionals who were eager to offer indicator evidence to validate abuse. A highly publicized example is the Kelly Michaels case, in which Eileen Treacy, an unlicensed but self-proclaimed expert on sexual abuse, “diagnosed” abuse in children on the basis of the presence of five characteristics she claimed to be common to abuse situations and parents’ responses to a checklist of 32 “indicators” of abuse (Rosenthal, 1995). The first issue on which the New Jersey Superior Court Appellate Division reversed the conviction of Michaels was the improper introduction of this expert testimony (although the court’s reversal on this issue was not accepted for review by the New Jersey Supreme Court; see Mason, 1995). Similarly, the 1993/1994 San Diego Grand Jury concluded that, “Child Sexual Abuse Syndrome (CSAS), or Child Sexual Abuse Accommodation Syndrome (CSAAS), or Post-Traumatic Stress Disorder (PTSD) or other theories utilizing behavior as a basis for proof of child sexual abuse is discredited and unacceptable” (San Diego County Grand Jury, 1994, p. 31).

The primary criticism of syndrome evidence is that such evidence is merely a description of some unknown proportion of abuse victims, with no data on whether the component characteristics of these syndromes or the symptom profiles actually discriminate between individuals who are
abused and individuals who claim to be abused, but are not. The courts have been more lenient in permitting the introduction of descriptive evidence for the second purpose, to explain seemingly incomprehensible behavior. Descriptions of the disclosure process are introduced to "rehabilitate" the witness after defense attacks the child's credibility on the grounds that the child denied charges at some point or was inconsistent in other ways. Legal experts are divided on whether such testimony should be admissible. For example, Mason (1995) reviewed three problems with admitting evidence to rehabilitate a witness that is not permissible for other purposes. First, Mason claimed that "it does not take an expert witness to explain that children may delay or recant the telling of an experience as sensitive as sexual abuse" (p. 408), and therefore, as one court determined, such expert testimony may "interfere with the jury's function to judge credibility" (Commonwealth v. Dunkle, 1992, p. 837). Second, characteristics such as delay or recantation may also be typical of fictitious accounts. Finally, Mason argued that syndrome testimony may prejudice the jury's decision making as much on rebuttal as it would have on direct examination.

Legal controversy about the value of indicators and syndrome evidence is consistent with the lack of consensus among practitioners about the appropriate use of such evidence. For example, Oberlander's (1995) survey of practitioners who specialized in child-sexual-abuse evaluations included two questions about syndrome evidence. First, evaluators were asked whether it was possible to determine whether a child's behavior and symptoms were consistent with typical responses to sexual abuse: 68% said they believed it was possible to make such a determination and 23% said it was not possible. Although making a determination of "typicality" is not the same as using a set of behaviors to validate abuse, this result suggests a belief about the consistency of children's reaction to abuse that is not endorsed by many experts on sexual abuse. Second, respondents were asked about the "utility of posttraumatic stress reaction or disorder as a framework for evaluations" (p. 482): 70% said they believed it was useful, and 10% said it was not useful. Therefore, the majority of respondents seemed unaware of the conclusions based on empirical reviews that, as Oberlander (1995) summarized, "Syndrome profiles do not account for victims who are asymptomatic [and] they do not discriminate among various clinical populations" (p. 482).

Summary

There are serious concerns about how practitioners evaluate cases of alleged child sexual abuse. Data indicate that a nontrivial percentage of abuse suspicions are ill founded, that there are several mechanisms that can lead children to make inaccurate event reports, that practitioners are widely variable in their decisions about whether abuse occurred in individual cases (particularly if evidence includes indicators that are only moderately or weakly correlated with abuse), and that many practitioners base decisions (at least partially) on information that has been dismissed by the courts as scientifically unfounded. This state of affairs has shifted the direction of research on child sexual abuse and children's event reports from a focus on description to a comparative approach that seeks evidence to distinguish between accurate and inaccurate reports. We now turn to the current status of that evidence.

Assessing the Accuracy of Young Children's Reports

In a summary of conceptual approaches to child-sexual-abuse evaluations, Berliner and Conte (1993) identified two approaches to improving the validity of professionals' judgments. The first, the indicator approach, seeks evidence that distinguishes between true and false reports. The second, the standards approach, specifies rules of conduct for professionals who complete the assessment process. This approach seeks to minimize opportunities for contaminating a child's reports by establishing best-practice standards for interviewing children. Berliner and Conte's organizational framework also provides a useful way to summarize research relevant to assessing the accuracy of children's reports: studies on the indicators of child sexual abuse and studies that have identified practices that affect the likelihood of reporting errors. Berliner and Conte used the term "standards approach" to refer to guidelines that are intended to minimize reporting errors, whereas this review concerns factors that influence judgments about the accuracy of an allegation. We will use the term "assessments approach" to refer to the process of forming an opinion about the reliability of an allegation by considering how the history of the investigation of a particular case favored either accurate or inaccurate testimony. Assessing the accuracy of children's reports about sexual abuse thus involves two complementary processes: identifying positive indicators of abuse and assessing the influence of factors that can contaminate event reports.

The following summary discusses articles selected to illustrate misunderstandings or methodological challenges that have stymied the search for valid assessment approaches, or to document well-established empirical phenomena relevant to assessing children's reports. Our goal is not to present a complete review of data on each topic, but rather to set the stage for recommendations about future directions for collecting and disseminating information on assessing the accuracy of children's reports.

The Indicator Approach

The methodological continuum. Articles claiming to report indicators of sexual abuse fall along a methodological continuum that represents stages in our understanding of the characteristics of abuse cases. Although each type of article has a place in our developing understanding of sexual abuse, there have been serious problems associated with
Assessing Children's Reports

drawing conclusions from studies that are not warranted by the studies' designs.

The first step in understanding a psychological phenomenon is often clinical description. Summit's (1983) Child Sexual Abuse Accommodation Syndrome falls into this category. Summit indicated that he was describing "the most typical female victim" (p. 180), but he did not offer estimates of the proportion of victims who match one or more of the five factors enumerated in his description of CSAAS, nor did he attempt to prove that cases that display these features are more often true instances of sexual abuse than cases that deviate from the prescribed pattern. Summit stated that his syndrome was derived from "the collective experience of dozens of sexual abuse treatment centers" (p. 190) and thus represented his impression of clinical consensus. Limitations of such "consensus" profiles are well known. For example, the tendency to better remember instances that are viewed as representative of a particular category can lead clinicians to overestimate the typicality of specific indicators. Similarly, the tendency to remember instances in which salient variables cooccur can lead to illusory correlations, or unwarranted conclusions drawn on the basis of specific indicators for discriminating between groups (for reviews, see Garb, 1994; Ploos, 1993).

Although this point might seem overworked, the limitations of clinical impressions are not consistently communicated to helping professionals, partially because the literature often fails to differentiate between claims based on intuition and claims based on systematic research. For example, a 1991 article on identifying sexual abuse from children's art work that appeared in The School Counselor stated that, "Nonverbal tools, such as art projective techniques, have been clinically successful in assessing the presence of sexual abuse" (Riordan & Verdel, 1991, p. 117), a conclusion that has been vehemently attacked (e.g., Wakefield & Underwager, 1994). Riordan and Verdel (1991) listed numerous specific features of children's drawings as indicators of abuse, such as elongated necks, eyes hidden by glasses, and the presence of objects that resemble genitalia. Similarly, a doctoral student in clinical psychology recently published a comment in which she reported being trained to identify sexual abuse by "clown" responses on a Rorschach test (Date, 1996). To the best of our knowledge there is no scientific evidence supporting the validity of these clinical intuitions regarding indicators of abuse; such intuitions are valuable as a source of testable hypotheses but cannot be accepted as diagnostic indicators until they are supported by relevant research.

A second methodological step toward understanding a psychological phenomenon, also representing the descriptive approach, is to code abuse cases to quantify the proportion of cases that evince specific characteristics. A good example of this development is Faller's (1988) article on the characteristics of children's reports of abuse. Faller analyzed 103 cases of child sexual abuse in which the offenders confessed to some level of abuse. Her purpose was to identify the extent to which three clinical assumptions about the characteristics of valid reports were accurate: a description of the context of the sexual abuse, a description of the sexual behavior, and an emotional reaction to the sexual abuse. The proportion of cases judged to contain each of these features, respectively, was 78.6%, 81.6%, and 81.6%. Faller also reported that 5.8% of the cases contained no features, 10.7% contained only one, 15.5% contained two, and 68% contained all three. This study provided valuable descriptive data, but its limitations are obvious. First, the high quality of the children's descriptions could be related to the case-selection criteria (i.e., offenders may more often confess when the child's statements are highly credible). Second, the study in no way demonstrated the usefulness of these criteria for distinguishing between true and false reports of abuse, because there was no comparison group of children who claimed to be abused but were not. It is possible that accurate reports are as likely to contain these features as inaccurate reports. In the absence of a comparison group, one can only say from such studies that many abused children evince these characteristics in their reports, that some do not, and that the value of these characteristics for determining abuse status is currently unknown. This study was recently highlighted in an article on "Identifying Sexually Abused Children," however, along with another study without a comparison group and one comparative study (Coulborn-Faller & Corwin, 1995). The juxtaposition of descriptive and comparative data conveys the impression that studies with and without comparison groups yield equivalent information, which perpetuates misunderstanding about how to identify diagnostic indicators.

The next methodological step is to compare abused children to a comparison group of nonabused children, which for practical reasons has usually been clinic controls (children who are brought to clinic for some purpose other than sexual abuse) or community controls (children from the same community who preferabably are matched on relevant socioeconomic variables). For example, Slusser (1995) reviewed six studies on behavioral and emotional indicators of child sexual abuse that met her criteria of having one or more comparison groups of nonsexually abused children and a breakdown of data by developmental level. Such studies are extremely useful for dispelling myths about the value of specific indicators. For example, physical indicators such as bedwetting, headaches, or constipation have not been found to reliably discriminate between abused and nonabused groups, often because these indicators have multiple etiologies and/or high base rates of occurrence (see R. D. Wells, McCann, Adams, Voris, & Ensign, 1995, for data on physical complaints).

Despite the value of studies that systematically compared abused children and matched nonabused comparison children, such research sometimes suffers from two limitations. First, an indicator that occurs significantly more often in the
abused group can still have little or no practical value for assessing abuse reports. For example, in some cases the difference between groups is not great enough to dramatically alter one's estimate of the probability of abuse (i.e., the difference is statistically reliable but not practically significant). Similarly, an indicator might have no incremental validity once routine case features are considered, and hence cannot increase one's ability to discriminate between abused versus nonabused children. Second, there are potential problems with comparing abused children who are involved in investigations or treatment with nonabused children who are not involved in such investigations or treatment. Emotional or behavioral changes, for example, could result from the disruption caused by child-protection actions (see Berliner & Conte, 1995). If this is the case, comparing abused children to nonabused children who are not involved in suspicions of abuse would tend to exaggerate the usefulness of the behavioral or emotional indicators.

Ultimately, the optimal comparison group for diagnostic research depends on the role of the professional who is assessing sexual abuse. For example, mandated reporters such as teachers may make judgments before allegations are investigated or children are involved in sexual-abuse therapy. The most useful research for their purposes would be comparisons of the classroom behavior of all nonabused children and the preinvestigation behavior of children who, after referral, were found to have been abused. For professionals who become involved after investigations are initiated, and therefore after children may have been exposed to a variety of conversations and interviews about abuse as well as to the stresses of an investigation, the critical comparison is between true cases of abuse and cases involving false suspicions or allegations.

Many controversies about indicators of sexual abuse can be attributed to "data drift," in which results from studies with one type of methodology and purpose were used to support conclusions that could not be made on the basis of that type of data. The most egregious error is assuming that descriptions of abused children automatically constitute indicators that can identify whether or not a child has been abused. The presence of a comparison group per se, however, does not remedy all the methodological challenges of indicator research. First, criteria for identifying sexually abused groups can influence the pattern of data in important ways. For example, studying a group of sexually abused children who were predominately referred by teachers (who often refer for sexual acting out or behavioral problems), therapists with a particular theoretical orientation, or any other homogeneous group could tend to overestimate the extent to which particular indicators are associated with sexual abuse. Therefore, the most useful results are those that withstand different criteria for identifying groups. Second, because participation in an investigation or sexual-abuse therapy may alter behavior, it is important to determine whether the target behaviors predated the investigation or included behavior change during the investigation. Finally, studies have tended to collapse over broad age ranges, and secondary citations of studies rarely clarify when particular age groups were not represented in the data set. This failure to attend to age trends is especially problematic for helping professionals who evaluate preschool children, because some behaviors that are more prevalent among abused than nonabused children in broad age-range samples have a rather high rate of occurrence in younger, nonabused, populations (e.g., hugs strange adults), which could lead to overidentifying sexual abuse. Conversely, other behaviors that are given a high degree of diagnostic validity in the literature are much less likely to be present in younger children (e.g., detailed and coherent event reports), which could lead to underidentifying abuse.

J. M. Wood (1996) recently provided a useful review of issues related to weighing evidence in sexual-abuse evaluations. He reminded readers that four criteria must be met before an indicator can be used: (a) the indicator must be valid (it must appear more frequently among abused than nonabused children); (b) the strength of the indicator must be known, at least approximately; (c) the prior odds of abuse must be known; and (d) the prior odds must be taken into account when conclusions are drawn from the evidence. For example, an indicator that is three times more likely among abused than nonabused would be considered only weak evidence, whereas an indicator that is 14 times more frequent among abused children would be moderate-to-strong evidence. Practitioners might be stunned to hear that a 3:1 likelihood ratio is considered "weak" evidence, until they work through an example of combining that information with information about prior odds. For example, consider a preschool teacher who has read that showing sex parts to other children is about three times more common among abused than nonabused children (normative versus clinical frequencies per hundred were 8.1 and 24.8 in Friedrich et al., 1992). Assume that this teacher estimates that about 1 child per a class of 20 has been sexually abused, or 5% of her students. If this teacher discovers a child exposing herself or himself to another child, it is still much more likely that the child is not abused than abused (1:19 × 24.8:8.1 = 24.8:153.9, or 14% chance of abuse). Evaluators can err by believing that characteristics distinguish abuse status when they do not, by failing to distinguish between indicators with varying degrees of evidentiary strength, by not combining information about indicators appropriately with base rates, or by assuming that two indicators are twice as much evidence as one (i.e., failing to realize that some indicators may have no incremental validity). With this

\footnote{An indicator (i.e., behavior, test result) possesses incremental validity if decision accuracy increases after that information is added to the information that is already available. For example, if psychologists made judgments based only on biographical information and then made a second set of judgments based on biographical information and Minnesota Multiphasic Personality Inventory (MMPI) profiles, one could describe the incremental validity of the MMPI (Garb, 1984).}
background in mind, consider the evidence for five general categories of indicators in sexual-abuse evaluations.

The evidentiary value of denials of abuse. Authors frequently claim that the interview with the child is the most important evidence available to professionals, yet they caution that it is common for children to deny abuse. More than any other assumption, the belief that children are highly reluctant to disclose abuse—and therefore that a denial is not indicative of abuse status—has fueled the repeated and sometimes coercive interviewing practices that typified highly publicized cases such as the McMartin and Michaels trials (see Ceci & Bruck, 1995, for a review).

Retrospective studies in which adults are asked about their sexual-abuse histories have often found that only a minority of respondents who report child sexual abuse indicate that they disclosed abuse when it occurred (e.g., Finkelhor, Hotaling, Lewis, & Smith, 1990), and these findings are frequently cited in discussions of disclosure patterns (e.g., Gonzalez, Waterman, Kelly, McCord, & Oliveri, 1993). The fact that many abuse victims never told anyone does not speak directly to the issue of disclosures during abuse evaluations, however, because it is likely that most of the adults in these retrospective studies were never asked about the possibility of abuse until the studies themselves. (For a study in which 85% of 12- to 17-year-olds reported having disclosed the abuse when it occurred, see Kellogg & Huston, 1995.) Our point here is not to claim that spontaneous disclosure rates are higher than suggested by the retrospective self-report studies, but rather to argue that the base rate of spontaneous disclosures may not be relevant to the question of how often abused children deny abuse when asked in a formal or informal interview.

The literature on children's disclosures during therapy or investigations has so heavily emphasized children's reticence to discuss abuse that Jones (1996) called belief in "the gradual unfolding process" a "sacred cow" for front-line practitioners (p. 879). Empirical studies of disclosure rates among suspected child victims have yielded varying results, suggesting that willingness to disclose abuse depends on several factors. The most frequently cited article is a retrospective analysis by Sorensen and Snow (1991) of files from 116 children who were referred for therapy by child-protective services, law enforcement, other mental health personnel, or private referral. The main results were that almost three fourths of the children denied having been abused, that they often moved into tentative disclosure prior to making a full disclosure of abuse, and that 22% recanted their allegations (with the majority reaffirming their allegations over time). Although this study stands as an interesting description of the disclosure process in a therapeutic setting, the question is whether it represents the general disclosure process of abused children during investigative interviews.

A large percentage of Sorensen and Snow's sample consisted of children who were referred for therapy based on suspicions of abuse stemming from factors such as time spent with an alleged or convicted offender (28%) or sexual behavior (14%), and it is not known how many children gradually disclosed abuse versus gradually constructed reports as a function of contact with adults who believed that abuse was likely. Second, the authors' examples of "tentative" disclosure included statements such as, "It only happened once," and "He tried to touch me but I hit him and ran away," with category assignments to these statements ("minimizing" and "empowerment," respectively) that seem to imply that children's statements should not be taken at face value. Such statements were considered "tentative" only because the children, over time and continued intervention, came to report more abuse. There are no data on how these children were questioned or evaluated, however, and it is not known how often the initial versus the final reports were more accurate.4

Contrasting Sorensen and Snow's data is a recent study by Bradley and Wood (1996). These investigators evaluated 249 validated sexual-abuse cases from the files of the Department of Protective and Regulatory Services (DPRS) in El Paso, Texas, ranging from preschoolers to 18-year-olds. Most (72%) reportedly had disclosed to someone prior to contact with DPRS or police. Only 6% of these victims initially denied abuse, and only 4% recanted previously reported abuse. Bradley and Wood suggested that data from therapy settings, such as the findings of Sorensen and Snow (1991), might be biased toward documenting reluctance, whereas the high rates of disclosure in their study could be a function of the different social expectations children have when they are interviewed for legal purposes (see also Sternberg et al., in press), or to DPRS bias to minimize denials or reluctance in their notes. The complete picture is likely to be a continuum between these extremes.

As these comments would suggest, one factor that likely affects disclosure rates during formal interviews is whether or not the child has already reported abuse to someone. In a study by Keary and Fitzpatrick (1994), the majority of children who were referred with a reported disclosure also disclosed during an evaluation that typically involved two or three sessions (percentage of disclosures by age group were 0–5 years, 59%; 6–10 years, 93%; 11–15 years, 95%; and 16+, 100%). Disclosure rates were low for children referred for emotional or behavioral problems, sexualized behavior, or physical signs "suggestive of child sexual abuse" (Keary & Fitzpatrick, 1994, p. 546) (0–5 years, 8%; 6–10 years, 15%; 11–15 years, 23%; and 16+, 0%). A second factor related to disclosure is parental beliefs or attitudes. D. M. Elliott and Briere (1994) found higher rates of disclosure among children whose mothers accepted the possibility of

4It is noteworthy that in an earlier article Snow and Sorensen (1990) described five cases of ritualistic abuse in different neighborhoods, noting that, "The cases of S victims (aged 5–16 yrs) illustrate how most Ss showed little symptomology at initial referral with significant increases during disclosure. Perpetrators involved religious leaders, women, and many juveniles that were considered conscientious, responsible members of their community" (p. 474).
abuse compared to mothers who were disbelieving despite the presence of confirmatory external evidence (see also Lawson & Chaffin, 1992).

We agree with Bradley and Wood’s (1996) conclusion that there are currently not enough data to understand the disclosure process and how it is affected by social context. There is no doubt that many abused children do not spontaneously disclose their abuse, but this fact must be clearly separated from the question of what they report when directly asked about abuse, or what percentage of children report abuse during investigative interviews when there is already a strong suspicion and public dialog about abuse. Some children are reluctant to disclose abuse, and this may be particularly likely when the child is quite young, when the abuse has not previously been discussed, when the child has a close and dependent relationship with the perpetrator, or when interviewing procedures are intimidating. Our impression, however, is that beliefs about children’s reluctance to disclose sexual abuse have often exaggerated their reticence. It is possible that greater understanding of the limitations of behavioral indicators for identifying sexual abuse will lower the proportion of unwarranted referrals, that improved understanding of the vulnerabilities of children’s reports will reduce the use of highly suggestive interviewing procedures, and that these changes may in turn alter expectations about gradual disclosure in legal settings. At the very least, recent data warn forensic evaluators to be cautious about assuming that a child who denies abuse is merely reluctant to disclose.

**Behavioral indicators of abuse status.** As noted by R. D. Wells et al. (1995), a parent’s report of behavioral change is often the first indicator that triggers a practitioner’s suspicion of sexual abuse. Numerous behavioral and emotional indicators of abuse have appeared in reviews, including poor academic functioning, sleep problems, somatic complaints, internalizing and externalizing disorders, and sexualized behaviors.

Studies of behavioral indicators have compared sexually abused children to nonabused community controls, physically abused or neglected children, or clinic controls. Several findings are typical. First, studies generally find that sexually abused children show a greater frequency of behavioral disorders than community controls (e.g., Dubowitz, Black, Harrington, & Verschoore, 1993; D. M. Elliott & Briere, 1994), but less than is found in other clinical samples (e.g., Cohen & Mannarino, 1988). Second, greater differences between sexually abused children and other groups are found when parent ratings are used rather than self-ratings. For example, Cohen and Mannarino found that parents of sexually abused children rated them as having more behavioral problems than did parents of a normative sample but less than did parents of a clinical sample, whereas the sexually abused children themselves did not self-report significant depression, anxiety, or low self-esteem. This null result for self-reports was replicated by D. J. Elliott and Tarnowski (1990). The discrepancy between parent and self-ratings is troublesome, because parents whose children have been identified as abused may have different criteria for reporting behavioral problems, and thus it is critical in the future to compare the behavioral symptoms of children who are identified as abused with a group with reported but unlikely abuse. Third, despite significant differences between abused and nonabused groups in the frequency of some behaviors, there is a great deal of overlap in the symptom profiles of these groups, and most behavioral correlates of sexual abuse are not specific to abuse (e.g., Gale, Thompson, Moran, & Sack, 1988). The large degree of overlap may help explain the considerable inconsistency in the literature on which indicators distinguish abuse status. For example, Dubowitz et al. (1993) concluded that sexually abused children more often evinced depression, sleep problems, and somatic complaints, whereas R. D. Wells et al. (1995) did not find that nightmares, sleeping more than usual, crying easily, or physical complaints were useful for distinguishing abuse status (despite a similar mean age in the two studies).

The behavioral predictor that consistently emerges as most useful is inappropriate sexual behavior. In her review of six studies, Slusser (1995) concluded:

While the sexually abused children studied exhibited significantly more indicators of emotional and behavioral disturbance than did comparison groups of “normal” children, when compared to non-sexually abused children receiving psychiatric services only one discriminating variable emerged. These studies empirically support the growing impression among clinicians that overt sexual behavior, inappropriate for age, is an indication of sexual abuse. (pp. 488–489)

In addition to the studies reviewed by Slusser, numerous other studies have also found that the frequency of sexual behavior discriminates between sexually abused children and comparison groups, often with frequency differences that are noticeably larger than those for other recorded behaviors (e.g., Dubowitz et al., 1993; D. M. Elliott & Briere, 1994; Gale et al., 1988; R. D. Wells et al., 1995).

Because sexual acting out has long been reported to occur more frequently among sexually abused children, there has been considerable interest in establishing normative rates of sexual behaviors and identifying specific behaviors that distinguish between abused and nonabused groups. Regarding normative data, sexual behavior in childhood occurs frequently and is often detected by parents. For example, S. Lamb and Coakley (1993) asked 300 undergraduates at an all-women’s college to complete a survey about childhood sexual play and games. Among the 128 who returned surveys, 85% reported that they had participated in a childhood sexual game. A typology derived from the respondents’ narratives included playing doctor (16%), exposure (15%),
experiments in stimulation (14%), kissing games (6%), and fantasy sexual play (which included feigning intercourse, 30%). Overall, 30% reported being persuaded, manipulated, or coerced to play sexually. Respondents reported being discovered 44% of the time, with most “discoverers” ignoring the game. These data are not inconsistent with parental reports. For example, Gordon, Schroeder, and Abrams (1990a) investigated age differences in children’s knowledge of sexuality between 2 and 7 years of age. A significant minority of parents (28.5%) reported that their children had been exposed to sexually explicit materials, soap operas, or adult magazines; 49.6% reported that their child had masturbated; and 30.1% indicated that their children had engaged in sex-exploratory play (see also Gordon, Schroeder, & Abrams, 1990b). In an extensive study of children between 2 and 12 years of age, Friedrich, Grambsch, Broughton, Kuiper, and Beilke (1991) concluded that some behaviors, such as exhibitionism and self-stimulatory behavior, were observed by parents with high frequency, whereas behaviors that are more aggressive or imitative of adult sexual behavior were rarely observed.

The magnitude of differences between sexually abused and nonabused children varies with study methodology (see Friedrich, 1993, for a review), but it is clear that some behaviors that are rarely observed in nonabused samples occur with much higher frequency among abused samples. Friedrich et al. (1992) contrasted a normative sample of 880 children with 276 sexually abused children on the Child Sexual Behavior Inventory (Friedrich et al., 1991), a checklist completed by parents that includes 35 items ranging in typicality from “uses opposite sex toys” to “puts mouth on sex parts.” Results revealed significant differences between groups on 27 of the items. Because sample sizes were large, many significant results represented differences in frequency that are small for predictive purposes, but other measures yielded quite large discrepancies between groups (e.g., imitates intercourse, with a percentage occurrence in the nonabused sample of 11.1, but 14.1 for the abused group). Overall, the instrument had good sensitivity (i.e., the ability to detect sexually abused children) for boys and girls ages 2–6 years and for boys ages 7–12 years. Specificity (i.e., the ability to discriminate abused and nonabused groups) was less than satisfactory, however, leading the authors to conclude that “the CSBI should not be relied on in isolation as the primary indicator of sexual abuse” (Friedrich et al., 1992, p. 311). Practitioners need to avoid overreacting to sexual behavior, because most of the frequency differences reported by Friedrich et al. are representative of weak indicators of abuse, because it is not known to what extent the abused children evinced elevated sexual interest due to having been involved in sexual-abuse suspicions, and because rates of sexual behavior in nonabused samples are influenced by family practices and norms regarding nudity and sexuality (Friedrich et al., 1991).

In summary, data on behavioral indicators of abuse generally find that discriminative validity is low. Although reliance on behavioral indicators raises obvious concerns about false identifications of abuse, the search for traumatic reactions can also lead practitioners to ignore children who have been abused. Many abused children do not show high rates of symptomatology (D. M. Elliott & Briere, 1994), yet professionals have sometimes failed to substantiate abuse due to apparently good mental health status that was deemed inconsistent with their profile of abused children (Haskett, Wayland, Hutcheson, & Tavana, 1995). In short, neither the presence of symptomatology nor its absence has yet been shown to improve decision accuracy beyond what would have been decided on the basis of the child’s report or nonreport of abuse. This is clearly an area in need of continued research.

The content and characteristics of children’s abuse reports. Numerous features of children’s reports are often said to bolster confidence in the validity of the account, including appropriate emotional expressions, consistency, and contextual detail. Regarding emotional presentation, Haskett et al. (1995) found that child-protection investigators sometimes justified decisions to substantiate on the grounds that the child readily disclosed abuse and was matter-of-fact, whereas others used hesitancy and embarrassment to justify substantiation. These authors concluded, “Given the wide range of affective and behavioral responses of abused children and the difficulty in separating the child’s reactions to the alleged abuse from their response to the interview process, the use of affect as evidence of the validity of allegations . . . seems highly problematic” (Haskett et al., 1995, pp. 41–42). The assumption that abused children typically react emotionally in interviews was challenged by a recent analysis of 55 videotaped interviews of high-risk sex abuse cases seen at a multidisciplinary assessment center (B. Wood, Orsak, Murphy, & Cross, 1996): These children were rated as attentive, on-task, relaxed, and displaying few emotions.

Criteria-Based Content Analysis (CBCA) is a system for coding children’s reports that has been used for over 35 years in Europe. CBCA rests on the assumption that statements based on experienced events will contain a set of 19 criteria more often than fabricated (or, perhaps, imagined) events. The CBCA criteria include many characteristics that appear in discussions of true and false allegations, including general characteristics (e.g., logical structure and spontaneous digressions), specific contents (e.g., unexpected complications or interruptions), and peculiarities of content (e.g., peripheral details). Studies supporting the ability of CBCA to distinguish between true and false reports (e.g., Boychuk, 1991; Raskin & Esplin, 1991a, 1991b) have been criticized on a number of grounds, including insufficient reliability (Anson, Golding, & Gully, 1993), lack of data on whether the technique is appropriate for analyzing transcripts from younger children (G. L. Wells & Loftus, 1991), and basing results on comparisons of cases that were highly improbable.
and highly probable, which might exaggerate the ability of the technique to discriminate (M. E. Lamb et al., 1997). Concerns about reliability have been alleviated in recent studies in which raters coded from written transcripts rather than tapes (Horowitz et al., in press), but there are still serious concerns about the forensic application of the technique. M. E. Lamb et al. (in press) tested whether CBCA could discriminate between plausible and implausible accounts. Starting with 1,187 interviews of alleged victims, they selected 98 cases (representing children ages 4–13 years) that contained sufficient independent evidence, apart from the quality of the child’s narrative, to enable the researchers to classify the allegations as plausible or implausible. Although statements from the plausible cases did contain more criteria than those from the implausible cases, the difference was smaller than previously reported and several of the criteria were nonsignificantly more frequent in the implausible group. As M. E. Lamb et al. (1997) explained, “the surprisingly small differences between the plausible and implausible accounts occurred not because the plausible cases obtained particularly low scores but because some of the implausible cases obtained unexpectedly high scores” (p. 262). M. E. Lamb et al. noted that a false account may contain many criteria if the specific allegation, such as vaginal penetration, is embedded in a description of a recent interaction with the alleged perpetrator. This conclusion is consistent with concerns raised by Ceci, Bruck, and their colleagues (e.g., Bruck, Hembrouke, & Ceci, in press; Ceci, Huffman, Smith, & Loftus, 1994) that reports implanted in laboratory studies often contain features considered diagnostic of true reports. As M. E. Lamb et al. concluded, “the findings . . . underscore how CBCA scores should not yet—and perhaps should never—be used in forensic contexts to evaluate individual statements” (pp. 262–263).

**Projective techniques.** Professionals who evaluate children for possible sexual abuse use a wide variety of techniques to supplement verbal interviews. For example, a study of 201 Boston-area professionals who were interviewed by telephone identified 29 techniques in addition to verbal interviewing and use of anatomical dolls, including various types of drawings, toys, coloring games, and verbal techniques such as story telling (mean number of techniques used was 4.10 for mental health professionals and 1.74 for law-enforcement professionals; Kendall-Tackett, 1992b). The literature on these techniques is sparse, and the small number of relevant studies are fraught with methodological problems including lack of data on reliability, lack of control groups, and no mention of whether the techniques have incremental validity over more direct assessment methods.

The most heavily researched ancillary technique is anatomically detailed (AD) dolls, a topic that has been extensively reviewed in recent years (e.g., Ceci & Bruck, 1995; Everson & Boat, 1994; Koocher et al., 1995; Wolfner, Faust, & Dawes, 1993). The limitations of AD dolls are the same as the aforementioned limitations of behavioral indicators. First, although the collective evidence suggests that sexually abused children produce more explicitly sexual behavior with the dolls, nonabused children do sometimes engage in sexual demonstrations. Even if the base rate of these behaviors is substantially lower in nonabused children, the higher proportion of nonabused children in the general population would cause a great number of false positives if the dolls were used as a preliminary screen for sexual abuse. Second, comparing referred and nonreferred children is hardly an appropriate comparison for testing the usefulness of the dolls in investigatory interviews. As Wolfner et al. put it, “The problem is that the differential behavior may be reflecting little or nothing more than the types of behavior and background variables that led to suspicion—or referred versus nonreferred status—in the first place” (1993, p. 8). Ceci and Bruck also questioned the extent to which mere suspicions of abuse followed by several investigative interviews—often using AD dolls—would contribute to sexualized behavior. In short, there is no evidence that doll play provides any incremental validity and no support in recent comprehensive reviews for the use of AD dolls to establish that abuse has occurred. (For a field study that failed to find benefits from doll use in the context of primary investigative interviews, see M. E. Lamb, Hershkowitz, Sternberg, Boat, & Everson, 1996).

**Family context.** Concerns about false allegations arose predominantly from prominent day-care trials and allegations of sexual abuse made during contested custody proceedings. Although only a small percentage of divorce cases involve allegations of child sexual abuse, allegations occur much more frequently in families with custodial visitation disputes than in families without such conflict (Theoennes & Tjaden, 1990). Researchers often report that the majority of child-abuse allegations in divorce situations are believed to be valid (e.g., 65% reported by Faller, 1991). There is disagreement, however, about whether the discrepancy between false allegation rates in divorce and nondivorce situations reflects a real difference or merely different criteria for substantiating abuse that discriminates against positive findings in divorce cases (e.g., Berliner & Conte, 1993), and whether the difference is great enough to warrant lowering one’s estimate of abuse probability in light of a custody
battle (e.g., J. M. Wood, 1996). In a study of decisions to prosecute, only a minority of biological fathers and mothers’ boyfriends were prosecuted compared to 50% or more in every other category of perpetrator–child relationship (Cross, De Vos, & Whitcomb, 1994). Because the presence of psychological or medical evidence did not significantly increase prosecution rates in this data set, the lower prosecution rate for biological fathers and boyfriends could mean that suspicions about validity when there is family conflict are widespread and difficult to overcome in court.

J. M. Wood (1996) offered an interesting analysis of weighing evidence about the context of an allegation that illustrates the difficulty of forming general guidelines about the value of single indicators. J. M. Wood selected an a priori probability of abuse in an investigation, before any information is collected, of 7:3 (7 valid cases for every 3 invalid cases). Assuming that the percentage of false allegations in custody disputes is close to 50%, J. M. Wood computed the evidentiary value of knowing there was a custody issue by using the odds form of Bayes’ theorem. The result was a likelihood ratio of 1:2.33, which is considered a weak indicator of abuse (7:3 × 1:2.33 = 1:1). J. M. Wood concluded that “evidence of custody dispute should have little or no influence on professional decisions regarding the allegations” (1996, p. 34). In contrast to J. M. Wood’s approach, however, information about the presence or absence of a custody dispute could be viewed as adjusting the base rate from 7:3 without this information, to 1:1 in the case of a custody issue, and some value higher than 7:3 in the absence of custody issues (because the prediction of 70% valid allegations included divorce cases). With a discrepancy of more than 20% in the estimated probability of abuse, there are circumstances in which this information could be extremely important. For example, if there was other evidence against the validity of the allegation, the probability of abuse in the custody situation (assuming sufficient information to apply Bayes’s theorem) could drop below threshold for pursuing the allegations.

Summary. When all is said and done, there is no scientifically validated procedure for combining indicators of abuse that can circumvent the absence of a verbal report of abuse by a child or remedy deficiencies in that report. As Paradise (1989) concluded in an article on medical indicators, “Clinician comfort aside, the diagnosis of sexual abuse inevitably rests . . . on descriptive statements made by a child” (p. 176).

The Assessments Approach

The assessments approach derives from data on children’s eyewitness reporting in laboratory and field studies. The predominant focus of this approach is to compare children’s reports across interview conditions that vary in such factors as the delay between the event and an interview, the number of interviews, the type of interviewing (free recall, specific questioning, leading questioning), and the presence of contaminating influences outside the formal interview environment (e.g., stereotype induction or other misleading information presented during informal conversations). Applications of these data in legal settings involve analyzing the “chain of custody” (R. Rosenthal, personal communication, November 1, 1996) of children’s memories, or the extent to which children have been exposed to influences that are known to reduce the reliability of their reports (e.g., Ceci & Bruck, 1995). Because studies show that many children are quite accurate despite a nonoptimal assessment history, information about the conditions associated with false reports is most useful when children have made unusual or mushrooming allegations, or in which there are other grounds for skepticism regarding the allegations. In these cases, laboratory and field studies explain how children might come to report unusual but nonexperienced events. Studies in this tradition have been plagued, however, with the same result that has limited practical application of indicator research—large variability in children’s responses to adult influence. This variability has led researchers in recent years back to the question of indicators; in this case, the characteristics of accurate and inaccurate reports. In the assessments approach, however, research on potential indicators of the accuracy of children’s reports has focused on studies in which the objective truth of a report is known.

Numerous recent reviews have summarized which assessment conditions are associated with more or less accurate reporting (e.g., Berliner & Conte, 1993; Ceci & Bruck, 1993; 1995; M. E. Lamb, Sternberg, & Esplin, 1994, Steward, Bussey, Goodman, & Saywitz, 1993), and these data have been translated, with some variability in recommendations, into protocols that outline best-practice standards for interviewing children (e.g., American Professional Society on the Abuse of Children, 1990; Home Office, 1992; M. E. Lamb, 1994; Poole & Lamb, in press). We briefly summarize these findings in preparation for asking two questions: What is the practical significance of information about a child’s assessment history, and what are the characteristics of accurate and inaccurate reports in laboratory studies?

Conditions associated with accurate and inaccurate reports. Although there is no doubt that some young children can and do lie when instructed to by an adult,6 eyewitness

---

6Across studies, many young children do lie when instructed to do so. For example, Quas, Denton, Goodman, and Myers (1990) arranged for 35 children (ages 4 to 7 years) to play with a research assistant who either did or did not touch them on their bare stomachs, noses, and necks. Children were assigned to one of three conditions: 13 children who had been touched were told by another research assistant to tell the truth during a subsequent interview, 11 children who had not been touched were told to tell the truth, and 11 children who had not been touched were asked to help the research assistant by “being actors” and pretending to have been touched. Children who were asked to lie complied during 44% of the target questions, whereas children who actually had been touched but were told to tell the truth correctly answered “yes” only 51% of the time, and children who had not been touched correctly answered “no” 82% of the time. Similarly, in a study reported by Houts (1994), 56% of 4- to 11-year-olds blamed an innocent individual for a theft to protect their parents’ from detection. Tate, Warren, and Hess (1992) found that children between the
Researchers have been predominantly concerned about false allegations arising from other mechanisms when young children are involved. There is a large body of evidence on three factors that affect the likelihood of unintentionally distorted or inaccurate accounts: memorability of the event (i.e., recency and salience), the influence of specific or leading questioning, and the influence of repeated misleading suggestions.

Regarding memorability, testimony accuracy generally declines over time, but the rate of decline is a function of age (with younger children being more adversely affected by delays), salience of the event (with fewer errors for more salient events), and the types of questions interviewers ask (with the ill effects of specific and misleading questions increasing with delay). Consider several illustrative studies. Goodman, Hirschman, Hepps, and Rudy (1991) reinterviewed 3- to 7-year-old children 1 year after a clinic visit during which they had received inoculations. Mirroring many studies, the amount of accurate information in free recall declined over time, whereas the amount of inaccurate information did not, yielding a lower percentage of accurate information after the delay (90%) than initially (99%). As these percentages indicate, free recall was generally accurate even after 1 year, but performance on specific action questions and misleading questions was problematic after the delay, with the children correctly answering 73% of specific action questions and only 51% of misleading questions. When asked to identify a nurse from photos, correct identification dropped from 50% initially to 14% at the follow-up interview. Problems can also emerge at shorter delays, especially when interviewers move from free recall to more specific questions. For example, in a study by Orinstein, Gordon, and Larus (1992), 6-year-olds were consistent in their performance across a 3-week interval, but 3-year-olds showed declines in accuracy by 1 week and an increase in false assertions (to a high of 50%). It is important to note, however, that even after a 6-week delay 3- to 7-year-olds responded correctly to 93% of nonexperienced events when they were interviewed about a highly invasive and novel medical procedure (Merrit, Orinstein, & Spicker, 1992).

The general finding, then, is that even young children perform very well when interviewed shortly after an experience with free-recall questions such as “Tell me what happened,” provided they have not been exposed to misinformation outside the interview environment. (Exceptions to this general trend often involve very young children; e.g., Goodman, Quas, Battmer-Faunce, Riddlesberger, & Kuhn, 1994). Young children’s accuracy declines more rapidly than older children’s over long delays, and errors can include serious false reports such as attributing actions to the wrong individual (Poole & White, 1993). Developmental patterns are more variable across delays of several weeks or months, being largely a function of how memorable the event was and how specifically adults question them.

A robust finding is that children’s accuracy declines as questioning moves from free-recall questions (e.g., “Tell me what happened”), to open-ended but more focused questions (e.g., “Can you tell me what the man looked like?”), to specific questions that inquire about a single detail (e.g., “What color was his hair?”) or that ask children to select from a limited set of options (such as multiple choice or yes–no questions). This finding has been replicated across studies with very different procedures, age groups, subject populations (i.e., normal versus cognitively impaired children), and delays between events and interviews (see Dent, 1991, for a review).

The problem with specific questions is that children often try to answer them whether they know anything about the target information or not. In a recent study by Walker, Lunnning, and Elits (1996), kindergartners, second graders, and fifth graders watched a videotape and answered 36 multiple-choice questions, such as, “Did you see a little girl or a little boy on the video?” The investigators varied the placement of the correct alternative such that it appeared as the first or second option, or not at all. The children showed a marked response set in favor of selecting the second option. When neither option was correct, kindergartners answered correctly only 17.5% of the time, second graders 37.5% of the time, and fifth graders 48.3% of the time. Similarly, Poole and Lindsay (1996) found that younger children were more likely than older children to assent to leading questions about science demonstrations they neither experienced nor heard about from their parents. (It is interesting that although older children less often assented to leading questions about these novel science demonstrations, those who did assent were more likely than younger assenting children to provide narrative descriptions of the novel events.)

Although children and adults alike often field answers to specific questions, this does not necessarily mean that children will falsely assent to questions about sexual matters. A famous study on this topic was conducted by Saywitz, Goodman, Nicholoh, and Moan (1991). They interviewed 72 girls ages 5 and 7 who had experienced a medical check-up. Half of the girls had received a vaginal and anal exam, the other half had not. The majority of children in the genital-touch condition disclosed genital contact only when asked direct questions, and only a small percentage of children falsely reported genital touch in response to direct questions (2.9% falsely disclosed vaginal touch and 5.6% falsely disclosed anal touch). However, 21.7% of children in the genital condition falsely reported tapping of the spine (scoliosis examination), indicating that questions about less salient events are likely to receive higher false assentions. Across studies, 3-year-olds are more likely than older children to assent to specific questions that could be indicative.

Of 28 and 8 years expressed willingness to go along with a lie that was suggested by an unfamiliar research assistant, but that younger children were less successful in actually carrying through with the trick. Out of 20 children who were asked to lie, only 7 were successful throughout questioning in maintaining the false story.
of abuse. For example, 3-year-olds averaged 14% commission errors across conditions in Goodman, Rudy, Bottoms, and Aman (1990, Experiment 2) as opposed to 5% for the 5-year-olds.

It is unclear whether children involved in actual investigations will be as resistant to specific questioning about abuse as these analog studies indicate. For example, the girls interviewed about genital touching in Saywitz et al. (1991) probably were not expecting questions about embarrassing issues and were not enmeshed in environments that would disinhibit them from talking about such issues. In another study, false reports of touching were not uncommon when younger children (ages 3–6 years) were asked direct questions with the interviewer pointing to either a doll or body drawing. Steward and Steward (1996) interviewed children immediately, 1 month, and 6 months after a pediatric clinic visit. During the immediate interview, 8% of the children falsely reported genital touching with prop-assisted direct questioning, 12% falsely reported touch to the buttocks, and 18% falsely reported anal touch. Corresponding percentages of commission errors at a second interview 1 month later were 15%, 12%, and 30%; at the final 6-month interview these percentages were 12%, 6%, and 36%. It could be argued that medical analog studies overestimate children’s willingness to make false disclosures of genital touch, because genital touching is not script-inconsistent in the context of a medical exam. However, the flip side of this issue is that these rates of commission errors might better reflect children’s willingness to report genital touching when other environmental conditions lead them to expect questions about touching (e.g., repeated exposure to sexual-abuse prevention materials followed by repeated questioning about touching). What we can conclude, however, is that children generally are less accurate with direct questioning, that younger children (especially 3- and 4-year-olds) are more likely than older children to acquiesce to questions about nonexperienced touching (e.g., Bruck, Ceci, & Francoeur, 1995; Bruck, Ceci, Francouer, & Renick, 1995), and that the majority of children who have not been primed about abuse issues are accurate on abuse-related questions when questioning is not repeated or misleading.

False reporting sometimes increases dramatically when young children are exposed to suggestive influences either outside formal interviews or within the interview environment, especially when misleading questioning is repeated. Earlier in this review, we described examples of procedures that implant false reports. Techniques include asking children to think repeatedly about events that did not occur (Ceci, Loftus, et al., 1994), telling children accusatory information about an individual prior to specific questioning (Leichtman & Ceci, 1995), and reading children books that describe nonexperienced events (Poole & Lindsay, 1995, 1996). Regardless of the procedure used to introduce misinformation, false reports tend to increase when interviewers ask leading questions or questions that imply that the nonexperienced events actually happened.

How many children actually succumb to misleading information and leading questions? The specific percentage varies greatly depending on the experimental procedure, including the content of the suggestion and the strength of the suggestibility manipulation. What is clear is that children are very accurate in the absence of adult suggestion, but that a significant proportion are susceptible to influence. In Leichtman and Ceci’s (1995) “Sam Stone” study, for example, some children (ages 3 to 6 years) watched as a stranger named Sam Stone visited their preschool, walked about, and left. After four interviews about this visit that included no suggestive questions, they were asked about two fictitious events: “Did Sam Stone do anything to a book or a teddy bear?” These children were highly accurate; for example, only 10% of the 3- and 4-year-olds said that Sam did something to the book or teddy, only 5% maintained that response when asked if they actually saw him do anything, and only 2.5% continued to do so when gently challenged (“You didn’t really see him do anything to the book/teddy bear, did you?”). None of the older children continued to claim to have actually seen Sam do these things when challenged in this mild way. A second group of preschoolers was exposed to repeated conversations about how clumsy and prone to breaking things Sam was, and were subsequently interviewed with suggestive questions (e.g., “Remember that time Sam Stone visited your classroom and spilled chocolate on that white teddy bear? Did he do it on purpose or was it an accident?”). In the final interview, 72% of the youngest preschoolers reported that Sam did something to the book or teddy, 44% said they actually saw him do these things, and 21% maintained their false stories when gently challenged. The 5- to 6-year-olds were less malleable: Only 11% said they actually saw the misdeeds, and less than 10% maintained that story when challenged. In Ceci, Loftus, et al. (1994), the mean percentages of false assertions to negative events (involving injury) were 17% and 10% initially for children ages 3–4 and 5–6 years, respectively, and 28% and 23% during the final (12th) interview. With more suggestive interviews in another study, over 70% of preschool children falsely asserted to both positive and negative nonexperienced events by the fourth interview (Bruck, Hembrooke, & Ceci, in press). Moreover, false assertions are not limited to nonabused children or to nonabuse questions. For example, a study of children who had participated in a 5-day hospital stay as part of an ongoing forensic investigation found that 3- to 5-year-olds asserted falsely to 20% of misleading abuse-related questions (questions that asked about a nonexperienced event that could be considered abuse), compared to only 4% for 6- to 10-year-olds (Eisen, Goodman, Qin, & Davis, in press). Finally, it is worth emphasizing that leading or coercive questioning is not necessary for eliciting false reports; we found errors in free recall merely by hav-
ing parents read their children books that described nonexperienced events (Poole & Lindsay, 1995, 1996).

The practical significance of information about a child’s assessment history. Studies of children’s eyewitness accuracy provide little basis for estimating the probability that a child would make a false report under the conditions of any particular sexual-abuse investigation. Reporting accuracy is influenced by multiple factors, including children’s ages, the delay between the target event and the interviews (and other factors that would influence the memorability of the events in question), and the number and type of interviews. Most suggestive conversations are not recorded in actual cases, which precludes analyses of suggestive influences, and even if they were recorded it is unlikely that the specific features of any particular case would exactly match a published study. Nonetheless, there is sufficient similarity in findings across studies to contrast rates of true and false reports under conditions of nonleading and leading interviews. Furthermore, this analysis illustrates how easily leading questioning could reduce estimates of the probability of abuse to below 90%, or well within “reasonable doubt.”

A Bayesian approach explains why leading questioning might seriously reduce decision accuracy—particularly when the majority of abused children respond “yes” to direct questions about abuse. Leading interviews will have no effect on decision accuracy only when those interviews proportionately change the rate of responding “yes” for abused and nonabused children.7 With a reporting rate of 65% to direct questions, for example, leading interviews could at most augment “yes” responses by 54% (i.e., 35% increase/65%), whereas the potential to increase false responses is less constrained. To understand this argument, assume that 65% of abused children disclose abuse in response to direct questions, but that an additional 25% disclose after leading interviewing. Leading interviewing therefore increases the disclosure rate to 90% (i.e., 25%/65% is a 38% increase in disclosures from using leading questions). Also assume, however, that 5% of nonabused children falsely report abuse in response to direct questions, but that an additional 20% report falsely under conditions of leading questioning, raising the rate of false reporting to 25% (i.e., 20%/5% is a 400% increase in false reports). With a prior odds of 70:30, the posterior odds of abuse given a “yes” response under no leading questioning is 97% (70:30 x 65:5), but only 89% under leading questioning (70:30 x 90:25). Bayes’s theorem therefore informs us that leading interviewing would most likely decrease estimates of the probability of abuse given a “yes” response; leading interviewing would increase confidence in a “yes” response only if it elicited proportionately more accurate “yes” responses than inaccurate “yes” responses.

What do studies indicate about the effects of leading interviews on accurate and inaccurate “yes” responses?

Consider three sets of examples of the relationship between base rates of abuse, the characteristics of interviews, and the probability that a child’s report of abuse is accurate. For illustrative purposes, we selected numerical values from published studies in which children aged 7 and under were questioned. Imagine that an expert is brought in during an ongoing sexual-abuse investigation involving a preschool child. For this jurisdiction, we selected an a priori probability of abuse of 7:3, or seven accurate allegations for every three invalid suspicions. To work the first scenario, assume that there is no evidence of misleading interviewing or other contaminating adult influence. When nontouched preschool children are asked directly about touch, we expect only a 5% false assertion rate (e.g., Leichtman & Ceci, 1995); when touched children are directly questioned, we expect a 74% accurate acknowledgment of those experiences (72% from Saywitz et al., 1991, 1-month delay with direct questions; 74% from the 1-month interview in Steward & Steward, 1996). Table 1a illustrates the results from this set of assumptions. Given a positive report of abuse, there is a 97% chance that the child was in fact abused. Table 1a then adjusts this percentage for children who have been exposed to misleading information and interviews. Under conditions of repeated leading interviewing, assume that 40% of nonabused children will acquiesce to abuse (between the 20% figure in Eisen et al., in press, for a single misleading question and the over 70% figure from Leichtman & Ceci, 1995), and that asking leading questions increases correct assertions to 95% for abused children. In this case, there is an 85% chance of abuse given a positive response. Thus with a base rate of abuse of 70% (and the other assumptions outlined in the example), the presence of leading interviews increased doubt about the child’s response from only 3% to 15%. There is still a high probability of abuse, however, showing that suggestibility manipulations would need to be strong to increase doubt substantially when the base rate of abuse among investigated cases is believed to be moderate to high.

Consider how these numbers pan out with a base rate of abuse of 50% (a frequently cited estimate for custody cases). Table 1b recalculates the probabilities of abuse given a positive response to direct abuse questions for (a) children who have had no suggestive influences and (b) children who have been exposed to misleading information. In this example, the probability of abuse given a positive report of abuse declines from 94% for the child interviewed neutrally, to only 70% for the child who had been exposed to misleading information sufficient to induce false reports in 40% of children.

Finally, how would doubt increase if the base rate we use is the base rate of abuse for all children in a particular jurisdiction? Imagine, for example, that concerns about an identified pedophile prompted officials to interview most of the children at a local day-care program. If we assume that

---

7We thank Robyn Dawes for suggesting this point.
Table 1. Hypothetical Probabilities of Abuse Given Assent to a Direct Abuse Question, as a Function of the A Priori Probability of Abuse and the Child’s Interviewing History

<table>
<thead>
<tr>
<th>Responses</th>
<th>Abused (n = 700)</th>
<th>Nonabused (n = 300)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes” responses, no leading interviews</td>
<td>518</td>
<td>15</td>
<td>$p(\text{abuse}</td>
</tr>
<tr>
<td>“Yes” responses, leading interviews</td>
<td>665</td>
<td>120</td>
<td>$p(\text{abuse}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responses</th>
<th>Abused (n = 500)</th>
<th>Nonabused (n = 500)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes” responses, no leading interviews</td>
<td>370</td>
<td>25</td>
<td>$p(\text{abuse}</td>
</tr>
<tr>
<td>“Yes” responses, leading interviews</td>
<td>475</td>
<td>200</td>
<td>$p(\text{abuse}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responses</th>
<th>Abused (n = 100)</th>
<th>Nonabused (n = 900)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes” responses, no leading interviews</td>
<td>74</td>
<td>45</td>
<td>$p(\text{abuse}</td>
</tr>
<tr>
<td>“Yes” responses, leading interviews</td>
<td>95</td>
<td>360</td>
<td>$p(\text{abuse}</td>
</tr>
</tbody>
</table>

Note. Examples assume that abused children correctly assent to 74% of direct questions with no prior leading interviews and 95% of direct questions with leading interviews; nonabused children falsely assent to 5% of direct questions with no prior leading interviews and 40% of direct questions with leading interviews.

10% of the children in this program have actually been abused at some point in time, what is the value of a positive response to abuse questions for children who were interviewed neutrally versus children who were exposed to misleading information? Table 1c shows that, for children who have not been exposed to misleading information, there is only a 62% chance of abuse given a positive response to abuse (a low value because the base rate of abuse is assumed to be low). Moreover, for children who have been exposed to misleading information, there is only a 21% chance of abuse given a positive response to a direct abuse question.

In summary, just as the presence of positive indicators of abuse increases confidence in the accuracy of a report of abuse, the presence of misleading interviews or suggestive influences lowers our confidence in the veracity of the report. Furthermore, given a constant false-assertion rate resulting from misleading information, the adjustment in our probability estimate becomes greater as the base rate of abuse becomes lower. In other words, we have the most doubt about cases in which there was little justification for questioning children in the first place, and in which questioning was known to be leading.

Debate among researchers often revolves around whether too much or too little attention is paid to the problem of leading interviewing. These debates often focus on interpretations of the research, specifically on whether data from eyewitness studies can legitimately be generalized to abuse situations. A look at the flip side to Table 1, however, reveals how professionals' work environments might influence their perspectives about interviewing. For each of the three examples in Table 1, consider the probability of abuse given that the child answers "no," that is, the probability of a false negative. With a 70% a priori probability of abuse, 39% of children who respond "no" after no leading interviews were abused, compared to only 16% after leading interviews. With a 50% a priori probability of abuse, these values are 21% and 8%. With a 10% a priori probability of abuse, these values are 3% and 1%. Note that the pattern of difference between nonleading and leading interview conditions is the reverse of that depicted in Table 1. Regarding the detection of abuse, leading interviewing dramatically decreases the number of false "no" responses when the base rate of abuse is assumed to be high, whereas the benefit of leading questioning to detect abuse is less noticeable when the base rate is assumed to be low. In other words, it makes sense that professionals who work in environments with a high base rate of abuse would favor leading interviewing techniques and downplay the risks from those procedures. Professionals who primarily analyze cases with lower base rates of abuse should have heightened concerns about the dangers of leading interviews, however.

The characteristics of accurate and inaccurate reports. Studies of child-sexual-abuse cases and analog memory studies suffer from different limitations. Sexual-abuse cases have no documented event against which to check the accuracy of a child's statement, but analog studies cannot mimic the social pressures that come into play with abuse investigations, such as embarrassment or encouragement to recant. Despite these differences, both literatures have stated that a primary indicator of accuracy is consistency. The term "inconsistency" is used in two different ways: to refer to new information that has not been previously mentioned (i.e., reminiscence), and to refer to information that conflicts in meaning with a previous statement (i.e., a contradiction). Consider the first type of inconsistency. In
two recent studies, Salmon and Pipe (1997a, 1997b), reinter- viewed young children 1 year after they had participated in studies on the influence of props on eyewitness reporting. In the first study (1997a), 3- and 5-year-olds participated in a quasi-medical event involving a sick teddy bear and were interviewed 3 days later and again after 1 year. Many stud- ies report that reminiscence, or reporting new information across interviews, is an extremely common phenomenon (e.g., Fivush & Schwarzmueller, 1995). The question of interest here, though, is the accuracy of that new information. Salmon and Pipe (1997b) found that information reported at the 3-day delay that was not repeated at the 1-year delay was highly accurate (84% correct), as was information reported during both interviews (97%). Information that popped up only at the 1-year delay was significantly less accurate (54%), however. The second study, involving interviews about a routine health assessment in school, showed a similar pattern. Only 68% of the information recalled at the 1-year delay had been previously recalled, but 95% of this redundant information was accurate. In contrast, only 51% of the information that was unique to the 1-year delay was accurate. Thus it is common to report new information, and much of this new information is accurate, but much of it is inaccurate. These data lend empirical support to the practice of questioning the accuracy of information that emerges only after a delayed interview.

The most comprehensive content analysis of children's reports was described by Bruck, Hembrooke, and Ceci (in press). These investigators asked 16 preschool children to describe a positive event that had been staged for each child individually at school (helping a visitor who tripped and hurt her ankle), a negative event from the past few weeks in which the child had been punished, and a positive and a negative fictitious event (helping a lady find her lost monkey in the park and a food-stealing incident). Data were analyzed for five interviews: an initial baseline interview in which the children simply reported on each event, two suggestive interviews (involving visualization techniques, repeated misinformation, and selective reinforcement), a fourth nonsuggestive interview in which the children told the events to a hand-held puppet, and a final interview by a different research assistant.

Bruck et al. (in press) looked at a number of variables that often are presumed to discriminate between true and false accounts, including number of details, spontaneous reminiscences and consistency, contradictions, and narrative cohesion. Children gave few details in the first interview regardless of event type, but by the third interview there were no significant differences in the number of details provided for the true-positive and the two false events. (Narratives regarding true-negative events were sparse, perhaps because there simply was not much to report about the punishment incident, and this condition was therefore deleted from subsequent comparisons.) Most reminiscences were spontaneous, and the children provided more reminiscences for false than for true events. Regarding narrative cohesion (as evidenced, for example, by temporal markers), there were few differences but the differences that were obtained were in the direction of false stories being more coherent than true stories. Also, children's stories became more elaborated with emotional terms and descriptives over time, such that by the third interview the false stories contained more elaborative detail than the true stories. True stories were more consistent initially, but by the fifth interview there was no difference in the proportion of consistent details for the false unpleasant and true pleasant events. Finally, there were few contradictions (details that conflicted with a previously described detail) in the children's stories, but the proportions of contradictory details were similar across events and increased with repeated interviewing. Regarding their extensive content analysis, Bruck et al. concluded:

To summarize, we examined the linguistic structure of the children's true and falsely suggested narratives; we examined changes in structure and content of these narratives as a function of repeated interviews. Replicating many of the results of a pilot project involving a different sample of children (Bruck, Ceci, & Hembrooke, 1995), our analyses revealed the following. First, it is the first narrative which was elicited by nonsuggestive techniques that allows the clearest differentiation between true and false stories. This is because children mainly deny the false stories and as a result these contain no details. However with repeated interviews, the false stories quickly come to resemble the true stories in terms of the number of details provided, the spontaneity of the utterances, the number of new details, [contradictions] across narratives, the elaborativeness of the details, and the cohesiveness of the narrative. It is only consistency across narratives that differentiates true from false events; it seems, however, that could become a less potent predictor when children are repeatedly interviewed. When false stories are told as a result of repeated suggestive interviewing, they take on additional qualities that make them seem more believable than true narratives: specifically, after a number of interviews, false narratives contain more descriptive, elaborative and emotional material than true narratives.

Summary. Research on children's eyewitness reports documents the double-edged sword of repeated leading or suggestive experiences: These experiences can elevate the rate of false reports while simultaneously obscuring the differences between accurate and inaccurate accounts. Among the many unanswered questions on this topic is the extent to which children's original testimonies can be "recovered" by special procedures, such as isolating children from suggestive influences followed by training that clarifies the importance of distinguishing between actual experiences and
knowledge from other sources. Unfortunately, such research programs are still in the early stages of development.

Conclusion and Recommendations

Controversy about the accuracy of children’s reports of sexual abuse reflects two challenges. First is the challenge of understanding development, which is characterized by resiliency, cognitive immaturity, and variability. These principles, which are the basis for our success as a species, make it extraordinarily difficult to “postdict” our pasts or forecast our futures. Early efforts to identify indicators of sexual abuse often ignored resilient children, and therefore failed to appreciate the range of reactions to abuse. Cognitive immaturity ensures that young children will often be markedly unconcerned about the sources of their knowledge, and that the act of assessing memory must therefore be viewed as an integral part of the history of that memory. Both the indicator and assessment approaches have confronted the reality of variability, or large overlap in the behaviors of children across groups.

Although investigators cannot alter the complexities of development, we can alter the types of questions we ask and the way we communicate our findings. The second challenge that has stymied progress in assessing sexual-abuse allegations is the difficulty of discussing findings in a way that preserves critical information about their value and limitations. The studies reviewed in this article suggest that attention to four principles might improve the interface between research and practice:

1. **Assessing the accuracy of reports requires comparative information.** Neither descriptions of abused children nor descriptions of the spontaneous true narratives of children who have witnessed events shed light on how to distinguish between true and false reports. This simple fact has implications both for assessment and for defining who is qualified to render judgments. As Dawes (1996) explained:

   Unhappily, many experts who testify about the characteristics of abused children claim to do so on the basis of their “experience,” but claim to have little or no experience with children who weren’t abused but who at one point claimed to have been. Hence, such experience is—not just “strictly speaking” but profoundly—irrelevant to “the determination of an action.” Ironically, such pseudo-experts often try to support their alleged expertise by claiming that almost all the children they see claiming abuse have in fact been abused. If taken seriously, such a claim should automatically disqualify them as having any experienced-based expertise in the matter of most urgent interest to the court, which is the rational determination whether or not the child has been abused as claimed. (pp. 6–7; emphasis in the original)

Unfortunately, descriptive data that are assembled for other purposes are frequently misinterpreted as conclusions become transformed, as in the game of “telephone,” through a chain of secondary articles and workshops. Consequently, authors need to take special care to flag potential inappropriate uses of their findings.

2. **The characteristics of true and false reports probably vary depending on children’s assessment histories.** Sexual-abuse assessments involve judgments at many points in time, from initial suspicions or disclosures of abuse through subsequent investigations. Throughout this process, children’s behaviors, and the characteristics of their reports, may evolve as a function of their experiences. In the best of all possible worlds, studies would select comparison groups that represent parallel points along this process. When this is impossible, authors should alert readers to possible limitations of their results.

3. **Developmental discontinuities are probably the rule rather than the exception.** There are reasons to expect that the characteristics that distinguish between accurate and inaccurate reports will vary across development. For example, numerous studies have found that 3-year-olds are much less accurate than older children (especially if they have been exposed to suggestive influences), and that preschool children have special difficulties with memory source monitoring. Older children show discontinuities as well. For example, school-aged girls may be more reluctant than other groups to discuss genital touching and less likely to act out sexually when abused (Friedrich et al., 1992). Because qualifications about age ranges are easily lost when findings are cited in secondary sources, authors should be especially sensitive to the need to preserve information about age when they summarize results from other investigators.

4. **It is an empirical question whether decision accuracy is improved by combining multiple indicators.** It is common for articles on sexual-abuse assessment to state that evaluators should tackle the uncertainty issue by taking multiple factors into account, including the consistency of a child’s report of abuse, the family context, and the presence of sexual behavior that is indicative of abuse. None of these factors has proved to be strongly diagnostic of accuracy in isolation, however, and we know of no data showing that these indicators in combination do a better job of distinguishing between true and false reports. As Wolfner et al. (1993) pointed out, the presence of varied unreliable indicators merely enhances confidence through the illusion of “convergence” (p. 9, emphasis in the original). A goal for future research is to evaluate
multivariate models of decision making. A sensible approach for designing models is to break data down by variables that are likely to increase variability, such as age, delay since the event, the type of event, and the number of prior interviews.

A critic might complain that this review sheds more shadows than light on the issue of assessing children’s reports. Yet even shadows only emerge when there is some illumination. Research to date has refined our understanding of the difficulties involved in assessment and identified key variables and procedural requirements for future studies. In the process, this research has demonstrated how the accuracy of children’s reports can be dramatically shifted by adult influences. Adults alter the a priori probability of abuse in investigated cases by their referral strategies, the content and characteristics of children’s reports by their interviewing strategies, and the outcome of individual lives by their tolerance for false-positive versus false-negative errors. It is important to note that there is no reason to assume that efforts to reduce the probability of false positives will necessarily increase the probability of false negatives. What is required is that assessors do not merely alter the stringency of their response criterion, but rather that they qualitatively change and improve the procedures by which assessments are made. Pursuit of this goal is clearly in the best interest of both abused and nonabused children and of society at large.

REFERENCES

Guidelines for psychosocial evaluation of suspected sexual 
abuse in young children. Chicago: Author.

Press release. APSAC Advisor, 6 (2), 23–24.

abuse allegations: Reliability of Criteria-Based Content Analy-
sis. Law and Human Behavior, 17, 331–341.

Conceptual and empirical obstacles. Child Abuse & Neglect, 
17, 111–125.

Berliner, L., & Conte, J. R. (1995). The effects of disclosure and 
intervention on sexually abused children. Child Abuse & 

dren’s statements about sexual abuse: A field-based validation 
study. Unpublished doctoral dissertation, Arizona State Univer-
sity, Tempe, AZ.

disclosure process in child sexual abuse. Child Abuse & 
Neglect, 9, 881–891.

Brainerd, C. J., & Poole, D. A. (1997). Long-term survival of 
children’s false memories: A review. Learning and Individual 
Differences, 9, 125–151.

detailed dolls do not facilitate preschoolers’ reports of 
touching. Paper presented at the biennial meeting of the Society 
for Research in Child Development, Indianapolis, IN.

cried when I got my shot!” Influencing children’s reports about 
a visit to their pediatrician. Child Development, 66, 193–208.

mically detailed dolls no not facilitate preschoolers’ reports of 
a pediatric examination involving genital touching. Journal 

Bruck, M., Hembroke, H., & Ceci, S. (in press). Children’s re-
ports of pleasant and unpleasant events. In J. D. Read & D. S. 
Lindsay (Eds.), Recollections of trauma: Scientific research 

Bussey v. Commonwealth, 697 SW 2d 139 (Ky. 1984).

Developmental patterns of eyewitness responses to repeated 
and increasingly suggestive questions. Journal of Experimental 
Child Psychology, 61, 116–133.

considerations. In D. L. Schacter (Ed.), Memory distortion: 
How minds, brains, and societies reconstruct the past (pp. 91– 

A historical review and synthesis. Psychological Bulletin, 113, 
403–439.

scientific analysis of children’s testimony. Washington, DC: 
American Psychological Association.

tions of sexual abuse: Forensic and scientific issues: A reply to 
commentators. Psychology, Public Policy, and Law, 1, 494– 
520.

Repeatedly thinking about a non-event: Source misattributions 
among preschoolers. Consciousness and Cognition, 3, 388– 
407.

The possible role of source misattributions in the creation of 
false beliefs among preschoolers. International Journal of Clin-
cial and Experimental Hypnosis, 42, 304–320.


evaluating children’s reports of sexual abuse: Results from a 
survey of professionals. American Journal of Orthopsychiatry, 
61, 428–437.

statements and behaviors: Role in identifying sexually abused 

child sexual abuse: Which cases are accepted? Child Abuse & 
Neglect, 18, 663–677.

Date, A. (1996). On maintaining an empirical orientation in an
"alternate ways of knowing" world. *Behavior Therapist, 19,* 86.


Hornor, T. M., Gyer, M. J., & Kalter, N. M. (1993b). The biases of
Poole, D. A., & Lindsay, D. S. (1996, June). *Effects of Parental Suggestions, Interviewing Techniques, and Age on Young Child-
dren's event reports. Paper presented at the NATO Advanced Study Institute, Recollections of Trauma: Scientific Research and Clinical Practice, Port de Bourgenay, France.


