CHAPTER 11: FORENSIC CHILD SEXUAL ABUSE EVALUATIONS

ACCURACY, ETHICS, AND ADMISSIBILITY

STEVE HERMAN

Forensic evaluators base their judgments about the validity of allegations or suspicions of child sexual abuse (CSA) on two types of evidence: (1) hard evidence such as perpetrator confessions, medical evidence, photographs or videos of the abuse, and other physical evidence and (2) soft psychosocial evidence. Considerable controversy exists with regard to which psychosocial case characteristics, if any, actually constitute probative evidence that either supports or weakens support for an allegation of sexual abuse. However, putting aside the question of probative value for the moment, here are some examples of case characteristics that many evaluators actually consider to be psychosocial evidence that is relevant to their judgments about the validity of CSA allegations:

- The contents of the child’s verbal statements. For example, does the child clearly report sexual abuse during investigative interviews? Is the child’s report consistent with what is known about typical sequences of events in CSA or does it include elements that are bizarre, improbable, or impossible?
- The narrative qualities of a child’s verbal report of sexual abuse and the child’s nonverbal behavior. For example, is the child’s narrative logical and are core elements consistent? Does the narrative include realistic contextual details? Is the child emotional or unemotional during investigative interviews?
- The psychosocial context in which the child’s report of abuse emerged. For example, was the child’s first report made spontaneously or only after prolonged questioning by adults who suspected that abuse had occurred? Did the child initially deny that abuse had occurred?
- The child’s psychosocial history, including any significant changes in the child’s behavior. For example, did the child begin engaging in unexpected and age-inappropriate sexual behaviors after the alleged onset of the abuse?
The psychosocial history and behavior of other parties involved in the case, and the contents, narrative qualities, and historical contexts of statements made by these other parties.

If there is clear and convincing hard evidence that either corroborates or contradicts an abuse allegation, then the soft psychosocial evidence is superfluous for the purpose of judging the validity of an abuse allegation (Horner & Guyer, 1991b). However, empirical research (summarized in Table 11.1) indicates that, in a substantial minority of forensic CSA evaluations, there is no strong independent evidence to either corroborate or contradict verbal reports of sexual abuse made by children, which means that judgments about the validity of abuse allegations in these cases are based primarily or solely on the psychosocial evidence.

### Table 11.1 Rates of Verbal Reports of Sexual Abuse, Evidence, and Substantiation Decisions in Forensic Child Sexual Abuse Evaluations

<table>
<thead>
<tr>
<th>Citation</th>
<th>N</th>
<th>REPORTS (R)</th>
<th>EVIDENCE (E)</th>
<th>SUBSTANTIATIONS (S)</th>
<th>R) –E</th>
<th>S) R) –E</th>
<th>S) R) –E</th>
<th>S) –R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elliott &amp; Briere (1994)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>399</td>
<td>.58</td>
<td>.30</td>
<td>.62</td>
<td>.38</td>
<td>.33</td>
<td>.20</td>
<td>.10</td>
</tr>
<tr>
<td>DiPietro, Runyan, &amp; Fredrickson (1997)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>179</td>
<td>.47</td>
<td>.35</td>
<td>.63</td>
<td>.21</td>
<td>.18</td>
<td>.26</td>
<td>.18</td>
</tr>
<tr>
<td>Dubowitz, Black, &amp; Harrington (1992)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>99</td>
<td>.60</td>
<td>.37</td>
<td>.64</td>
<td>.33</td>
<td>.26</td>
<td>.26</td>
<td>.11</td>
</tr>
<tr>
<td>Total N and weighted means</td>
<td>677</td>
<td>.55</td>
<td>.32</td>
<td>.62</td>
<td>.33</td>
<td>.28</td>
<td>.22</td>
<td>.12</td>
</tr>
</tbody>
</table>

Note: R) –E is the proportion of all cases evaluated that included both a verbal report during formal interviews and no corroborating evidence; S) R) –E is the proportion of all cases evaluated that included a verbal report, no corroborating evidence, and in which evaluators judged the allegations to be likely to be true (substantiated); S) R) E is the proportion of all cases evaluated that included a verbal report, corroborating evidence, and in which evaluators substantiated; S) –R is the proportion of all cases evaluated that included no unrecanted verbal report during formal interviews and in which evaluators substantiated.

<sup>a</sup>This study only provides enough information to determine that the report rate (R) was between .52 and .64; the midpoint of this range, .58, is used as an approximation to R. Similarly, .33 <= R) –E <= .44; the midpoint is .38. Corroborating evidence in this study included medical evidence, confessions, witnesses, and other strong independent evidence.  
<sup>b</sup>The report of the study only provides enough information to determine that .12 <= R) –E <= .31; the midpoint of this range, .21, is used as a reasonable approximation to the actual report rate. Similarly, .09 <= S) R) –E <= .28; the midpoint is .18. Corroborating evidence reported in this study included only medical evidence.  
<sup>c</sup>Full data was available for 99 out of the 132 children in the study. Corroborating evidence reported in this study included only medical evidence.

In the studies summarized in Table 11.1, researchers examined a total of 677 forensic CSA evaluations performed by mental health, social work, and medical professionals. These are the only available empirical studies that provide sufficient information to estimate the joint distribution
of three variables: (1) verbal reports of sexual abuse by children during investigative interviews, (2) the presence or absence of corroborative evidence, and (3) subsequent judgments about the validity of abuse allegations by forensic evaluators. Evaluators in these studies decided that the abuse allegations were likely to be true—they “substantiated”—in 62% of all of the cases they examined. The studies in Table 11.1 are consistent with the widespread perception among researchers and practitioners that the most important evidence of abuse in many cases of alleged CSA is the child’s verbal report (e.g., Brown & Lamb, Chapter 13; Goodman, Batterman-Faunce, Schaaf, & Kenney, 2002; London, Bruck, Ceci, & Shuman, 2005). In 33% of all of the cases evaluated there were uncorroborated verbal reports of abuse by children; 85% of these uncorroborated reports were judged to be likely to be true by evaluators. Almost half (45%) of all of the cases that were judged likely to be true were uncorroborated report cases.

This chapter focuses on these more difficult cases—cases in which there is an uncorroborated report of sexual abuse by a child. It is in these cases that the opinions of forensic evaluators about the validity of abuse allegations are most likely to play a decisive role. The key questions addressed in this chapter are:

- How accurate are the judgments of forensic evaluators about the validity of children’s uncorroborated reports of CSA?
- Are evaluators’ judgments about the validity of uncorroborated reports accurate enough to be used as a basis for making potentially life-altering legal decisions in child protection, civil disputes, or criminal investigations and prosecutions?
- Ethically, what role should mental health and social work professionals play in investigations of allegations of CSA and in related legal proceedings?
- Should expert witness testimony that purports to either confirm or support the validity of an uncorroborated sexual abuse allegation be considered legally admissible evidence in civil or criminal legal proceedings in the United States?
How Accurate Are Evaluators’ Judgments?

There are two general ways in which psychosocial evidence may be used to judge the validity of an allegation or suspicion of CSA: actuarial (mechanical, statistical, formal) and clinical (informal). Actuarial methods use explicit, well-defined rules for weighing data in order to estimate the probability that an allegation of abuse is true. The outcome of an actuarial procedure is generally a numerical score that is supposed to index the likely veracity of an allegation. Clinical methods rely on an evaluator’s holistic judgment of all of the data, and decision rules are usually implicit rather than explicit. Actuarial decision methods have generally been shown to produce more accurate classification decisions than clinical methods in a large body of empirical research on clinical judgment tasks in child welfare, psychology, criminology, medicine, and other fields over the past 50 years (Baird & Wagner, 1999, 2000; Blenker, 1954; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson, 1998; Hanson & Bussiere, 1996; Meehl, 1954; National Council on Crime and Delinquency, 2000).

The most well-known and well-researched actuarial method for evaluating cases of alleged CSA is Statement Validity Analysis (SVA) and its core component, Criteria-Based Content Analysis (CBCA). CBCA is described in some detail by Brown and Lamb (Chapter 13). The use of SVA/CBCA to evaluate cases of alleged CSA is common in Germany and other European countries (Ruby & Brigham, 1997; Vrij, 2005). The accuracy of CBCA has been evaluated in numerous empirical studies (see Vrij, 2005, for a review of 37 studies). Vrij concludes that SVA/CBCA credibility assessments are “not accurate enough to be admitted as expert scientific evidence in criminal courts but might be useful in police investigations” (p. 3). Reality Monitoring (Granhag, Strömwall, & Landström, 2006) is an alternative, but less well researched, actuarial method for evaluating the veracity of verbal reports.

Most decisions about the validity of uncorroborated allegations of CSA in forensic evaluations conducted in the United States, Israel, England, and other countries are made using informal clinical judgment rather than actuarial methods. There have been only a handful of empirical studies that shed some light on the probable accuracy of informal clinical judgments about the validity of allegations of CSA (Finlayson & Koocher, 1991; Horner, Guyer, & Kalter, 1992, 1993a, 1993b; Jackson & Nuttall, 1993; McGraw & Smith, 1992; Realmuto, Jensen, & Wescoe, 1990;
Realmuto & Wescoe, 1992; Shumaker, 2000). There are also a number of relevant theoretical analyses and commentaries (Faust & Ziskin, 1988; Herman, 2005; Horner & Guyer, 1991a, 1991b; Melton & Limber, 1989; Poole & Lindsay, 1998; Ziskin, 1995). Taken together, these empirical studies and theoretical analyses indicate that the reliability, validity, and accuracy of evaluators’ clinical judgments about the validity of uncorroborated allegations of CSA are low. In an earlier review (Herman, 2005), the author concluded that at least 24% of professional judgments about uncorroborated allegations of CSA are erroneous—either false positive errors (wrongful substantiations of false allegations) or false negative errors (failures to substantiate true allegations).

In that earlier analysis (Herman, 2005), it was impossible to provide even a rough estimate of the ratio of false positive to false negative errors based on data from empirical studies that were available at that time. Furthermore, all of the empirical evidence considered in that review was indirect to a greater or lesser extent: in 2005, there were no published empirical studies that directly assessed evaluators’ ability to judge the validity of actual reports of sexual abuse by children. That situation changed with the publication in 2007 of “Improving Credibility Assessment in Child Sexual Abuse Allegations: The Role of the NICHD Investigative Interview Protocol” by Hershkowitz, Fisher, Michael Lamb, and Horowitz. These researchers conducted the single most important empirical study of evaluators’ clinical judgments about the validity of children’s reports of sexual abuse that has been published to date. Unfortunately, some of the most important implications of this study are not made adequately clear in their report.

Hershkowitz et al. (2007) did not set out to evaluate the general accuracy of evaluators’ judgments about the validity of children’s reports of sexual abuse. They had a more limited objective for their study, to test the hypothesis that forensic evaluators’ judgments about the validity of children’s reports of sexual abuse would be more accurate when they were based on investigative interviews conducted using the National Institute of Child Health and Human Development (NICHD) forensic interview protocol (the 2007 revision of the NICHD protocol is reprinted in its entirety in the Appendix) than when they were based on unstructured investigative interviews.

The NICHD protocol (Brown & Lamb, Chapter 13, this volume; Lamb, Orbach, Hershkowitz, Esplin, & Horowitz, 2007; Orbach, Hershkowitz, Lamb, Esplin, & Horowitz, 2000; Sternberg, Lamb, Orbach, Esplin, & Mitchell, 2001) is one of several protocols that are used for forensic interviews with children in cases of alleged CSA. Some of these protocols, such as the NICHD protocol, provide specific wording for
questions and a fairly detailed and rigid format for interviews, while others (e.g., the Flexible Interview Protocol; Poole & Lamb, 1998) provide flexible interview guidelines, rather than a completely structured interview. Brown and Lamb (Chapter 13) provide a comprehensive overview of existing forensic interview protocols and describe each protocol’s research support. The most well-researched and well-supported interview protocol for conducting forensic interviews with children in cases of alleged CSA is the NICHD protocol. There is no other protocol that comes close to the NICHD in terms of empirical validation for the purpose of interviewing alleged child victims of sexual abuse, especially since the publication of the Hershkowitz et al. (2007) study.

The importance of Hershkowitz et al. (2007), and what distinguishes it from all prior empirical studies in this area, is that these researchers used trained and experienced forensic evaluators to evaluate transcripts of actual investigative interviews that were conducted with alleged child victims of sexual abuse. The researchers selected 12 transcripts from historical cases in which there was strong independent evidence that the child’s report was true and 12 transcripts from cases in which there was strong independent evidence that the child’s report was false. Half of the interviews in each category had been conducted using the NICHD protocol, the other half were unstructured, nonprotocol interviews. Hershkowitz et al. asked 42 well-trained Israeli “youth investigators” to read four transcripts each, one from each of the four experimental groups formed by crossing the two levels of the report status variable (true vs. false) with the two levels of the interview type variable (protocol vs. nonprotocol). Participants were asked to judge the validity of the reports of sexual abuse in each of the four transcripts they read. The 42 evaluators provided a total of 168 validity judgments.

Clinical judgment can be conceptualized as a diagnostic procedure that evaluators use to classify allegations and reports of sexual abuse into one of two or more predefined probability categories (e.g., substantiated, inconclusive, or unfounded). The accuracy of these classifications can be studied using the same methods and statistics that are used to assess other types of diagnostic tests and procedures (e.g., Mossman, 1994). There are many statistics that are used to describe, compare, and evaluate the accuracy of diagnostic tests and procedures. Many common accuracy statistics, such as the hit rate, describe an aspect of the relationship between a dichotomized diagnostic variable (e.g., a medical test result that is higher or lower than a predetermined threshold value) and a dichotomous variable that specifies the real state of the world (i.e., an illness is actually present or absent). An alternative approach to quantifying accuracy that does not depend on a single artificial dichotomization of a continuous diagnostic
variable uses signal detection theory and models diagnostic accuracy using the Receiver Operating Characteristic (ROC) methodology (Swets, Dawes, & Monahan, 2000).

A detailed statistical reanalysis and rethinking of the data presented in Hershkowitz et al. (2007) using both ROC and dichotomization approaches was conducted by the author. A supplemental technical document spelling out the assumptions and results of this analysis is available online at http://herman-research.blogspot.com. This reanalysis leads to five conclusions:

1. Study participants in Hershkowitz et al. (2007) showed no ability to discriminate between true and false reports made by children during the course of nonprotocol interviews. The hit rate (total percentage correct) for judgments about reports made in nonprotocol interviews was 49%, essentially the same as the expected chance hit rate of 50% (the expected percentage correct if judgments had been based on random coin tosses). The correlation between judgments and actual report status (true or false) was $r = .11$.

2. Study participants were able to discriminate between true and false reports made during NICHD protocol interviews with fairly high levels of overall accuracy compared with the accuracy of clinical judgments in other areas of psychology and medicine (Garb, 1998; Grove et al., 2000; Herman, 2005). The hit rate for protocol interviews was 74%—an error rate of 26%. The correlation between judgments and report status was $r = .58$.

3. Considering protocol and nonprotocol judgments together, the overall hit rate was 61%; the overall correlation between judgments and report status was $r = .35$.

4. The accuracy advantage for judgments about reports made during NICHD protocol interviews as compared with judgments about reports made in nonprotocol interviews was due solely to a reduction in the percentage of true reports that were not judged to be likely to be true—the false negative error rate—from 62% for the nonprotocol interviews to 5% for the NICHD protocol interviews. The percentage of false reports that were wrongly judged to be likely to be true—the false positive error rate—was high for both nonprotocol and protocol interviews, 40% and 48%, respectively. The NICHD protocol did not reduce the false positive error rate. For protocol interview judgments, 24% of all judgments were false positive errors, whereas only 2% were false negative errors. Put another way, for each false negative error, there were 12 false positive errors.
5. Because of the large difference in false negative and false positive rates for the protocol interviews, there is a marked asymmetry in the accuracy of evaluators’ judgments about the protocol interviews. When evaluators judged a child’s protocol interview report to be likely to be true, they were correct 67% of the time. However, when evaluators judged a child’s report to be unlikely to be true, they were correct 100% of the time.

The relatively low overall accuracy of validity judgments in Hershkowitz et al. (2007) is consistent with results from a large body of empirical research on the human detection of truth and deception. In a meta-analysis of 206 deception detection studies, Bond and DePaulo (2006) found that the average rate of correct classification was only 54% (47% for false messages; 61% for true messages), only slightly higher than the expected chance accuracy rate of 50%. In the 19 studies in that meta-analysis that compared persons who are presumed to be experts in deception detection (police officers, detectives, judges, interrogators, criminals, customs officials, mental health professionals, polygraph examiners, job interviewers, federal agents, and auditors) to laypeople, the experts had slightly lower (but not statistically significantly lower) average accuracy than the laypeople—laypeople were correct 55.74% of the time, the experts, 54.09% (p. 229).

Hershkowitz et al.’s (2007) findings are also consistent with numerous empirical studies showing that human judges often demonstrate unwarranted overconfidence in their ability to discriminate between true and false messages (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997). In Hershkowitz et al., many participants were unjustifiably confident that their judgments were correct—the average confidence level across the 168 judgments was 3.9 on a 5-point scale from very unconfident to very confident, but both accuracy and the correlations between accuracy and confidence were low (p. 106). The correlations between confidence and accuracy were $r = -.07$ for nonprotocol/true cases, $r = .07$ for nonprotocol/false cases, $r = .37$ for protocol/true cases, and $r = -.12$ for protocol/false cases. Evaluator overconfidence can compound the potential for damage from evaluator errors, because legal decision makers who rely on evaluators’ opinions may be influenced by evaluators’ high levels of confidence in their own (erroneous) judgments.

In summary, a reanalysis of empirical data from Hershkowitz et al. is consistent with earlier empirical studies and theoretical analyses. Taken together, these studies and analyses indicate that:
A substantial minority, about one third, of all forensic evaluations in cases of alleged CSA include an uncorroborated verbal report of sexual abuse by a child.

Error rates in evaluators’ judgments about children’s uncorroborated verbal reports of sexual abuse are high. The overall error rate in Hershkowitz et al. was 39%. Herman (2005) estimated that the overall error rate for these judgments is at least 24%.

False positive error rates in Hershkowitz et al. were high, even when children are interviewed using the NICHD protocol. The overall false positive error rate in Hershkowitz et al. was 44%. The use of the NICHD protocol did not reduce the false positive error rate.

The use of the NICHD protocol had a dramatic effect on the false negative error rate in Hershkowitz et al., virtually eliminating false negative errors. The false negative error rate was 62% for nonprotocol interviews, but only 5% for NICHD protocol interviews.

**Ethical Implications of False Positive Errors**

Hershkowitz et al. (2007) discuss the problem of the high false positive rate in their report: “[study participants’] inability to identify implausible statements is worrying. . . . The failure to identify implausible cases means that false allegations may inappropriately elicit child protection and law enforcement actions which have serious consequences . . . nonabused children fabricating allegations were disconcertingly often perceived as real victims by [study participants]” (pp. 107–108). However, Hershkowitz et al. do not adequately explore the real-world implications of this high false positive rate. Specifically, they tend to overstate study participants’ accuracy in judging protocol interview transcripts, as will be described.

Hershkowitz et al.’s (2007) study participants’ judgments about true protocol interview reports were correct (substantiations) 92.5% of the time. However, this fact does not have the significance that Hershkowitz et al. attach to it. It is misleading to consider judgments about true cases in isolation from judgments about false cases, describing the former as having “high levels of accuracy” (p. 108). We could ensure 100% accuracy for all
judgments about true allegations by simply substantiating all allegations or suspicions of abuse. There would be no need to interview children or even to examine any of the evidence in order to achieve a guaranteed 100% hit rate for judgments about true abuse allegations. However, what evaluators are (or should be) trying to do is to correctly classify both true and false reports, rather than simply validating true reports (cf. Horner & Guyer, 1991b).

Hershkowitz et al. (2007) also assert that: “the broader goal [of the designers of the NICHD protocol] was to help forensic professionals assess children’s statements, thereby promoting justice for abused children.” This implies that the goal of forensic professionals and forensic assessments is (or should be) “promoting justice for abused children.” This is not the goal of forensic assessments. The goal of forensic assessments in cases of alleged CSA is (or should be) to help legal decision makers evaluate the validity of allegations of sexual abuse in order to promote justice for all children and adults involved in these cases, including abused children, nonabused children who are the objects of false allegations or suspicions of abuse, and juveniles and adults accused (rightly or wrongly) of sexually abusing a child.

Some professionals argue that the goal of protecting nonabused children and falsely accused juveniles and adults from the potentially severe consequences of the substantiation of a false allegation take precedence over the goal of substantiating true allegations in order to protect abused children and punish child molesters. This is not a scientific argument about facts, of course, but a sociolegal one about social values, policies, morality, and ethics (cf. Horner & Guyer, 1991b, p. 405). Are there related societal value judgments that can help clarify this issue?

One related fundamental sociolegal value that is embodied in many countries’ criminal justice procedures is the perception that protecting the innocent from wrongful criminal convictions should have a much higher priority than the conviction of the guilty. British jurist William Blackstone (1769) put it this way, “it is better that ten guilty persons escape than that one innocent suffer.” American icon Benjamin Franklin was even more emphatic (1785/1907, p. 293), “it is better that a hundred guilty Persons escape than one innocent Person should suffer.” This is the deeply held social value behind the presumption of innocence and the high “reasonable doubt” standard of proof in criminal cases, which are both designed to protect the innocent from wrongful convictions, even at the cost of letting many guilty parties go free. This important sociolegal value is codified in judicial decisions and statutes in the United States (Coffin v. United States, 1895; In re Winship, 1970), England (Woolmington v. Director of Public Prosecutions, 1935), Israel (Mann & Weiner, 2003), and many other
countries, as well as in international criminal law (Rome Statute of the International Criminal Court, 2002).

In some cases, the erroneous substantiation of false allegations by CPS workers or mental health experts contributes directly to a wrongful criminal conviction (e.g., Rosenthal, 1995). Furthermore, anecdotal reports indicate that the investigation, substantiation, and prosecution of false allegations of sexual abuse can have disastrous effects on the lives of the children and adults involved, even if the wrongfully accused persons are never convicted of a crime (Bernet, 1997; Besharov, 1994; Bruck, 1998; Ceci & Bruck, 1995; Fincham, Beach, Moore, & Diener, 1994; Fukurai & Butler, 1994; Pillai, 2002; Rabinowitz, 2004; Richardson, 1990). Because of the risk of severe harm to innocent children and adults, there is a direct analogy between the state interest in protecting innocent children and adults from state sponsored investigations and evaluations that result in wrongful substantiations of false allegations of sexual abuse and its interest in protecting innocent citizens from wrongful criminal convictions.

There is another ethical dimension to decisions about allegations of CSA that needs to be made clearer to practitioners, policymakers, and researchers: When a sexually abused child comes to the attention of the authorities, the major harm to that child has already been done, and this harm was caused by the perpetrator of the abuse. Substantiating a true abuse allegation may protect the abused child from future abuse, prevent the perpetrator from abusing other children, and satisfy society’s need to punish child molesters; however, it may do little to mitigate any harm that has already been caused to the child. Ethically, if the available evidence is insufficient to substantiate a true abuse allegation, this is not the fault of the evaluator or investigators, assuming they made diligent, but unsuccessful, attempts to corroborate the allegations.

The ethical implications are very different when an evaluator substantiates a false allegation, possibly by creating or reinforcing a false report of sexual abuse through overzealous, suggestive questioning. When that happens, the evaluator becomes a primary agent of serious harm and injustice to the nonabused child and to those falsely accused of abuse. If the evaluator is an employee or contractor for a state agency, or has a professional license from the state, then the harm inflicted is also partially the responsibility of the state and its citizens. In becoming an agent of direct harm as a result of the application of a faulty diagnostic technique that is (now) known to have a very high false positive error rate, the evaluator contravenes two key ethical principles of the helping professions, justice and nonmaleficence (American Psychological Association, 2002; Beauchamp & Childress, 1989; Lilienfeld, 2007; Sen, Gordon, Adshead, & Irons, 2007). Because of the high risks of harm to the innocent and
vulnerable, ethical CSA evaluators should not employ faulty diagnostic techniques known to have very high false positive error rates, even if this makes it impossible to substantiate some true, but uncorroborated, allegations of sexual abuse.

Note that “the faulty diagnostic technique” referred to in the previous paragraph is not the investigative interview. The “technique” referred to is the process of making inferences about the validity of abuse allegations on the basis of psychosocial evidence, including children’s reports in investigative interviews. The objection is not to investigative interviewing, but to one use that is often made of data collected in investigative interviews—to make validity judgments. Investigative interviews are critical elements of good forensic CSA evaluations, but the purpose of such interviews should be to elicit information that is designed to lead to obtaining or discovering hard evidence that either supports or contradicts an abuse allegation. For example, well-conducted, videotaped interviews can be used to elicit confessions from guilty perpetrators.

Importance of Corroboration

The findings shown in Table 11.1 are inconsistent with an idea that is often expressed as a corollary to the widespread assumption among mental health practitioners and researchers that the child’s verbal report is the key evidence in many or most cases of CSA. The corollary is the assumption that corroborative evidence is rarely available in cases of CSA. In the three Table 11.1 studies, hard corroborative evidence was quite common, available in 218 (32%) of the 677 cases evaluated; all but one of these 218 cases were judged to be likely to be true by evaluators. Put another way, there was corroborative evidence in 52% of the cases that were judged to be likely to be true by evaluators.

If hard corroborative evidence is available in approximately one third of all forensic CSA evaluations and in half of all cases judged likely to be true by evaluators, why do researchers and practitioners often describe corroboration as “rare” or “seldom found”? It may be because the primary expertise and interests of mental health researchers and practitioners is in interviewing and diagnosis, rather than in law enforcement investigations, and, consequently, mental health professionals and researchers place relatively more emphasis on the forensic interview and the child’s verbal report of abuse and tend to underestimate the prevalence and importance of corroborative evidence.
Conclusion

The publication of Hershkowitz et al.’s (2007) study may eventually prove to be a defining historical moment for research on—and for the practice of—forensic evaluations of allegations of CSA. This study provides the most direct empirical evidence to date that something is very wrong with the way that children’s reports of sexual abuse are evaluated in many countries. The analysis of Hershkowitz et al.’s findings presented here is consistent with earlier studies and conclusions about the low accuracy of judgments about uncorroborated allegations of CSA. This analysis indicates that, in the absence of corroborative evidence, forensic evaluators are (a) unable to discriminate between true and false reports of sexual abuse based on children’s reports during unstructured investigative interviews at greater than expected chance accuracy (the level of accuracy that could be achieved by making judgments based on flipping a coin) and (b) a limited ability to discriminate between true and false reports based on children’s reports during NICHD protocol interviews. Of particular concern with respect to both protocol and nonprotocol interviews are the high false positive rates: If the results of Hershkowitz et al. are generalizable, forensic evaluators are likely to judge about half of all false reports of sexual abuse by children to be likely to be true.

In view of the consistent findings of high error rates, possibly more than 24% (Herman, 2005), for evaluator judgments about the validity of uncorroborated allegations of CSA and, particularly, the high false positive error rate observed in Hershkowitz et al. (2007), expert testimony that the psychosocial evidence in a case of alleged CSA is “consistent” with or supports the validity of an allegation, would not appear to meet the criteria for legal admissibility specified in the Daubert (1993), Joiner (1997), and Kumho (1999) decisions of the U.S. Supreme Court. Daubert refers to the magnitude of the known or potential error rate as one of the criteria that should be considered by judges in order to evaluate the legal admissibility of expert testimony. It is unlikely that testimony that is based on a diagnostic procedure with a false positive error rate that is approximately 50% would be considered admissible under the Daubert standard, or under any other reasonable legal standard.

Researchers, practitioners, policy- and lawmakers cannot afford to ignore these results, which suggest that drastic reforms of current practices in forensic evaluations of allegations of CSA are necessary. Perhaps the most important conclusion of this analysis is that policymakers and legal decision makers need to clarify and reform guidelines for processing allegations of CSA in order to bring current practices more in line with our
fundamental sociolegal values. We cannot rely on individual forensic evaluators to make these important sociolegal decisions on a case-by-case basis—this must be a top-down process. In particular, investigations of alleged CSA should be refocused on the traditional law enforcement investigative priority, the search for hard corroborative (or refuting) evidence, and away from the current emphasis on evaluating the validity of children’s verbal reports. Shifting the primary investigative responsibility for CSA investigations away from mental health professionals and back to law enforcement personnel might be a step in the right direction (Cross, Finkelhor, & Ormrod, 2005; Walsh, 1993).

Echoing the words of Hershkowitz et al. about CBCA (2007, p. 101), the key conclusion of this review is that professionals’ informal clinical judgments about the validity of children’s reports of sexual abuse are currently too inaccurate to “make forensic application appropriate.” Because of the high false positive rate, these judgments pose a particularly severe risk of harm and injustice to nonabused children who are the objects of false allegations of abuse and to those who are wrongly accused of CSA. To protect these children and adults from potentially severe harm, the current widespread practice of substantiating uncorroborated allegations of sexual abuse on the basis of evaluators’ clinical judgment should be halted.

Because existing empirical studies indicate that professionals and nonprofessionals perform at equally low levels in deception detection tasks, there is no good reason to believe that legal fact finders and investigators (judges, jurors, law enforcement personnel, and prosecutors) achieve higher levels of accuracy in judging the validity of children’s uncorroborated reports of sexual abuse than do mental health, medical, or social work professionals, or that they are less subject to a false positive error bias. Making policy changes that eliminate or restrict forensic evaluators’ ability to “substantiate” uncorroborated allegations to CSA will be pointless if steps are not taken to reduce the likelihood that legal fact finders will take up the slack by making their own decisions to substantiate uncorroborated allegations. Ethical mental health experts should clearly inform individual legal fact finders and decision makers about the dangers of error inherent in such judgments.

Guidelines

Considerations and Cautions

The practical implications of the current analysis and a previous analysis (Herman, 2005) by the author are as follows:
A large body of empirical research strongly supports the assertion that structured forensic interviews with alleged child victims of sexual abuse that are conducted using the NICHD protocol reduce the impact of interviewer biases and elicit more accurate and more detailed information from children than unstructured interviews. If the results of Hershkowitz et al. (2007) are generalizable, then the use of the NICHD interview also leads to a dramatic reduction in false negative errors. Although other interview protocols and guidelines exist, none of them have the same level of empirical support as the NICHD protocol.

Intensive training and ongoing supervision are desirable in order to achieve optimal results with the NICHD protocol, or any other interview technique or protocol, for that matter. Unfortunately, adequate training and ongoing expert supervision are simply not available in many settings. Fortunately, the NICHD protocol is self-explanatory and, if carefully and rigidly followed, even by inexperienced evaluators without specific training, should reduce the impact of interviewer bias and lead to more accurate and detailed statements by children than unstructured interviews. The NICHD interview protocol is currently the gold standard for forensic interviewing of children in cases of alleged or suspected CSA. It should be used, and rigidly adhered to, in all (or almost all, cf. Brown & Lamb, Chapter 13) forensic interviews with alleged or suspected child victims of sexual abuse, even when specific training and ongoing supervision of interviewers is not available.

False positive error rates in forensic interviews are too high for these interviews to be used as the basis for making validity judgments about children’s reports of CSA. The purpose of forensic interviews should be limited to the collection of information that will be used to attempt to obtain hard evidence that either supports or contradicts an allegation of sexual abuse.

The error rates in expert judgments that uncorroborated allegations of CSA are true are too high for these judgments to be used as the basis for important legal decisions. Furthermore, although more research on this point is needed, it is doubtful that using these error-prone opinions to bolster weakly corroborated allegations will lead to more accurate decisions. No legal decisions in child protection, civil, or criminal contexts should be based solely on an evaluators’ judgment that an uncorroborated allegation of CSA is likely to be true. Legal decision makers should be extremely cautious about allowing the use of error-prone expert judgments about psychosocial evidence to bolster weakly corroborated allegations.

Expert opinions that psychosocial evidence is “indicative of,” “consistent with,” or “not inconsistent with” sexual abuse are too inaccurate to be legally admissible in civil or criminal legal proceedings under the
Daubert standard and are unlikely to meet other legal standards for the reliability of scientific testimony. *Expert testimony that purports to support or demonstrate the validity of an allegation of CSA on the basis of an analysis of psychosocial evidence is not scientifically reliable, and should be barred in civil and criminal proceedings. On the other hand, testimony that informs fact finders about (a) the research discussed in this chapter about high false positive error rates in professional judgments about the validity of uncorroborated reports of sexual abuse, (b) the generally poor performance of both professionals and laypersons in deception detection tasks, and (c) how false or distorted memories can be unintentionally created in both children and adults as a result of suggestive questioning techniques has a firm basis in reliable scientific knowledge and should be considered legally admissible under the Daubert standard and other legal standards.*

- Judgments by mental health experts that uncorroborated reports of CSA are true are subject to high false positive error rates. Neither available research nor common sense provide any good reason to believe that judgments by laypersons or professionals in other fields will be more accurate than judgments made by mental health professionals. *There is currently no good reason to believe that legal fact finders, gatekeepers, and law enforcement investigators will be more accurate than mental health professionals in judging the validity of uncorroborated allegations of CSA.*

Following are a number of additional recommended guidelines for conducting forensic evaluations in cases of alleged CSA. These guidelines are generally supported by empirical evidence and/or by expert consensus (cf. Herman, 2005):

- Children should be interviewed no more than three times, and the interviewer should be the same person across multiple interviews.
- All contacts between forensic interviewers and children should be electronically recorded from start to finish, preferably on videotape, including any interactions that precede the actual formal interview phase. Similarly, all forensic interviews or interrogations conducted with any other parties—witnesses or suspects—should also be recorded from start to finish.
- If possible, forensic child interviews should be conducted by—or at least supervised by—an advanced forensic mental health professional (a doctoral level forensic psychologist or social worker, or a forensic psychiatrist) with specialized training in interviewing children, preferably using the NICHD protocol.
- The roles of therapist and forensic interviewer should be kept separate...
and should never be performed by the same individual (except in unavoidable, rare circumstances).

- Practitioners should be familiar with the vast empirical research literature that is relevant to conducting forensic CSA evaluations, and they should be able to accurately evaluate the methodological quality of empirical studies.
- Court-appointed experts with a thorough knowledge of the research on child interviews and expertise in interviewing children, rather than partisan experts hired by one side or another, should be used whenever possible in civil and criminal cases involving allegations of CSA.
- If possible, evaluations should be conducted by multidisciplinary teams that include an advanced forensic mental health professional and a medical professional who is an expert in the medical diagnosis CSA. It may also be desirable to include CPS caseworkers, and law enforcement personnel. It may be desirable, as mentioned earlier, for law enforcement personnel to take primary responsibility for directing these investigations.
- Evaluators should always keep an open mind about the validity of abuse allegations during the entire course of an investigation. Evaluators should always attempt to collect information and data that may help to either support or weaken support for each of a set of alternative explanations for an allegation of abuse, including, but not limited to these hypotheses:

  - The child has been sexually abused by the suspected perpetrator.
  - The child has been sexually abused, but not by the suspected perpetrator.
  - The child has been sexually abused, but has denied or recanted abuse.
  - The child has not been sexually abused, but has developed false memories for events that never actually occurred.
  - The child has not been sexually abused, but is deliberately lying about being abused.
  - The child has not been sexually abused, and the allegation is based on a sincere misunderstanding of the child’s verbal or nonverbal behavior by a concerned or mentally ill adult.
  - The child has not been sexually abused, and the allegation is based on a malicious false accusation by an adult or other child who is trying to achieve some specific goal.

COMP: Set the following as an unmarked chapter footnote at the bottom of the first page of the chapter.

The author wishes to thank Irit Hershkowitz and Michael E. Lamb for their comments on an earlier draft of this manuscript.
Although some evaluators may view psychosocial evidence as “corroborating” a child’s verbal report, the term *corroborated* is used here in a narrower sense—consistent with law enforcement usage—to refer only to reports that are supported by hard evidence.

In Israel, all forensic child interviews in cases of alleged CSA are conducted by these specialized investigators (Sternberg, Lamb, & Hershkowitz, 1996). Information presented in Hershkowitz et al. (2007) and elsewhere (Hershkowitz, Horowitz, & Lamb, 2005) suggests that Israeli youth investigators are better trained and more closely supervised than most Child Protective Services (CPS) caseworkers who investigate allegations of CSA in the United States (cf. Lawlor, 1998; Shumaker, 2000).

References


Franklin, B. (1907). The writings of Benjamin Franklin (Vol. 9). New York: Macmillan. (Original work published 1785)


*In re Winship, 397 U.S. 358 (1970).*


Woolmington v. Director of Public Prosecutions, A.C. 462 (1935).