Interviewing children vs. tossing coins: Accurately assessing the diagnosticity of children’s disclosures of abuse

Thomas D. Lyon
Elizabeth C. Ahern
Nicholas Scurich

University of Southern California

*Journal of Child Sexual Abuse, forthcoming*
Revised July 25, 2011.

Authors’ note: Preparation of this chapter was supported in part by National Institute of Child and Human Development Grant HD047290. Correspondence concerning this article should be addressed to Thomas D. Lyon, University of Southern California Law School, University of Southern California, 699 Exposition Boulevard, Los Angeles, CA 90089-0071. Electronic mail may be sent to tlyon@law.usc.edu.
Abstract

We describe a Bayesian approach to evaluating children’s abuse disclosures, and review research demonstrating that children’s disclosure of genital touch can be highly probative of sexual abuse, with the probative value depending on disclosure spontaneity and children’s age. Several contributors to the Kuehnle & Connell (2009) volume understate the probative value of children’s disclosures by: confusing the probability of abuse given disclosure with the probability of disclosure given abuse, assuming that children formally questioned about sexual abuse have a low prior probability of sexual abuse, misstating the probative value of children’s abuse reiteration when formally questioned, and confusing the distinction between disclosure and non-disclosure with the distinction between true and false disclosures. We review interviewing methods that increase the probative value of disclosures, including interview instructions, narrative practice, non-contingent reinforcement, questions about the perpetrators’ and other adult’s statements, and questions about the child’s reactions to the alleged abuse.

Keywords: Sexual Abuse, Interviewing, Disclosure, Bayes’ Theorem,
Interviewing Children p. 3

Introduction

Because the child victim tends to be the only eyewitness to sexual abuse (other than the perpetrator), the child’s report is an extremely important piece of evidence in any abuse case. Researchers have identified a number of ways in which child interviewing can be improved, both by increasing the number of true details in reports of children who have been abused and by decreasing the likelihood that children who have not been abused make false reports. We will argue that children’s disclosures of abuse can be highly probative of abuse, particularly when they are elicited using techniques supported by research.

Kuehnle and Connell (2009) brought together a diverse group of scholars with widely varying views regarding the potential to distinguish between true and false suspicions of child sexual abuse. The first three chapters, authored by David Faust and his colleagues, took a Bayesian approach to understanding the process by which evaluators attempt to determine if abuse has occurred (Bridges, Faust, & Ahern, 2009; Faust, Bridges, & Ahern, 2009 a & b). They painted a pessimistic picture regarding the ability of research on abused and non-abused children to increase the diagnostic accuracy of sexual abuse assessments. We agree that the Bayesian approach is an excellent framework for understanding the difficulties of evaluating abuse allegations. However, we will argue that Faust and colleagues’ examples may lead one to underestimate the value of disclosures of abuse.

Chapters by Brown and Lamb (2009) and Herman (2009) discuss the National Institutes of Child Health and Human Development (NICHD) protocol for interviewing children about abuse. Brown and Lamb (2009) reviewed evidence that the NICHD protocol increases the quantity and quality of information that children questioned about abuse provide. We agree, and will elaborate on interviewing methods that can increase the diagnosticity of disclosures. Herman (2009) argued that the NICHD protocol does a poor job of distinguishing between true and false allegations of abuse. We believe that Herman understated the evidentiary value of a disclosure of abuse and the advantages of the NICHD protocol.

First, we will present a primer on Bayesian thinking about probabilities. This is intended to introduce the subject to novice readers. Second, we’ll critique Faust and colleagues’ discussion of the probative value of sexual abuse indicators. We’ll explain how they fall prey to the inverse fallacy in their argument about the limited probative value of indicators. We’ll also argue that they underestimate the likelihood that children seen for sexual abuse assessment have been abused, because of the likelihood that assessment is triggered by disclosure. Third, we’ll discuss the probative value of children’s disclosures of genital touch, and argue that such disclosures often have very high probative value, taking into account the child’s age and the type of questions asked in assessing disclosure. Fourth, we’ll discuss innovations in interviewing that increase probative value, including interview instructions and open-ended narrative practice. Fifth, we’ll take issue with Herman’s pessimistic conclusions about the probative value of children’s disclosures.

The Bayesian approach: understanding how to assess evidence of abuse

Before we discuss other authors’ treatment of the diagnosticity of disclosures, it is important to describe the Bayesian approach. A mainstay of probability theory is Bayes’
Bayes’ theorem formally prescribes the extent to which one ought to update one’s belief in a given proposition in light of receiving new information. According to the theorem, one updates one’s belief by multiplying the prior probability of the proposition by the value of the new information. Let’s imagine that we’re attempting to determine the probability that a child was abused, and that the new information is “this child disclosed sexual abuse in a forensic interview.” The prior probability is the probability that the child was abused based on all we knew about the child before considering the child’s disclosure. If the child was randomly selected from the population to be questioned about sexual abuse, then the prior probability would be equal to the “base rate” of abuse among children in the population. Faust and colleagues (2009 a & b), for example, assumed an abuse base rate of 5%. We will return to the issue of base rates later, but for now it is important to understand that we would use the base rate as the prior probability if we had no reason to suspect abuse before we learned that the child disclosed abuse in an interview.

Bayes’ Theorem is easiest to understand if one speaks in terms of odds rather than percentages. For example, if the percentage likelihood of abuse is 50%, this is even odds; the likelihood the child was abused (50%) is equal to the likelihood that the child was not abused (50%). Odds can be expressed as a ratio: even odds would be 1:1.

It is not very difficult to convert percentages to odds ratios. The percentage likelihood is the first value in the ratio and 100 minus the percentage likelihood is the second value in the odds ratio. Hence, if the likelihood of abuse is 50%, then this is the first value in the odds ratio, 50. The second value is 100 – 50 or 50. So for a 50% likelihood, the odds ratio is 50:50, which simplifies to 1:1. If one assumes a base rate of 5%, then the odds would be 5:(100-5) or 5:95, which simplifies to 1:19.

The new information’s value (which can be called the probative value or the diagnosticity) is typically quantified as a likelihood ratio. In order to calculate the probability that abuse occurred after learning that the child disclosed abuse in an interview, one multiplies the prior probability of abuse by the diagnostic value or likelihood ratio of the disclosure. The prior probability should be expressed as odds, so if the prior probability is 50%, then the prior odds are 1:1, and if the prior probability is 5%, the prior odds are 1:19.

The likelihood ratio is the ratio of the true positive rate and the false positive rate. The true positive rate in our example is the likelihood that children disclose abuse when they have been abused. The false positive rate is the likelihood that children disclose abuse when they have not been abused. In other words, the likelihood ratio will be the proportion of abused children who disclose abuse divided by the proportion of non-abused children who disclose abuse. For example, assume that 40% of abused children disclose abuse, whereas 2% of non-abused children falsely claim abuse. The likelihood ratio for disclosure is the proportion of abused children who disclose abuse (40%) divided by the proportion of non-abused children who disclose abuse (2%), or 20.

1 The reader might want to work through some more examples. 10% is equivalent to 10: (100-10), or 10:90, or 1:9. 25% is equivalent to 25: (100 – 25) or 25:75 or 1:3. 75% is equivalent to 75: (100 – 75) or 75:25, or 3:1.
The likelihood ratio indicates how much the odds of abuse are increased by the new information (the disclosure). A likelihood ratio of 20 means that the disclosure increases the odds of abuse 20 times. Hence, in order to calculate the probability that a child who discloses abuse was in fact abused, we multiply the prior odds by the likelihood ratio. Hence, if the prior odds of abuse are 1:19, and the likelihood ratio is 20, then the probability of abuse is 20:19, which is slightly higher than 1:1 odds.

Imagine that a child is randomly selected from a group of children, only 5% of whom were abused. This would mean that the prior probability of abuse, before we interview the child, would be equal to the base rate of abuse: only 5%, or 1:19 odds. Then imagine that the child is interviewed about abuse, and discloses abuse, and the likelihood ratio of a disclosure is 20. We would then conclude that the odds that the child really was abused was 20:19, or only slightly better than 1:1. In percentage terms, it would be 51%.²

Commentators sometimes describe likelihood ratios as being strong or weak evidence. Generally, a likelihood ratio of 3 is considered weak, 5 weak-to-moderate, 7 moderate, 14 moderate-to-strong, and 20 strong (Wood, 1996). But even strong evidence may not be convincing if the prior odds of abuse are very low (Faust et al., 2009 a & b; Myers, 2005; Poole & Wolfe, 2009). Indeed, the example we just discussed is an illustration of this problem. We assumed that disclosure has a likelihood ratio of 20, which makes disclosure strong evidence. But because the prior odds of abuse were only 1:19, disclosure only makes it slightly more probable than not that the child was abused.

On the other hand, if one starts with a high prior odds even a likelihood ratio that is considered weak can convince one that abuse occurred. It is unusual for children to be plucked at random from the population and given a forensic interview for abuse. To the extent that evidence of abuse exists before the child is interviewed, then the prior probability of abuse is likely to be much higher than the base rate. For example, if the prior odds are 1:1, then a likelihood ratio of 3 increases the odds to 3:1 or 75%. In other words, one might be on the fence before the disclosure, but firmly convinced after the disclosure.

There are two other important points to make about the likelihood ratio. The reader will recall that we assumed 40% of abused children disclose abuse (the true positive rate) and 2% of non-abused children disclose abuse (the false positive rate). 40% is a minority of abused children. But 2% is a very small number, so even if most abused children deny abuse, disclosure can be strong evidence of abuse. Similarly, the fact that evidence is common or rare among abused children tells us very little about the diagnosticity of that evidence. Imagine that 60% of abused children experience nightmares. This may mean nothing at all, if 60% of non-abused children experience nightmares. On the other hand, imagine that 1% of abused children suffer from gonorrhea. This may be very strong evidence of abuse, if the percentage of non-abused children who suffer from gonorrhea is much less than 1%. Hence, the likelihood ratio teaches us

² To convert odds back to percentages likelihoods we divide the first number (20) by the sum of the first and the second numbers in the odds ratio (20 + 19). That is, we divide 20 by 20+19, which is 20/39 or 51%. Again, the reader will want to try some more examples. 1:1 converts to 50% because 1 divided by (1+1) is ½ or 50%. 2:1 converts to 67% because 2 divided by (2 + 1) is 2/3 or 67%. 100:1 converts to 99% because 100 divided by (100+1) is 100/101 or 99%.
that we can’t assess the diagnosticity of any piece of evidence without knowing both the true positive rate and the false positive rate (Lyon & Koehler, 1996).

A second point is that it is very important not to confuse the different terms. The true positive rate and the false positive rate are quite different. Novices will sometimes assume that they necessarily sum to one, which is false. Similarly, knowing the true positive rate does not tell us what the false positive rate is. They are calculated based on different groups of children: the false positive rate is calculated based on abused children, and the false positive rate is based on non-abused children.

It is also easy to confuse the true positive rate with the probability of abuse given disclosure. The true positive rate is the probability of disclosure given abuse. The reader should notice that this sounds similar to the probability of abuse given disclosure. Confusion between these two probabilities has been given different names: the inverse fallacy (Kaye & Koehler, 1991), transposing the conditional (Evett, 1995), the prosecutor’s fallacy (Thompson & Schumann, 1987). We’ll use the term inverse fallacy.

Coin tosses vs. tests that are 75% accurate

Now we are ready to consider the arguments made by Faust and his colleagues (Bridges, Faust, & Ahern, 2009; Faust, Bridges, & Ahern, 2009 a & b.) They purported to show that a test that is 75% accurate may nevertheless be worse than a coin toss in distinguishing between true and false allegations of abuse. If their proposition is true, it would indeed undermine one’s confidence in assessing child sexual abuse claims. However, the “75% accurate” argument suffers from two principal problems that may mislead the novice reader. First, the “75% accurate” test is actually only weakly probative, with a likelihood ratio of 3. Understanding how Faust and colleagues define “accurate” will make this clear. Second, the “75% accurate” test is in fact stronger evidence than a coin toss. Because of looseness in the meaning of the term “accurate,” Faust and colleagues fall prey to the inverse fallacy.

Faust et al. (2009a) wrote:
Assume there is an evaluative method or test with a 75% accuracy rate in separating [abused children from non-abused children]. Although 75% is far from perfect, it might well seem that the test has much to offer and could help us. (p. 10).

75% certainly sounds like a good test. The reader might assume that if a test is 75% accurate, then this means that if it says the child is abused, it is 75% likely to be correct. Surely, if it is correct 75% of the time, then it could be usable as a means of distinguishing abused children from non-abused children.

However, Faust and colleagues used the term “accuracy” in a specific and somewhat idiosyncratic way. They were describing a test which has a 75% true positive rate and a 75% true negative rate. So, if the child was abused, the evidence would be present 75% of the time (the true positive rate). If the child was not abused, the evidence would be absent 75% of the time (the true negative rate). In order to assess the value of the “75% accurate” test, one needs to know the likelihood ratio. The reader will recall that the numerator of the likelihood ratio is
the true positive rate, and the denominator is the false positive rate. The true positive rate of the “75% accurate test” is 75%. The false positive rate can easily be calculated using the true negative rate. The false positive rate of the “75% accurate” test is 25%. This means that the “75% accurate” test has a likelihood ratio of only 3.

Translating the “75% accurate” test into a likelihood ratio gives us a better picture of how useful the test is. The reader will recall that a likelihood ratio of 3 is considered weak evidence. Furthermore, a false positive rate of 25% is quite high. For example, if the evidence in question was a disclosure of abuse, then a false positive rate of 25% would mean that the questioning method elicited false claims of abuse from 25% of non-abused children. Indeed, a false positive rate of 25% ensures that the evidence will be weak; the highest possible likelihood ratio is 4, because the true positive rate cannot be greater than 100%.

Furthermore, translating the “75% accurate” test into a likelihood ratio emphasizes the need to consider the prior odds of abuse. Faust and colleagues made the correct point that if one starts with a low probability of abuse, and finds weak evidence of abuse (the “75% accurate” test), then the likelihood of abuse will remain low. They discuss two different cases, either assuming a prior probability of abuse of 17% (equal to a prior odds of 1:5) or 50% (prior odds of 1:1). In the case in which the prior probability of abuse is 17%, a positive result on the “75% accurate” test would increase the odds of abuse by three, from 1:5 to 3:5, or 38%. If the prior probability is 50%, then a positive test result would increase the odds of abuse by three, from 1:1 to 3:1, or 75%.

With respect to the 38% figure, Faust et al. (2009) conclude:

“we will be wrong far more often than we are correct: our accuracy rate will only be 38%, or worse than a coin toss, an outcome that is hardly satisfactory.” (p. 12).

This is an example of the inverse fallacy. The problem lies in the ambiguity of the term “accuracy rate.” The authors start with the assumption that the “accuracy rate” of a coin toss is 50%. But by comparing coin tosses to the 38% “accuracy rate,” the authors confuse the probability of evidence given abuse (50%, using a coin toss) to the probability of abuse given evidence (38%).

To demonstrate the inverse fallacy, consider how a coin toss can be considered both 50% accurate (likelihood of evidence given abuse) and less than 50% accurate (likelihood of abuse given evidence). Imagine that you take a group of 100 children, 10 of whom have been abused, and you toss a coin to determine who is abused and who is not abused: heads abused, tails non-abused. If you take any child who has been abused and toss a coin, the likelihood that you will call the child abused is 50%. Hence the coin appears to be 50% “accurate.” But now consider the entire group of children. By tossing coins, you will randomly select 50 of the 100 children “as abused.” The proportion of 50 children who really will have been abused will be the same as

---

3 The false positive rate is the inverse of the true negative rate: they sum to 100%. If the child is not abused and the evidence is absent, this is a true negative. If the child is not abused and the evidence is present, this is a false positive. Hence, the likelihood that the evidence is present plus the likelihood that the evidence is absent will be 100%. If the true negative rate is 75%, the false positive rate is 25%. 
the original group, 10%. Therefore, the coin toss will only be correct 10% of the time. Hence, the probability of abuse, given heads, is 10%, but the probability of heads, given abuse, is 50%. The two probabilities are quite different, and should not be confused.

By committing the inverse fallacy, Faust and colleagues are able to dismiss the use of relevant evidence. The truth of the matter is that evidence with a likelihood ratio of 3 increases the odds of abuse by 3 times. A coin toss does not affect the odds of abuse (and we do not recommend it!). If the prior probability is low, it will remain low in the face of this weak evidence. But if the prior probability is close to 50% (or 1:1) odds, this evidence can make the difference. In sum, Faust and colleagues’ argument purports to show that a test can be 75% accurate and yet worse than a coin flip. Their “75% accurate” test is only weakly diagnostic (likelihood ratio of 3) and nevertheless clearly better than a coin toss.

*Unrealistic assumptions about the likelihood of abuse among children seen for evaluation*

Faust and colleagues rightly speculate that children are typically evaluated for sexual abuse because “someone suspects abuse” (p. 25). The relevance of the suspicion, however, is very important. In their hypothetical example, Faust and colleagues discuss the use of a “screening method with robust validity” (p. 11), and assume the following numbers:

For purposes of this example, assume that the base rate for child sexual abuse in the catchment area for Clinic 1 is 5%. Thus, there are 19 non-abused children for every sexually abused child. Assume that all children known to display explicit sexual behavior, such as imitating sexual relations with dolls, are referred for sexual abuse evaluations. Assume further that explicit sexual behavior occurs in about 20% of sexually abused children and 5% of non-abused children (figures that roughly align with research estimates…).

The evidence, sexual play with dolls, has a likelihood ratio of 4 (true positive rate = 20%; false positive rate = 5%) which would be considered “weak” evidence. If one starts with a base rate of 5% (or 1:19 odds), multiplying this by a likelihood ratio of 4 leaves us with 4:19 or approximately 1:5 odds of abuse. This is only 17%. The 17% figure is the one that the authors use, in conjunction with the “75% accurate” test, to argue that coin tosses are better than abuse evaluations.

We will not challenge the assumption that 5% is a reasonable base rate for the population. As others in this special issue emphasize, the prevalence of abuse among girls is much higher than 5%, and boys substantially higher (Faller & Everson, this issue). However, we are not sure that the prevalence rate is the correct base rate. The prevalence rate is the percentage of children who are abused at any time during childhood. But many children are questioned long before their childhood ends, and therefore the likelihood they have been abused will be lower.

What we find questionable about the hypothetical is that it assumes communities screen for sexual abuse by giving children at large anatomically correct dolls with which to play. It is hard to imagine that anyone would ever undertake such an approach. First, we are not aware of any screening of children at large for sexual abuse. Second, if we did conduct population
screening, it is unlikely that we would have children play with anatomical dolls. A fairly broad consensus has emerged that observation of doll play is not a valid means of evaluating children for sexual abuse. Rather, supporters of the use of dolls advocate their use as a demonstration aid for children who are disclosing abuse (Pipe & Salmon, 2009).

The fact is that disclosure is a common, if not the most common reason, children are evaluated for sexual abuse. For example, Heger and colleagues (2002) examined records of 2400 children referred for evaluation of sexual abuse and found that 70% of the children had previously disclosed abuse.

The diagnosticity of the disclosure of genital touch

It is therefore most relevant to consider the diagnosticity of the disclosure of abuse in determining the likelihood that professional evaluations of abuse allegations are correct. The evidence suggests that children’s spontaneous reports of genital touching are strong evidence that touching has occurred, but that as the questions become more direct and potentially leading, and as younger children are questioned, the diagnosticity of a report of genital touch decreases.

For example, an oft-cited study by Saywitz and colleagues examined 72 5- and 7-year-olds girls’ memories of a pediatric examination (1991). For half of the girls, the examination included genital touch, and for the other half, the examiner substituted an examination for scoliosis. When asked free recall questions about the event one month afterwards, 22% of the girls who had been touched mentioned vaginal touch, and none of the girls who had not been touched did so. When asked a direct question about genital touch with the aid of an anatomically correct doll (“Did that doctor touch you there?” while “pointing to the doll’s vagina”), 86% of the girls who had been touched acknowledged genital touch, and 3% of the girls who had not been touched falsely claimed that they had. The results suggest that a spontaneous report of genital touch (in response to a very general question about one’s interactions with a person) is very strong evidence. We might say that the likelihood ratio is infinitely large, although it would be safer to say that in this study recall of genital touch provided conclusive proof that touching occurred. Even a simple “yes” response to a direct question with an anatomically correct doll constituted strong evidence of genital touch (86/3 or a likelihood ratio of 29).

Steward and colleagues conducted a similar study, in which children were either touched or not touched during a physical exam (Steward, et al., 1996). However, they included children from 3 to 6 years of age, used either anatomically detailed dolls or anatomical body diagrams, and interviewed the children one day, one month and six months after the exam. Direct questions about genital touch asked with the assistance of a doll or drawing elicited relatively weak evidence of touching; the likelihood ratios ranged from 5 to 9. True positive rates ranged from 73 to 86%, whereas false positive rates ranged from 8 to 12%. In contrast, when children were asked “did the doctor touch you,” and then, if they answered yes, were asked “where were you touched,” there were no false reports of genital touch in the verbal-only interviews at any of the time periods, whereas the true positive rates ranged from 18 to 33%. Disclosure of genital touching without the use of dolls or drawings constituted conclusive evidence of touch. The Steward (1996) study found that 3-year-olds performed worse than 6-year-olds. Indeed, as younger children are tested, the false positive rates increase, making direct questions with dolls
Interviewing Children p. 10

still less probative of touching. Bruck and colleagues questioned preschool children immediately after a pediatric examination, and included direct questions with the aid of an anatomically detailed doll (Bruck, Ceci, Francouer, & Renick, 1995; Bruck, Ceci & Francouer, 2000). Some of the questions were highly suggestive, but we will only consider the questions about genital touch most analogous to those asked by Saywitz et al., and Steward et al. (i.e., “Did Dr. F. touch you here?”) Although Bruck and colleagues did not find significant age differences, their sample sizes were small, and the percentages suggested improvement with age. The 3-year-olds (mean age 2-11) performed close to chance, such that the likelihood ratio was not greater than 1. Specifically, children exhibited a 50% true positive rate and a 42% false positive rate. The 4-year-olds (mean age 4-1) exhibited a 45% true positive rate and a 14% false positive rate, which suggests a likelihood ratio of 3, or weak evidence of touching.

Critics of the Saywitz et al. (1991) and Steward et al. (1996) studies have pointed out that they did not contain a number of other suggestive elements that may occur when children are questioned about abuse, particularly when the questioner strongly suspects that abuse has occurred (Bruck & Ceci, 1996; Bruck & Ceci, 1999). Critics also focus on the false positive rate for anal touch, which are often inexplicably quite a lot higher (Bruck & Ceci, 1996). Indeed, because of the increased false alarm rate with the use of anatomical dolls or drawings, we have cautioned against their use (Lyon, Lamb, & Myers, 2009). Hence, this discussion is not intended as a defense of the use of dolls, but rather illustrates the fact that even methods criticized as unduly suggestive will often elicit reports that increase the likelihood that genital touching occurred, particularly with children older than preschool age.

To our knowledge, researchers have not asked about genital touch when using more suggestive methods. They have, however, asked questions that would strongly intimate that genital touching may have occurred. For example, Bruck & Ceci (2004) summarize one study as finding a 68% false positive rate for “misleading questions with abuse themes (e.g. “he took your clothes off, didn’t he?”)” when the questions are asked in a high-pressure interview (p. 231; citing Finnila, Mahlberga, Santtilaa, & Niemib, 2003).

Finnila and colleagues (2003) tested 4- and 8-year-old children who were either the most suggestible or least suggestible in their age group. Children participated in a 10-minute interaction with an adult male, and were questioned one week later. In the “low pressure” group children were asked questions such as, “he took your clothes off, didn’t he?” In the “high pressure” group, children were first told, “I have already spoken to the big kids and they told me that he did some bad things that he shouldn't have done. Now I would like to know if you also have such a good memory and can help me, because I really need your help to find out what happened” For each question, the interviewer first told the child what other children had reported (e.g., “The other kids told me that he took their clothes off. He took your clothes off too, didn’t he?”). If the child said “yes” or “maybe” the interviewer “praised the child’s memory.” If the child said “no,” the interviewer “implied that the child’s answer might be wrong, repeated what ‘the other kids’ had said and re-asked the question.” Again, if the child now said “yes” or “maybe,” the interviewer praised the child’s memory.

Although Bruck and Ceci (2004) report a 68% false positive rate for misleading questions generally, citing the “clothes off” question as an example, Finnila and colleagues (2003) report
specific percentages for the “clothes off” question. In the “high pressure” condition, 9% of children answered “yes” to “He took your clothes off, didn’t he?” In the “low pressure” condition 3% did so. The rate is in the “high pressure” condition is significantly higher than in the “low pressure” condition, but at the same time, given the pressures inflicted upon the child subjects, remarkably low. Moreover, the “clothes off” question elicited lower rates of false positives in the “high pressure” condition than the other “misleading abuse” questions.

We would never advocate the use of the high pressure tactics, but if they have been used, an important question is whether the child’s responses are non-diagnostic. Given a 9% false positive rate, if the high pressure interview elicited a true positive rate of 60% or more, then a “yes” response would be considered moderately strong evidence. However, because Finnila and colleagues (2003) did not test a condition in which clothes were in fact removed, one cannot assess the diagnosticities of children’s “yes” responses to the “clothes off” questions.

Similar to the anatomical dolls research, the questions are likely to elicit higher error rates in younger children. Goodman and Aman (1990) interviewed 3- and 5-year-old children (some with anatomical dolls) about a ten minute interaction with a man, and asked the “clothes off” question as part of a group of suggestive questions (“He took your clothes off, didn’t he,” “He kissed you, didn’t he?” and “How many times did he spank you?”). Whereas 3-year-olds false alarmed 21% of the time, 5-year-olds false did so only 2% of the time (p. 1867). Similarly, the authors asked children more directly about genital touch (“Did he touch your private parts?” “Did he keep his clothes on?” and “Show me where he touched you”). Whereas 14% of the 3-year-olds’ responses “would be likely to raise suspicions of abuse,” 5% of the 5-year-olds responses were so coded (p. 1865). In contrast, Finnila and colleagues (2003) report no age differences. However, they did find that children’s accuracy was related to their performance on a suggestibility scale, and that scale, in turn, was related to age. Hence, by testing the effects of both the suggestibility scale and age in the same model, the effect of age may have been captured (and thus obscured) by the effect of the scale.

In sum, the fact that a child has disclosed genital touch has substantial probative value in concluding that the child was sexually abused. If the child reports abuse in free recall, this is arguably very strong evidence of abuse, and even if the child’s disclosure is only in response to more direct questions, this has some probative value. As children mature, their “yes” responses to direct questions increase in probative value.

The potential redundancy of disclosure

When children are referred for evaluation of sexual abuse, they are obviously questioned about abuse. If most referred children are referred because of a disclosure, this means that the evaluation will include both the child’s previous statements and any disclosure the child makes during the evaluation. Faust and colleagues (2009a) raised a potential problem with double-counting evidence, one that they called double-dipping. The evaluator should not assume that the child’s responses during the evaluation are independent of the child’s prior disclosure. As they write,
“[T]he problem is that once the variable or variables have been used during Phase 1 as a basis for referral, then any positive qualities they might have had for the detection of abuse are neutralized when they are reapplied in Phase 2, with the result often being to reduce judgmental accuracy.” (Faust et al., 2009, p. 29)

Faust and colleagues note that redundant information adds no incremental validity to the assessment, but may lead to overconfidence in the judgment about the occurrence of abuse because of the evaluator’s failure to recognize redundancy.

Faust and colleagues apply this reasoning to sexual behavior. Does it also apply to interviewing? In order for the new disclosure to add nothing to the evaluator’s judgment, the second disclosure must be truly redundant. However, there are a number of reasons why the evaluation interview will not be redundant, particularly if it relies as much as possible on open-ended questions. First, the child’s initial disclosure is likely to have contained few details. Children’s first reports are rarely to the authorities, and non-professionals are unlikely to conduct exhaustive forensic interviews when children first disclose. Second, if the child’s disclosure was to an interested person (such as a parent involved in a custody dispute with the accused), then there are sure to be doubts about the credibility of that person’s report. The disclosure may never have occurred at all, or may have been ambiguous and subject to misinterpretation, or may have been elicited through leading questions. Third, even if the recipient of the report was unbiased, the child’s report may have been elicited through direct questions, which would reduce the diagnosticity of the disclosure because responses to direct questions tend to be less accurate than free recall. To the extent that the evaluation interview is more open-ended and less leading than the interaction that led to the child’s initial disclosure, the diagnosticity of the evaluation interview increases. Some suggestibility research documents that errors in response to suggestive questions may perseverate or increase over multiple interviews, but this is far from inevitable: these effects emerge when each interview is itself suggestive and highly leading questions are asked (La Rooy, Lamb, & Pipe, 2009; Goodman & Quas, 2008. A large component of suggestibility is due to errors in the child’s responses rather than impairment of the child’s memory, which means that non-leading questions will not elicit repeated errors. On the other hand, repeated open-ended interviews are likely to reveal a great deal of new and accurate information (La Rooy, et al., 2009). Furthermore, the evaluation interview need not yield any new information to be valuable. The interview is likely to be what Schum and Martin (1982) called corroboratively redundant. The redundant report is corroborative of abuse because it reduces worries regarding the validity of the initial disclosure.

Interviewing methods for increasing the diagnosticity of abuse disclosures

Although the bulk of research on children’s suggestibility has emphasized the ways in which children’s accuracy can be impaired, a growing number of studies have explored means of increasing the accuracy and completeness of children’s reports. These techniques will increase the diagnosticity of disclosures of abuse to the extent that they increase the likelihood that abused children will disclose (true positives) and decrease the likelihood of false disclosures (false positives).
Interviewers who begin with interview instructions can increase the accuracy of children’s reports. It is recommended that interviewers: a) teach children that they can say “I don’t know” (e.g., Cordon, Saetermoe, & Goodman, 2005) because children may be reluctant to respond “I don’t know” to yes/no questions (e.g., Poole & Lindsay, 2001) or specific wh-questions (e.g., Memon & Vartoukian, 1996), b) teach children that they can say “I don’t know what you mean” (e.g., Saywitz, Snyder, & Nathanson, 1999; Peters & Nunez, 1999) because children frequently fail to ask for clarification (Carter, Bottoms, & Levine, 1996; Perry, McAuliff, Tan, & Claycomb, 1995; Saywitz, et al., 1999), and c) teach children that they can correct the interviewer (Saywitz & Moan-Hardie, 1994; Warren, Hulse-Trotter, & Tubbs, 1991). Interviewers are encouraged to not only explain the rules of the interview but to provide children examples because an unelaborated instruction (e.g., “it’s o.k. to say I don’t know”) is unlikely to be effective (Ellis, Powell, Thomson, & Jones, 2003; Geddie, Fradin, & Beer, 2000; Memon & Vartoukian, 1996; Moston, 1987). Furthermore, interviewers should reinforce giving an answer when one can, so that children do not over-utilize the “don’t know” or “don’t understand” responses (Gee, Gregory, & Pipe, 1999; Nesbitt & Markham, 1998; Saywitz & Moan-Hardie, 1994). Eliciting a promise to tell the truth from children has been found to increase honesty without increasing errors (Evans & Lee, 2010, Lyon & Dorado, 2008; Talwar, Lee, Bala, & Lindsay, 2002, 2004), even among maltreated children who have been coached to either falsely deny or falsely claim that events occurred (Lyon, Malloy, Quas, & Talwar, 2008).


When children are properly questioned, questions tapping free recall can elicit surprisingly large amounts of information. In forensic interviews, children’s responses to free recall prompts elicited three to five times more information than responses to more focused prompts (e.g., Lamb, Hershkowitz, Orbach, & Esplin, 2008). The key is to provide the child some guidance. Laboratory studies demonstrating very poor recall performance among younger children led some commentators to conclude that direct questions may be necessary in order to elicit abuse details (Lyon, 1999), but the limitation of those studies is that free recall questions did little more than ask “what happened,” “what else happened,” and “tell me more.” In contrast, an interviewer who asks follow-up questions in the form of “you said x; tell me more about x” or “you said x; what happened next?” elicits free recall while providing the child needed scaffolding.

Children are not accustomed to being asked open-ended questions, and they benefit from practice. A useful tool is narrative practice, in which the interviewer asks the child about an innocuous event before moving to the allegation. For example, the interviewer asks the child to “tell me everything that happened” on the child’s last birthday, and seeks elaboration through “you said x, what happened next” and “you said x; tell me more about x” questions. In the field, Sternberg et al. (1997) found that when interviewers used narrative practice during rapport-
building with open-ended rather than closed-ended questions, children provided longer and richer responses to the first substantive question about abuse, and longer responses to free recall questions throughout the interview. Laboratory research has demonstrated that children’s responses are also more accurate when narrative practice is provided before the substantive interview (Roberts, Lamb, & Sternberg, 2004).

Narrative practice can profitably be combined with non-suggestive forms of interviewer encouragement and support. Addressing the child by his or her name and providing non-contingent reinforcement (e.g., “you really help me understand”) is related to greater elaboration by the child (Hershkowitz, 2009). Another form of reinforcement is the use of facilitators, also known as back-channel responses, in which the interviewer encourages additional information through simple utterances that communicate that the interviewer is listening without taking the floor (e.g., uh huh, o.k.) (Cautilli, Riley-Tillman, Axelrod, & Hineline, 2005). Laboratory research has also found that interviewer social support, such as smiling often, using warm vocal intonations, and sitting in close proximity helps children resist misleading questions (Davis & Bottoms, 2002).

Reinforcement is non-contingent to the extent that the interviewer does not selectively reinforce desired responses. Contingent reinforcement, on the other hand, can be highly leading, particularly when the interviewer overtly praises acquiescence to yes-no questions (Garven, Wood, Malpass, & Shaw, 1998; 2000). Analogously, general reassurance about disclosing negative information increased children’s true disclosures of a minor transgression (Lyon, et al. 2008), but reassurance that specifically mentioned the transgression increased false positives (Lyon & Dorado, 2008).

We recommend that interviewers encountering reluctance inform the child that “it is really important that I know everything that happened,” which is less leading than providing specific and potentially leading reasons for disclosure (e.g., “it is really important that you tell me so I can put the suspect in jail”). Simply stating that “it is really important” is analogous to “placebic” requests that increase compliance without providing explicit justification (e.g., “May I use the Xerox machine, because I have to make copies” is more effective than “May I use the Xerox machine?”) (Langer, Blank, & Chanowitz, 1978).

Interviewers should strive to ask children to describe individual abusive events (Lamb et al., 2008). When children narrate individual events, they are less likely to provide skeletal and generic reports, and more likely to disclose idiosyncratic details, such as interruptions (e.g., the perpetrator stopped because a parent was heard coming home). Idiosyncratic details are harder to attribute to some sort of adult coaching. On the other hand, interviewers should avoid asking children to estimate the number of times abuse occurred or temporal information about the abuse, because this is likely to make true reports more difficult to distinguish from false reports. Children have difficulty providing numerical estimates in general. Children may also have particular difficulty with time and number if they are asked to recall incidents that occurred on multiple occasions over a long period of time (Wandrey, Lyon, Quas, & Friedman, in press). Moreover, children’s responses to direct questions about time and number are likely to be cursory whether their report is true or false.
It is also useful to ask children what occurred before and after the abusive contact. An adult who suggestively questions a child or who coaches the child to make a false report is likely to focus on eliciting the abusive act itself, rather than the context in which the abuse occurred. Children who experienced sexual abuse are frequently able to provide information about the perpetrator’s preparatory actions (e.g., closing the bedroom door), the immediate after-effects of the abuse (e.g., washing up, difficulty in falling asleep), and the perpetrator’s efforts to conceal what occurred (e.g., telling the child not to tell). Interviewers can also ask more direct questions about what the child and the suspect did following the abuse (e.g., “What did he do after [the touching]?” “What did you do after he left?”), and what children thought or felt about the incident (e.g., “How did you feel when he x?” “What did you think when he x?”). Children who report multiple episodes of abuse can be asked about their thoughts and feelings during early and later abuse incidents. This may elicit children’s naivete regarding the perpetrator’s intentions initially and their fearful expectations on later occasions. Children are often remarkably articulate with respect to describing their emotional and physical reactions to abuse (Lyon, Scurich, Choi, Handmaker, & Blank, under review).

Children often report that genital touching occurred during caretaking, such as toileting or bathing, which may make it difficult to distinguish between abusive and innocuous touch. In addition to asking questions about what the suspect said about the touching, which might support sexual intent, the interviewer can ask the child about what occurs when others bathe or care for the child, which may enable the interviewer to determine if the suspect’s actions were innocent or abusive.

Interviewers can better understand the social and emotional pressures on the child by inquiring into the child’s prior disclosures and the child’s reasons for disclosing (or for not disclosing sooner) (Hershkowitz, Lanes, & Lamb, 2007). Information about the disclosure can be elicited by continuing to ask “what happened next” questions until children report telling another person (Hershkowitz et al., 2007), or the interviewer can ask the child in a more focused manner “who did you tell” and “what did you say to them.” The interviewer should also ask the child about the disclosure recipient’s reactions (“what did she do/say when you told her”) and what the disclosure recipient has told the child about talking to the interviewer (“what did your mom tell you about talking to me”). The interviewer should also ask the child what the parent and other interested adults have said about the alleged perpetrator (“what has x said about y”).

The responses of the people to whom the child disclosed are very important. Children, particularly young children, are likely to disclose abuse first to a parent (Kogan, 2004; Hershkowitz et al., 2007). Children are less likely to disclose and more likely to recant when non-offending parents refuse to believe that abuse occurred or otherwise fail to support the allegation (Lawson & Chaffin, 1992; Malloy, Lyon, & Quas, 2007). On the other hand, children’s reports are often doubted because of the assertion that a parent is influencing the child to make a false claim of abuse. Hence, the parent’s reaction can play an important role in determining if the child’s report is consistent over time. The interviewer can also ask the child what the perpetrator has said about others (including caregivers) and about the abuse, as this may reveal threats and other inducements to keep the abuse a secret, and thus help to explain delays and inconsistencies in the child’s report.
In order to distinguish between events the child reports because he or she has heard about events from others and events the child has actually experienced, the interviewer can ask the child “how do you know” questions. These are called “source monitoring” questions. Although these questions are very difficult for young preschoolers, who have limited abilities to identify the source of their beliefs (Gopnik & Graf, 1988), older children can report whether they actually saw or merely heard about events. Indeed, the Sam Stone study (Leichtman & Ceci, 1995), which documented high rates of false reports in preschoolers exposed to both repeated suggestive questioning about and negative stereotyping of a stranger who briefly visited the classroom, found that the rate of false reports decreased from 72% to 44% among 3- to 4- year olds and from approximately 30% to 10% among 5- to 6- year olds when children were asked “did you see Sam Stone rip the book?” and “did you see Sam Stone soil the teddy bear?” (see also Poole & Lindsay, 2001).

Some source monitoring questions may be less effective, such as questioning whether a recipient “told” the child what happened (Bruck, Melnyk, & Ceci, 2000). Young children often confuse “tell” and “ask,” (Walker, 1999) and may not understand the implication of a question such as “did your mother tell you what to say?” The question implies that the mother coached the child, but speaking literally, a mother who tells the child to “tell the truth” or “tell them what really happened” is telling the child “what to say.” Moreover, “how do you know” questions are preferable to “did you see” questions insofar as the latter are yes/no questions and therefore less likely to elicit less accurate responses.

All of these inquiries enable the interviewer to test alternative hypotheses regarding the child’s actual experiences. Asking about other’s reactions, for example, helps the interviewer to explore the possibility that the child is either alleging or denying abuse because of pressures from adults close to the child. Suggestibility researchers have argued that “[w]hen an interviewer avoids confirmatory biases by posing and testing alternative hypotheses, the suggestive techniques do not seem to result in as many serious problems” (Bruck, Ceci, & Melnyk, 1997, p. 304). We would caution, however, that the questions designed to test alternative hypotheses should themselves be worded in as open-ended a fashion as possible (e.g., “Tell me about a time your mom gave you a bath.” “What did your mom say about talking to people?”). Questions that are sometimes recommended (e.g., “Who else beside your teacher touched your private parts? Did you mommy touch them, too?”) (Bruck, Ceci, & Hembrooke, 1998) risk eliciting inaccurate responses because of their directness.

The most widely researched guideline for conducting child interviews is the NICHD investigative interview protocol (Lamb et al., 2008). Under the NICHD protocol child interviewers administer scripted interview instructions, rapport building questions and non-leading allegation questions. International research testing the NICHD protocol demonstrates its superiority to non-protocol interviews (Lamb et al., 2008). For example, interviewers using the NICHD protocol use at least three times more open-ended and approximately half as many option-posing and suggestive prompts as they do without the protocol, considering comparable incidents involving children of the same age (Lamb et al., 2008).

The Ten-Step interview (Lyon, 2005) is a modified and simplified version of the NICHD protocol. Because of research warning that children may over-use responses they are taught are
acceptable (such as the “I don’t know” response) (Gee, Gregory, & Pipe, 1999; Nesbitt & Markham, 1999; Saywitz & Moan-Hardie, 1994), the Ten-Step includes a counterexample for each instruction. For example, the “I don’t know” instruction reinforces both a “don’t know” response (to “what is my dog’s name”) and a responsive answer (to “do you have a dog?”). Because of research demonstrating that eliciting a promise to tell the truth from children increases children’s honesty (Evans & Lee, 2010; Lyon & Dorado, 2008; Lyon, Malloy, Quas, & Talwar, 2008; Talwar, Lee, Bala, & Lindsay, 2002, 2004), the Ten-Step includes the promise.

In sum, children’s disclosures of sexual abuse can be highly probative evidence that abuse occurred. Laboratory research supports the conclusion that if a child discloses genital touching in response to free recall questions about their interactions with an individual, this is strong evidence that such touching did in fact occur. Even “yes” responses to recognition questions that would be considered inappropriately leading by many researchers (such as questioning with the assistance of an anatomically correct doll) are often strong evidence, but the strength of the evidence diminishes as the suggestiveness increases and the age of the child decreases. Laboratory research also supports the conclusion that interviewers can elicit more accurate reports from children with the use of open-ended questions, interview instructions (including a promise to tell the truth), and narrative practice rapport-building. Interview protocols that incorporate these elements are therefore likely to produce more accurate reports, in particular by reducing the likelihood that a false allegation will either be created or perpetuated by poor interviewing.

Differences between our analysis and Herman (2009)

Our conclusions are quite different than one put forward by Herman (2009), who argues that “[f]alse 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” (p. 261). Herman (2009) describes a study by Hershkowitz and colleagues (2007) that found that although trained evaluators exhibited “fairly high rates of overall accuracy compared with the accuracy of other areas of psychology and medicine” when considering true and false reports elicited using the NICHD protocol, they were unable to discriminate between true and false reports elicited with non-protocol interviews (p. 253). Moreover, for both the NICHD protocol and the non-protocol interviews, evaluators judged a large percentage of the false disclosures as true.

The problem with Herman’s analysis is that the study by Hershkowitz and colleagues (2007) did not assess the probative value of disclosures of sexual abuse. Rather, they looked at whether true disclosures could be distinguished from false disclosures. This is an important issue, but it begs the question of whether disclosures increase the likelihood that abuse occurred. The Hershkowitz et al. (2007) study only examined cases in which a child disclosed sexual abuse. It then selected equal numbers of true and false disclosures for both the NICHD protocol and non-protocol interviews. In order to assess the probative value of a disclosure, however, it is important to recognize that a disclosure is itself highly probative of abuse, and that good interviewing is less likely to elicit false disclosures than bad interviewing.

An analogy might help to make this point more clear. Imagine someone claimed that DNA tests are worthless, because true matches look the same as false matches. In true matches,
the perpetrator really did contribute the DNA sample found at the scene of the crime. In false matches, the perpetrator didn’t contribute the DNA sample found at the scene of the crime, but the test result is positive; perhaps there was a lab error (for example, we don’t have the perpetrator’s actual DNA) or perhaps we have a random match (in which someone with DNA identical to the perpetrator actually contributed the DNA sample found at the scene of the crime). If the study compared true matches to false matches, they would be very difficult (perhaps impossible) to distinguish. Indeed, if they are both actual matches, they should look identical. It would be wrong, however, to conclude that a DNA match means nothing. Rather, one would ask what is the likelihood of a true match compared to a false match? As long as true matches are much more likely than false matches, then a match is highly probative evidence.

Imagine a second DNA study was conducted following the implementation of improved procedures which ensured that the only false matches are random matches. But again, if the study compared true matches to false matches, they would be difficult if not impossible to distinguish. It would be wrong, however, to conclude that improved procedures are ineffective. Rather, one would ask whether the number of true matches compared to the number of false matches increased, thus increasing the probative value of a match.

A disclosure of sexual abuse is analogous to a match. A true disclosure is like a true match. A false disclosure is like a false match. There is good evidence that disclosure of sexual abuse is much more likely to be true than false and that improving interviewing can reduce the likelihood of false disclosures. But these virtues are not detectable if one only compares true and false disclosures.

A second problem with using the Hershkowitz et al. (2007) study to argue that child interviews are not helpful is that the determination of whether sexual abuse occurred is not solely based on the word of the child. Herman (2009) is careful to make this point with respect to the existence of corroborative evidence, including perpetrator confessions, medical evidence, and eyewitness reports. However, he argues that unless this kind of corroborative evidence (which he calls “hard” evidence) is present, a child’s disclosure (which he calls “soft” evidence) cannot suffice to substantiate an abuse report.

In a supplemental technical document (available on the internet), which includes material that did not find its way into the chapter, Herman (2008) also acknowledges that evaluators “might have interviews with the alleged perpetrator and other parties involved in the case, data about the context and manner in which the concern about possible sexual abuse first arose, data regarding the number of people who had talked with the child about suspected abuse before the recorded interviews occurred, as well as other case history information. It is possible that access to this additional information could improve decision accuracy across all cases.” Herman (2009) groups this type of information with the child’s disclosure, thus categorizing it all as “soft.” However, this is precisely the kind of information that can enable an interviewer to determine the risk that the child’s prior and current disclosure are the product of suggestion, or coercion, or insincerity. Moreover, as we have emphasized, the child interview itself should inquire into the child’s initial disclosure and the pressures that have been placed on the child to disclose or not to disclose.
Discussion

Bayesian approaches, if correctly understood, can help us understand the probative value of evidence. Children’s disclosures of genital touch often constitute strong evidence that touching did in fact occur. If a grade-school child recalls genital touch in response to free recall questions, this is particularly strong evidence. More suggestive questions asked of younger children have less probative value.

Happily, researchers have moved beyond identifying methods that undermine children’s accuracy, and have developed positive prescriptions for effective interviewing. Interviews that utilize the tools incorporated into the NICHD structured interview protocol, including instructions, narrative practice report building, and open-ended questions regarding abuse will lead to more accurate and complete reports. Furthermore, these methods can be further improved, though the use of instructions with counter-examples and a promise to tell the truth (e.g., in the Ten-Step Interview). Interviewers may elicit information from the child that helps the interviewer assess the likelihood that the child’s report has been distorted by others, including the suspect and prior recipients of the child’s disclosure.
References


