Child Abuse & Neglect 35 (2011) 287–298

Contents lists available at ScienceDirect

Child Abuse & Neglect

Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements☆

Mark D. Everson[a,*], Jose Miguel Sandoval[b,1]

[a] Department of Psychiatry, University of North Carolina at Chapel Hill, 104-A Market Street, Chapel Hill, NC 27516, USA
[b] Injury Prevention Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ARTICLE INFO

Article history:
Received 4 October 2008
Received in revised form 5 January 2011
Accepted 6 January 2011

Keywords:
Forensic child sexual abuse evaluations
Professional judgements
Child sexual abuse
Subjectivity and bias
Decision-making
Sensitivity
Specificity
Skepticism

ABSTRACT

Objectives: Evaluators examining the same evidence often arrive at substantially different conclusions in forensic assessments of child sexual abuse (CSA). This study attempts to identify and quantify subjective factors that contribute to such disagreements so that interventions can be devised to improve the reliability of case decisions.
Methods: Participants included 1106 professionals in the field of child maltreatment representing a range of professional positions or job titles and years of experience. Each completed the Child Forensic Attitude Scale (CFAS), a 28-item survey assessing 3 forensic attitudes believed to influence professional judgments about CSA allegations: emphasis-on-sensitivity (i.e., a focus on minimizing false negatives or errors of undercalling abuse); emphasis-on-specificity (i.e., a focus on minimizing false positives or errors of overcalling abuse); and skepticism toward child and adolescent reports of CSA. A subset of 605 professionals also participated in 1 of 3 diverse decision exercises to assess the influence of the 3 forensic attitudes on ratings of case credibility.
Results: Exploratory factor analysis identified 4 factors or attitude subscales that corresponded closely with the original CFAS scales: 2 subscales for emphasis-on-sensitivity and 1 each for emphasis-on-specificity and skepticism. Attitude subscale scores differed significantly by sample source (in-state trainings vs. national conferences), gender, years of experience, and professional position, with Child Protective Service workers unexpectedly more concerned about overcalling abuse and more skeptical of child disclosures than other professionals—a pattern of scores associated with an increased probability of disbelieving CSA allegations. The 3 decision exercises offered validation of the attitude subscales as predictors of professional ratings of case credibility, with adjusted R^2s for the three exercises ranging from .06 to .24, suggesting highly variable effect sizes.
Conclusions: Evaluator disagreements about CSA allegations can be explained, in part, by individual differences in 3 attitudes related to forensic decision-making: emphasis-on-sensitivity, emphasis-on-specificity, and skepticism toward child reports of abuse. These attitudes operate as predispositions or biases toward viewing CSA allegations as likely true or likely false. Several strategies for curbing the influence of subjective factors are highlighted including self-awareness of personal biases and team approaches to assessment.

☆ The Injury Prevention Research Center at the University of North Carolina at Chapel Hill underwrote the statistical analysis of this research. Much of the national conference data for this study were collected at the Beyond Finding Words Conference in Indianapolis, IN in August 2006 and in Atlantic City, NJ in August 2007.

* Corresponding author.
[1] Josè Miguel Sandoval is now at the Center for Child and Family Policy, Duke University, Durham, NC, USA.

0145-2134/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chiabu.2011.01.001

Introduction

Discrepant opinions about the validity of child sexual abuse allegations have long been a topic of concern (Corwin, Berliner, Goodman, Goodwin, & White, 1987; Ney, 1995; Poole & Lindsay, 1998). One evaluator or investigator's belief that child sexual abuse (CSA) has occurred is contradicted by a second evaluator's equally strong conviction of the allegation's falsehood. When such disagreements cannot be explained by variations in training or experience, speculation often centers on the intrusion of subjective factors, including personal bias, in the forensic process (Faller, 2003; Herman, 2005; Jackson & Nuttall, 1997).

A substantial body of research, conducted primarily in the 1990s, offers evidence of such subjectivity in CSA assessments (see reviews by Herman, 2005, 2009). The most common methodology has been to ask professionals to rate the credibility of CSA allegations based upon a review of a shared set of facts provided in a case summary. Professional judgements are reliable to the degree that professionals examining the same evidence reach similar conclusions. Subjectivity or bias is inferred to the degree that case ratings or determinations are unreliable, or are systematically related to characteristics of the professional, such as discipline, professional position or gender (Herman, 2005).

Decision-making studies have varied in their methods of presenting case evidence to study participants. Their methods have included written vignettes (Finlayson & Koocher, 1991; Jackson & Nuttall, 1993, 1997; Shumaker, 2000); videotaped child interviews (Realmuto, Jensen, & Wescoe, 1990; Realmuto & Wescoe, 1992); written transcripts of child forensic interviews (Hershkowitz, Fisher, Lamb, & Horowitz, 2007); and detailed case presentations (Horner, Guyer, & Kalter, 1993a, 1993b). In addition, McGraw and Smith (1992) conducted record reviews of Child Protective Services (CPS) files to compare professional judgements. In each of these studies, professionals having access to the same evidence widely disagreed about the credibility of the cases. Several studies have also reported ratings of case credibility varying significantly by professional discipline (Horner, Guyer, & Kalter, 1993a, 1993b; Jackson & Nuttall, 1997), years of experience (Jackson & Nuttall, 1997), the professional's gender (Finlayson & Koocher, 1991; Horner, Guyer, & Kalter, 1993a, 1993b; Jackson & Nuttall, 1997), and personal history of childhood sexual abuse (Jackson & Nuttall, 1997).

Herman (2009) has emerged as a leading critic of current forensic practice. Citing the consistency of research findings that professional judgements lack reliability and therefore validity, Herman has raised serious questions about the legitimacy and ethics of making decisions to substantiate CSA allegations without corroborating evidence. Herman defines "corroborating evidence" narrowly, to include only "hard evidence," with examples such as perpetrator confessions, definitive medical findings, and photographic evidence. He has also called for "drastic reforms of current practice," including severe restrictions on the type of evidence considered by CPS or admissible in court (Herman, 2009).

Relatively little is known about the identities of the specific subjective factors that wreak such havoc on professional judgements. Such knowledge may be essential for implementing intervention strategies to lessen the influence of subjective factors. This paper focuses on three forensic attitudes that have been proposed as possible contributors to disagreements in professional judgements. Runyan (1998) has suggested the epidemiological concepts of "sensitivity" and "specificity" to explain the differing perspectives of medical and legal professionals in regard to CSA allegations. Both terms refer to indices of diagnostic accuracy (Fletcher, Fletcher, & Wagner, 1996). Sensitivity focuses on minimizing false negative errors or errors of undercalling, while specificity emphasizes minimizing false positive errors or errors of overcalling. High sensitivity and high specificity are both desirable diagnostic goals, but achieving one often demands a trade-off from the other. Such a trade-off underlies what Runyan describes as a "clash of cultures" between the medical and legal professions in their approaches to CSA cases: The medical diagnostic process, at least in its initial phases, emphasizes sensitivity (lest a child victim go unidentified and untreated), while the legal system favors specificity (lest an innocent person be imprisoned). The question arises whether similar, inherent differences in attitudes may be prevalent in professional groups other than medicine and law. However, a more crucial question is whether individual differences in the emphasis placed on avoiding false negative errors vs. false positive errors may alter the odds that evaluators will come down on one side or the other in substantiation decisions.

The third forensic attitude of interest, "skepticism toward child disclosures," refers to beliefs about the likely truthfulness of child and adolescent claims of sexual abuse. Professionals high on skepticism approach cases with an a priori belief that a large percentage of children and adolescent reports of CSA are false. Previous research on skepticism has been limited by reliance on assessment instruments with unknown psychometric properties and on small, geographically-restricted convenience samples. Specifically, three studies have compared the level of skepticism (or conversely, belief in credibility) among assorted professional groups, including CPS personnel (Boat & Everson, 1988; Everson, Boat, Bourg, & Robertson, 1996; Saunders, 1988). In all three studies, CPS workers were found not only to be significantly less skeptical than law enforcement, but also to rank as the least skeptical among the professional groups assessed.

It is important to note that the skepticism and emphasis-on-specificity attitudes are conceptually related, but not synonymous. Individuals who score high on skepticism are likely to emphasize specificity in abuse decisions. However, the inverse is not necessarily true. One can have significant specificity concerns without having a bias against believing the abuse accounts of children. Alternatively, emphasizing sensitivity over specificity does not require a belief that all, or even most, child disclosures of abuse are true. We will use exploratory and confirmatory factor analysis and regression analysis to empirically examine the distinctiveness of the three forensic attitudes, as operationalized.

The purpose of this study is to determine whether the three forensic attitudes—emphasis-on-sensitivity, emphasis-on-specificity, and skepticism—influence professional judgements and thus contribute to evaluator disagreements in CSA

Table 1
Demographic characteristics of overall sample and three decision exercise subsamples.

Demographic	Overall sample		Decision exercise subsamples					
			Case vignettes		Mock evaluation		Record review	
	n	%	*n*	%	*n*	%	*n*	%
Source								
In-state trainings	557	50.4	239	100	114	100	0	0
National conferences	549	49.6	0	0	0	0	252	100
Professional position[a]								
CPS	415	37.5	205	85.8	13	11.4	84	33.3
MH	141	12.7	9	3.8	82	71.9	16	6.3
LE	225	20.3	15	6.3	1	.9	66	26.2
VA	55	4.9	2	.8	1	.9	6	2.4
ATT	66	5.9	2	.8	0	0	27	10.7
CFE	39	3.5	1	.4	7	6.1	7	2.8
CFI	165	14.9	5	2.1	10	8.8	46	18.2
Experience								
0–2 yrs	267	24.5	98	42.0	15	13.4	46	18.5
3–10 yrs	531	48.8	113	48.0	52	46.4	130	52.4
>10 yrs	290	26.7	24	10.2	45	40.2	72	29.0
Gender								
Male	284	26.2	54	22.7	25	22.1	64	25.4
Female	800	73.8	184	77.3	88	77.9	188	74.6

[a] CPS = Child Protective Services, MH = mental health, LE = law enforcement, VA = victim advocate, ATT = attorney, CFE = child forensic evaluator, CFI = child forensic interviewer.

assessments. The specific objectives include: (a) developing a brief, psychometrically-sound instrument for assessing the three identified forensic attitudes; (b) determining whether the forensic attitudes differ by professional position, years of experience, gender, or sample source (i.e., in-state trainings vs. national conferences); and (c) assessing the influence of the three attitudes on case credibility ratings, using subsets of the study sample participating in three diverse decision exercises.

Method

Participants

Participants included 1106 professionals in the field of child maltreatment. Slightly over half were recruited at 30+ in-service or continuing education trainings throughout the State of North Carolina from 2005 to 2008. The remaining 49% were recruited at 3 national professional conferences within the same time period. Participation rates, estimated from the percent of professionals in attendance who submitted completed attitude scales, ranged from approximately 70% of those at conference plenary sessions to 100% of professionals in workshops or other trainings. Advantages of this recruitment method included high participation rates and multi-state representation of professionals from diverse backgrounds, disciplines, and experience levels.

Table 1 provides a demographic comparison of the overall sample as well as the three decision exercise subsamples. Seven professional positions or job titles are represented, including Department of Social Services personnel, primarily in Child Protective Services (CPS); mental health professionals, primarily self-identified as therapists or counselors (MH); law enforcement personnel (LE); victim advocates (VA); attorneys, almost exclusively prosecutors or Department of Social Services agency attorneys (ATT); child forensic evaluators, primarily psychologists self-identified as evaluators (CFE); and child forensic interviewers from a variety of settings including child advocacy centers (CFI). CPS workers comprised 37.5% of the overall sample with LE being the second largest subgroup at 20.3%. Almost 49% of participants in the overall sample reported between 3 and 10 years of professional experience with abused children, with approximately 25% reporting 2 years or less and 26.7% reporting more than 10 years experience.

Subsamples for the Case Vignettes and Mock Evaluation exercises were recruited exclusively during in-state trainings while the Record Review subsample was recruited at a national conference plenary session. Professional roles varied widely across decision exercise subsamples from 85% CPS in the Case Vignettes exercise to 71.9% MH professionals in the Mock Evaluation exercise. Case Vignettes participants were less experienced than participants in the other 2 decision exercises, with 42% reporting 2 or less years of experience.

Measures

Scale development. The emphasis-on-sensitivity attitude was conceptualized to include 3 core beliefs: (a) Many CSA victims are reluctant to disclose their abuse; (b) Many true cases of CSA are not believed or substantiated; (c) It is better to err on the side of the child when cases are unclear. Similarly, the emphasis-on-specificity attitude was seen to include 3 beliefs: (a) Many allegations of CSA are untrue; (b) Many false allegations are mistakenly believed and substantiated; (c) It is better

to err on the side of the alleged perpetrator unless the evidence is definitive. An initial pool of 20+ items was developed for each attitude and pared through pilot-testing on the basis of item clarity, response variation, and internal reliability. The skepticism attitude was conceptualized as a continuum of doubts about the truthfulness of child and adolescent disclosures of CSA, varying with the victim's age and gender. The current skepticism subscale was adopted from Everson et al. (1996).

Scale description. In final version, the Child Forensic Attitude Scale (CFAS) is a 28-item, self-administered and self-scored attitude survey that is designed for use both as a research instrument and as a self-assessment tool for training purposes. The CFAS is composed of 3 subscales corresponding to the 3 forensic attitudes of interest, but re-named for mnemonic purposes: Fear of Undercalling Abuse (F-Under); Fear of Overcalling Abuse (F-Over); and Skepticism of Child Reports (Skep).

The F-Under and F-Over subscales each have 10 items, plus a primer item which is unscored. Participants were asked to rate agreement on a *5*-point scale from "Strongly Agree" to "Strongly Disagree." Sample items for the F-Under subscale include:

- Failing to believe or substantiate true cases of child sexual abuse is a common error in our field.
- It is better to err on the side of the child in sexual abuse investigations even if some cases are substantiated that should not be.

Sample items from the F-Over subscale include:

- Substantiating false allegations of child sexual abuse is a common error in our field.
- Accusing an innocent person of child sexual abuse is potentially so damaging that it is better to err on the side of the alleged perpetrator unless evidence of guilt is pretty clear.

The Skep subscale included 6 items that are identical in format, except for variations in child gender and age. Participants rated each Skep item on a 5-point scale that was anchored with specific percentages (less than 25%; 25–49%; 50–79%; 80–94%; 95% or more).

The first item was:

- Of 100 girls between the ages of 3 to 5 who disclose being sexually abused, how many would be <u>true</u> victims of sexual abuse?

The CFAS also includes brief questions on professional demographics, including job title and years of experience.

Psychometric properties. Cronbach's alphas ($N = 910$) for the subscale total scores were .69 for F-Under, .84 for F-Over, and .90 for Skep. Test–retest reliability based on a sample of 42 drawn from training workshops ranged from .75 for F-Under, .85 for F-Over, and .81 for Skep over a mean period of 4 weeks. Test–retest reliability using a convenience sample of 30 colleagues ranged from .78 for F-Under, .83 for F-Over, and .79 for Skep over a mean period of 6 weeks.

An exploratory factor analysis (EFA) using a sample of 910 was conducted to identify the factor structure of the 26-scored items. An unweighted least squares method with promax (oblique) rotation extracted 4 factors (see Table 2). Factor I was comprised of the 6 items from the Skep subscale, all with loadings of .53 or above. Factor II was comprised of the 10 items from the F-Over subscale, all with loadings of .39 or above. Factors III and IV were comprised of 6 items and 3 items from the F-Under scale, respectively. (One item, Under 2, failed to load significantly on any factor and was deleted from further analyses.) The items of Factor III cluster around the common theme that many true CSA cases are missed or not substantiated. This factor will be given the label, F-Under1 or "Missed Cases Common." The 3 items of Factor IV share the common theme that it is better to err on the child's side in CSA investigations. The label for this factor will be F-Under2 or "Err on Child's Side." As shown in Table 2, only 1 non-F-Under item loaded above .30 on these 2 factors. The factor analysis therefore indicated 4 distinct factors that mapped closely with CFAS subscales, with minimal item overlap. To facilitate comparisons, Cronbach's alphas for the 2 F-Under factors (.66 and .57, respectively) were adjusted using the Spearman-Brown Prophecy formula, to correspond to the original subscale length of 10 (Nunnally & Berstein, 1994). The adjusted alpha coefficients were .76 for F-Under1 and .81 for F-Under2.

Despite the EFA results, the moderately high correlation of .50 between the F-Over and Skep total scores raised concerns whether the 2 subscales represented a single underlying attitude dimension or 2 factors or dimensions as suggested by the EFA. As a result, a confirmatory factor analysis (CFA) was performed on a different sample of 264 professionals from the EFA sample, using the Mplus, version 6.0 statistical program with weighted least squares estimation of parameters. The 4-factor model from the EFA with F-Over and Skep as separate factors was compared with a 3-factor model combining F-Over and Skep as a single factor. The chi-square test for difference testing was highly significant, indicating that the four factor model provided a better fit than the 3 factor model ($\chi^2 = 106.5$, df = 1, $p < .0001$). Thus, F-Over and Skep were confirmed as separate factors.

Table 2
Exploratory factor analysis of f-under, f-over, and skep subscales.

Item number		Factor I	Factor II	Factor III	Factor IV
Skep 5	% true victims among 6–12-yr-old disclosing boys	**.94**	−.01	.02	.04
Skep 4	% true victims among 3–5-yr-old disclosing boys	**.91**	−.06	.11	−.15
Skep 1	% true victims among 3–5-yr-old disclosing girls	**.86**	−.03	.09	−.14
Skep 2	% true victims among 6–12-yr-old disclosing girls	**.83**	.08	−.04	.07
Skep 6	% true victims among 13–17-yr-old disclosing boys	**.73**	.03	−.07	.11
Skep 3	% true victims among 13–17-yr-old disclosing girls	**.53**	.19	−.16	.16
Over 7	Substantiations of false cases are common	−.01	**.70**	.09	−.09
Over 5	Many people wrongly convicted of CSA	.02	**.68**	.10	−.02
Over 4	Children falsely reporting is a common problem	.10	**.66**	−.00	.10
Over 11	I worry a lot about false allegations being believed	.02	**.62**	.08	−.10
Over 3	Many are too quick to believe children	−.05	**.59**	−.06	−.01
Over 9	Many false substantiations because of interviewer errors	.05	**.57**	−.10	.06
Over 10	Substantiating false case more frequent than missing true case	−.02	**.57**	.09	−.04
Over 8	More skepticism would reduce substantiation errors	.08	**.54**	.09	−.15
Over 2	At least 1/3 to 1/2 of CSA allegations are untrue	.07	**.53**	−.16	.11
Over 6	Better to err on side of alleged perp	−.05	**.39**	.02	−.33
Under 2	Except possibly for custody cases, most allegations true	−.24	−**.25**	.21	.12
Under 4	Failing to believe true case is common	.03	−.02	**.64**	−.01
Under 7	Failing to substantiate true case more frequent than substantiating false case	−.03	−.16	**.50**	−.05
Under 3	Over 50% of child victims too traumatized to disclose	−.01	.03	**.47**	.16
Under 11	I worry a lot about true allegations being doubted	.03	−.06	**.43**	.11
Under 5	Most sexually abused children never disclose	.03	.24	**.42**	.13
Under 8	Many true cases missed due to over-cautious interviewers	−.06	.27	**.38**	.15
Under 10	Better to err on side of child	−.07	−.04	.07	**.55**
Under 6	More harmful to miss true case than substantiate false case	.07	−.07	.14	**.50**
Under 9	Failure to believe true disclosure is most damaging error	−.01	−.03	.15	**.46**

Note: Skep 5 = item #5 of Skep subscale; Over 7 = item #7 of F-Over subscale.
The highest loading for each item is in bold-face type.

Procedures

The research was described to participants as an assessment of attitudes related to decision-making in cases of alleged abuse. At one conference plenary session, attendees were solicited directly as research participants to complete the attitude scale and Record Review exercise. In all other cases, including one conference plenary session, participants completed and self-scored the CFAS as part of a conference presentation or training workshop. They were then asked for the use of their completed scales (and Case Vignette or Mock Evaluation ratings) for the research. Participants conveyed consent by checking a "permission to use for research" box on the first page of the scale and by submitting the completed scale. This study was approved by the Institutional Review Board, Office of Human Research Ethics, at the University of North Carolina at Chapel Hill.

The 3 decision exercises were designed to assess the validity of the CFAS as a measure of forensic attitudes. Each exercise involved professionals completing a decision-making task after taking the attitude scale. The exercises were structured to vary widely in methodology to increase their generalizability to real world circumstances. The exercises also differed in the amount of case information provided to participants for their credibility ratings. The case summaries used in the Case Vignettes exercise were very limited. More extensive case information was available for the Mock Evaluation. The Record Review exercise offered the most comprehensive case summary, plus explicit case facts suggesting possible alternative explanations for the abuse allegation. The Case Vignettes exercise was completed immediately after the survey. The Mock Evaluation and Record Review exercises were completed 1 day after the survey. Brief descriptions of each exercise are as follows.

Case vignettes exercise (N = 239). Four case vignettes from the Jackson and Nuttall study (1993, 1997) were used with author permission. The 4 were selected from the 16 vignettes employed in the original study to represent a wide range of credibility ratings and included the vignettes with the highest and lowest ratings and 2 in the median range. Each vignette was composed of a 1-page case description, including the initial allegation, relevant case characteristics, and brief summary of statements by the alleged victim and perpetrator. The alleged victims in the vignettes were an 8-year-old boy, a 13-year-old boy, a 4-year-old girl, and a 13-year-old girl. As was true in the original research, participants were asked to rate the credibility of each vignette: "How confident are you that the sexual abuse did occur?" on a 6-point scale ranging from "very confident it did not occur" to "very confident it did occur." The credibility ratings for the 4 vignettes were summed to create a total credibility score.

Mock evaluation exercise (N = 114). As part of a 2-day workshop on forensic evaluations of alleged abuse, participants spent 3 h working in multidisciplinary groups of 4 on a mock evaluation based upon a real case of alleged abuse. The case involved allegations of sexual abuse of a 4-year-old by her biologic father, resulting in a parental divorce and legal battles over

visitation. The small groups were given the task of designing and conducting a forensic evaluation with the goal of producing a set of written conclusions and recommendations. The evaluation was a reiterative process in which groups requested and reviewed specific data packets (e.g., CPS records, medical file, interviews with child), before asking for additional data packets as the case unfolded. Up to 30 different data packets were available for review. After the groups had completed the exercise, participants were asked to set aside their group's conclusions about the validity of the case and to complete ratings of what they believed "in their heart of hearts." Specifically, participants were asked to rate the likelihood that the allegation of CSA was true on a 7-point scale.

Record review exercise (N = 252). This exercise was designed as a mock case record review involving a 4-year-old who made allegations that she had been fondled by her 13-year-old cousin. The record provided to participants was a detailed, 8-page, single-spaced summary of the initial allegation report, case background, medical examination, collateral interviews, and interviews with the alleged perpetrator. It also included a partial transcript of a forensic interview with the alleged victim. Participants were asked to rate the probability of the allegation being true on a 10-point scale.

Data analysis

A total score for each of the 4 attitude subscales, F-Under1, F-Under2, F-Over, and Skep, was computed by summing the corresponding CFAS items. For the demographic analyses, the total scores were converted into T-scores with a common mean and standard deviation of 50 and 10, respectively. T-scores offer the advantage of simplified interpretation of score similarities and differences, both within and across subscales and demographic variables. An added benefit is the ease of calculation of Cohen's *d* as an effect size estimate for the difference between group means. Mean T-score differences of 2, 5, and 8 points correspond to Cohen's *d* of .2, .5, or .8 or small, moderate, and large effect sizes, respectively.

One-way ANOVA analyses and *t*-tests were used to assess the impact of demographic characteristics on subscale T-scores. As a secondary analysis, a series of 3, 2-way ANOVA models were run for each attitude subscale in order to assess for interaction effects. The models included professional position (the primary demographic variable of interest) and 1 of either source, gender, or experience as main effects, plus the 2-way interaction term. Analysis of the validation exercises included two phases: (1) Pearson correlations between the total scores for each attitude subscale and the credibility ratings; (2) multiple regressions predicting the credibility ratings using the 4 subscale total scores as predictors. This analysis strategy included exploratory regression runs to assess the contribution of several 2-way interactions in predicting the credibility ratings. A simultaneous entry model for the 4 subscale main effects and the centered, multiplicative interactions terms was used. Only 1 of 20 interaction terms (4 subscales $\times$ 5 interactions) was significant at the $p < .05$ level, leading to a decision to exclude interactions from the regression runs. Because of the interest in isolating the contribution of Skep, a 2-step entry was selected for the final regression model process. At step 1, F-Under1, F-Under2, and F-Over were entered simultaneously. At step 2, Skep was added to produce the final 4-variable model. Forensic attitudes are hypothesized to mediate the impact of demographic variables such as gender and professional position on case decisions. As a result, no demographic variable was included in the regression models as either a control or predictor variable lest the impact of the attitudes be attenuated.

Results

Demographic comparisons

Forensic attitudes among professional positions differed substantially. One-way ANOVA analyses and post hoc tests are summarized in Table 3. Victim advocates and law enforcement officers ranked highest and lowest, respectively, in beliefs about the frequency of missed cases of CSA (F-Under1). The mean difference between VA and LE subgroups was 8.6, corresponding to a Cohen's *d* of .86 or a large effect. Victim advocates also had significantly higher scores than CPS, ATT, and CFI (moderate to large *d*'s between .55 and .71). The second F-Under subscale, whether evaluators should err on the side of the child, revealed a rift between the legal and non-legal disciplines. Attorneys were least likely to endorse such a position and were separated from 5 of 6 professional groups by 1/2 of an S.D. or more, corresponding to *d*'s of .5 or higher. CPS personnel achieved the highest ranking on the emphasis-on-specificity subscale, with a significantly higher mean score than all 6 other professional positions (*d*'s ranging from .25 to .82). Law enforcement personnel, ranked second, were significantly higher than ATT, CFI, or VA groups. CPS and LE professionals were also ranked 1 and 2 in their level of skepticism. Post hoc tests revealed that CPS scored higher on the Skep subscale than all other groups except child forensic evaluators, with *d*'s ranging from up to .63.

Table 4 presents the demographic comparisons for sample source, gender, and years of experience. Sample source was significant for 3 of the 4 attitude subscales. Professionals recruited during in-state trainings scored higher on F-Under2, F-Over, and Skep than professionals recruited at national conferences. The corresponding *d*'s ranged from a modest .15 to a near moderate .47. There was no evidence of interaction effects between sample source and professional position in the 2-way ANOVA models. In addition, a comparison of attitude subscale means among professional positions revealed fairly comparable patterns for in-state vs. national conference recruits in the professional positions ranked high and low. As a result, professionals from the 2 sample sources were combined for most subsequent analyses.

Table 3
Comparisons of *t*-score means for attitude subscales by professional position.

Subscales	Professional position							
	CPS (*N* = 415)	MH (*N* = 141)	LE (*N* = 224)	VA (*N* = 55)	ATT (*N* = 65)	CFE (*N* = 39)	CFI (*N* = 164)	*F*(6, 1087)
F-Under1								
Mean[a]	48.7	52.0	47.2	55.8	48.1	50.7	50.3	8.15***
Rank	5	2	7	1	6	3	4	
F-Under2								
Mean[b]	51.1	50.0	49.2	52.7	43.1	48.1	47.7	4.12***
Rank	2	3	4	1	7	5	6	
F-Over								
Mean[c]	52.7	49.0	50.2	44.5	45.4	47.0	45.2	11.55***
Rank	1	3	2	7	5	4	6	
Skep								
Mean[d]	52.0	47.0	49.6	45.7	47.7	48.7	47.4	6.71***
Rank	1	6	2	7	4	3	5	

[a] Significant differences on Tukey's HSD test at $p < .05$: VA > CPS, ATT, LE, CFI; CFI > LE; MH > CPS, LE.
[b] Significant differences on Tukey's HSD test at $p < .05$: CPS, VA > CFI, ATT; MH, LE > ATT.
[c] Significant differences on Tukey's HSD test at $p < .05$: CPS > all others; LE > ATT, CFI, VA; MH > CFI, VA.
[d] Significant differences on Tukey's HSD test at $p < .05$: CPS > all but CFE.
*** $p < .001$.

Professional experience also proved to be an important differentiator of forensic attitudes. Professionals reporting more than 10 years of experience were significantly less likely to emphasize specificity or to be skeptical of child reports than colleagues with less experience and more likely to believe that many true cases of CSA are missed by the system (F-Under1) than their colleagues reporting minimal experience. Effect sizes for the experience variable fell in the small category, with *d*'s ranging from .21 to .42. Women professionals were more likely to emphasize sensitivity as measured by both F-Under subscales than men, while men were more likely to emphasize specificity. The 2 sexes, however, were not significantly different in their level of skepticism.

The 2-way ANOVA models used to assess interaction effects between professional position and the 3 other demographic variables revealed only 1 interaction in 12 ANOVA models that was significant at the $p < .05$ level: Position × Experience for F-Under1 [$F(12, 1062) = 2.03$, $p = .02$]. An examination of group means for this interaction clarified an earlier finding: Among professionals with >10 years of experience, only the MH, LE, and CFE groups had higher F-Under1 scores, while the remaining groups exhibited mixed patterns of results.

Intercorrelations between study variables

Table 5 presents the intercorrelations of the major study variables. The 2 factors of F-Under were moderately correlated with each other ($r = .31$), but minimally correlated with either F-Over or Skep. As noted earlier, emphasis-on-specificity and skepticism, as measured by F-Over and Skep, were found to be moderately highly related ($r = .50$).

F-Under2 and F-Over were significantly correlated with the credibility ratings for all 3 validation exercises. Professionals endorsing the position that evaluators should err on the child's side rated the allegations as more likely credible. Professionals expressing higher specificity concerns rated the allegations in the exercises as less likely true. F-Under1 and Skep were significantly correlated with the Case Vignettes and Record Review credibility ratings. Higher scores on the "Missing Cases Common" factor, but lower scores for skepticism, were associated with higher credibility ratings in both exercises.

The Case Vignettes exercise, which provided the least case information for decision-making, produced the highest attitude-behavior correlations with a mean r of .29, with 2 of its 4 *r*'s greater than .30. In contrast, the Record Review exercise, providing the most comprehensive case summary among the exercises, produced the smallest attitude-behavior correlations (mean $r = .16$), while the mean *r* for the Mock Evaluation exercise ($r = .20$) was intermediate. These *r*'s fall in the small to moderate effect size range according to the Rosnow and Rosenthal (2002) classification system (i.e., $r = .10$, small; $r = .30$, moderate; $r = .50$, large).

Regression models predicting credibility ratings

As shown in Table 6, the 2-step multiple regression model predicting the Case Vignettes total credibility score was highly significant [$F(4, 234) = 19.32$, $p = .0001$], with the highest R^2 of the three models (adjusted $R^2 = .24$). Forensic attitudes, as measured by CFAS subscales, thus accounted for 24% of the variance in credibility ratings. All four attitude subscales were found to be significant contributors to prediction. The significant ΔR^2 at step 2 indicated that Skep contributed significantly to prediction [$F(1, 234) = 4.62$, $p = .03$], over and above the contribution of the other three subscales, thus supporting the distinctiveness and heuristic value of both the F-Over and Skep subscales.

The regression model predicting the Mock Evaluation Credibility rating was also significant [$F(4, 109) = 4.81$, $p = .001$], with both F-Under2 and F-Over contributing to the successful prediction. The R^2 of .15 (adjusted $R^2 = .12$) suggested that the

Table 4
Comparison of *t*-score means for attitude subscales by source, gender, and experience.

Subscales	Source			Gender			Experience			
	In-state ($N = 557$)	Nat. confer. ($N = 546$)	$t(1101)$	Men ($N = 284$)	Women ($N = 797$)	$t(1079)$	0–2 yrs ($N = 267$)	3–10 yrs ($N = 529$)	>10 yrs ($N = 289$)	$F(2, 1082)$
F-Under1										
Mean	49.3	49.7	NS	48.4	49.8	2.01*	48.4[a]	49.6	50.5[a]	3.06*
F-Under2										
Mean	50.3	48.8	2.65***	47.7	50.2	3.64**	50.1	49.4	49.2	.632
F-Over										
Mean	51.9	47.2	7.99***	52.2	48.7	5.13***	51.4[a]	49.8[b]	47.2[a,b]	12.86***
Skep										
Mean	51.2	47.8	5.72***	50.2	49.4	NS	50.9[a]	50.1[b]	47.1[a,b]	12.00***

Means within rows sharing the same superscript (a, b) were significantly different at $p < .05$ on Tukey's HSD test.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 5
Intercorrelations between study variables.

Study vriables	F-Under1	F-Under2	F-Over	Skep
F-Under1		.31***	−.05	−.12***
F-Under2			−.07	−.06
F-Over				.50***
Skep				–
Case vignette ratings ($N = 239$)	.20**	.29***	−.32***	−.37***
Mock evaluation ratings ($N = 114$)	.04	.25**	−.34***	−.17
Record review ratings ($N = 252$)	.14*	.16*	−.13*	−.21**

Note: No study participant took part in more than one decision exercise so there are no intercorrelations between credibility ratings.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 6
Summary of multiple regression analyses predicting credibility ratings in three decision exercises.

	B	SE B	β	R^2	ΔR^2
Case vignettes					
Step 1				.19	
F-Under1	.13	.04	.18**		
F-Under2	.21	.07	.19**		
F-Over	−.15	.03	−.32***		
Step 2				.25	.06*
F-Under1	.11	.04	.15*		
F-Under2	.21	.07	.19**		
F-Over	−.10	.03	−.21***		
Skep	−.12	.03	−.26***		
Mock eluation					
Step 1				.15	
F-Under1	−.03	.04	−.06		
F-Under2	.14	.07	.21*		
F-Over	−.07	.02	−.30**		
Step 2				.15	.00
F-Under1	−.03	.04	−.06		
F-Under2	.14	.07	.21*		
F-Over	−.07	.03	−.31**		
Skep	.00	.03	.02		
Record review					
Step 1				.05	
F-Under1	.11	.04	.07		
F-Under2	.04	.06	.13		
F-Over	−.05	.02	−.14*		
Step 2				.08	.03
F-Under1	.04	.04	.06		
F-Under2	.12	.06	.14*		
F-Over	−.01	.03	−.04		
Skep	−.08	.03	−.20**		

* $p < .05$.
** $p < .01$.
*** $p < .001$.

forensic attitudes accounted for a small, but meaningful amount of the total variation in credibility ratings. Entry of the Skep subscale at step 2 did not improve prediction.

Prediction of the credibility rating from the Record Review exercise was also successful [$F(4, 235) = 5.01$, $p = .0007$] though the adjusted R^2 was only .06. Only F-Under2, a belief in erring on the child's side, and Skep were significant contributors in the final model. It is noteworthy that F-Over was a significant predictor in the first step of the model, but was replaced by Skep in step 2, suggesting considerable overlap in their contribution to this model.

Discussion

This study sheds considerable light on the question of whether, and to what degree, the 3 identified forensic attitudes contribute to evaluator disagreements. First, the study demonstrates that the attitudes—emphasis-on-sensitivity, emphasis-on-specificity, and skepticism—can be quantified using a brief questionnaire with acceptable psychometric properties. Despite a moderately high intercorrelation between 2 of the 4 subscales, exploratory and confirmatory factor analysis, multiple regression analysis, and T-score subscale comparisons confirm that the 4 subscales of the CFAS are statistically distinct.

Second, significant attitude effects were found for all 4 of the demographic variables assessed. Several of these effects were predicted and thus offer support for the validity of the CFAS. The following are examples: (a) consistent with their professional roles, victim advocates ranked highest among professional groups on the 2 emphasis-on-sensitivity subscales and lowest on the specificity and skepticism subscales; (b) our finding that women emphasize sensitivity while men favor specificity may explain the results from prior studies that women are more likely than men to view sexual abuse allegations as credible (Finlayson & Koocher, 1991; Jackson & Nuttall, 1997); and (c) consistent with speculation that national conference attendees are likely to be better trained/experienced than in-state workshop participants, significant attitude differences were found for 3 of 4 CFAS subscales.

Other demographic findings, though unexpected, were understandable in retrospect. For example, attorneys scored the lowest of all seven professional positions on F-Under2, Erring on Child's Side. This finding suggests that despite their victim-oriented role, prosecutors reject the notion of anything less than an objective stance. Such a position is consistent with their responsibilities as officers of the court.

The third central finding of this study is that the influence of forensic attitudes on professional judgement is widespread, but highly variable in magnitude. Despite the diverse evaluation formats represented, the forensic attitudes were found to predict case ratings for all three decision exercises. These findings provide support for the validity of the CFAS and for the primary thesis of the study that individual differences in the strength of the forensic attitudes help explain evaluator disagreements. The pattern of findings suggests that the emphasis-on-sensitivity and emphasis-on-specificity dimensions, respectively, operate as predispositions to view sexual abuse allegations as likely valid or likely invalid; skepticism functions as a bias against believing child and adolescent accounts of abuse. It is possible that high levels of these attitudes may bias a professional's view of a case, even sight unseen.

Both the correlational and regression analyses contribute to our understanding of the magnitude of attitude influences. Ten of 12 attitude-behavior correlations were statistically significant, with *r*'s ranging up to .37. This correlational pattern suggests broad, but only small to moderate effects for individual attitudes. Interestingly, the size of *r*'s is consistent with those commonly reported in more traditional attitudinal research. For example, Kraus (1995) reported a median attitude-behavior correlation of .33 in his meta-analysis of 88 attitude studies.

In contrast, the regression analyses revealed substantial variation in the impact of attitude combinations. The regression analyses indicate that attitude profiles are more predictive of case decisions than individual attitudes. The adjusted R^2 for the Record Review, Mock Evaluation and Case Vignettes regression models were .06, .12 and .24, respectively. The impressive proportion of variance explained by the Case Vignettes model and the 4-fold difference between the smallest and the largest R^2 are both noteworthy. It seems remarkable that a 10–12-min attitude survey can predict 24% of the variance in credibility ratings. This portion of variance is particularly striking when contrasted with the idealized standard for forensic assessments of 0% subjectivity.

The large adjusted R^2 for Case Vignettes, especially relative to the other exercises, has at least 3 explanations: Just as the inherent bias in a weighted coin may not be detectable in a single toss, predispositions in professional judgements might not be evident in a single case decision. Ajzen and Fishbein (1977) made a similar point in noting that attitudes influence the overall pattern of behavior, but they may not be predictive of any specific behavior. The Case Vignettes exercise was unique among the 3 exercises, in that its total credibility rating was a summation of 4 case ratings and thus reflected a pattern of professional judgement rather than a single case decision. Second, the absence of strong case facts in the Case Vignettes exercise may have encouraged or forced participants to rely on subjective factors in lieu of case facts. Last, Case Vignettes participants were less experienced and likely less well trained than Record Review participants and therefore perhaps more susceptible to subjective influences.

A sea change at CPS?

CPS professionals exhibited an overall attitude profile that cannot easily be reconciled with traditional views of their role. Specifically, they scored significantly higher on the specificity subscale than all other professional groups and significantly higher on skepticism than all but one group. CPS also scored well below the mean on one of the two emphasis-on-sensitivity subscales. As we have seen, such a pattern of scores is associated with a higher probability of disbelieving sexual abuse allegations. This finding is troubling in light of the role of CPS as one of the primary gatekeepers for sexual abuse cases entering the system. Other players in the system include law enforcement, prosecuting attorneys, judges, juries, and various mental health professionals. Ideally, these other players, in combinations that vary with case characteristics, function as checks and balances for CPS substantiation decisions. However, in most cases, there are no checks and balances for CPS decisions against substantiating allegations of abuse. As a result, if CPS sets standards for accepting or substantiating allegations that are too high, there is a risk of many true cases of child sexual abuse being screened out or unsubstantiated, leaving little recourse for abuse victims.

The elevated rate of skepticism among CPS personnel in the current study sharply contrasts with the rates obtained in three earlier studies. Using widely varying measures of skepticism, Boat and Everson (1988), Saunders (1988), and Everson et al. (1996) found that CPS scored significantly lower on skepticism than law enforcement and either achieved the lowest ranking, or were tied for the lowest ranking, among the assorted professional groups sampled. Although all three studies suffer from substantive methodological weaknesses as described earlier, the consistency of findings across the three studies is noteworthy. It is also notable that the North Carolina sample of CPS workers in the current study and the CPS samples in

two of the three earlier studies (i.e., Boat & Everson, 1988; Everson et al., 1996) were drawn primarily from the same county Department of Social Services agencies. The question therefore arises whether the current findings may represent a major transformation or sea change in CPS attitudes since the mid-1990s. Further exploration of this issue is warranted.

Limitations of study

This study has at least five limitations of note. First, the sample was recruited from workshop and conference participants rather than randomly selected from the larger population of maltreatment professionals. Second, the study relies on a newly-developed, untested attitude scale of modest scope and design. Third, shared method variance from the use of self-report to assess both attitudes and professional judgement may have artificially inflated the strength of study findings. Fourth, the attitudes assessed by the CFAS subscales have been shown to predict credibility ratings in the classroom and the conference center. The unanswered question is whether they predict dichotomous substantiation decisions in the field.

Fifth, it is possible that this study seriously underestimates the contribution of subjective factors to evaluator disagreements. In actual practice, subjective factors not only affect the substantiation decision, but also numerous, earlier choices related to evaluation content and structure. These include the number of child interviews, the focus of questioning, and the records chosen for review. By standardizing the case information provided, the decision exercises eliminated a major source of subjective influence.

Implications for practice

This study provides support for Herman's (2005, 2009) criticism of the limited reliability of professional judgements. However, the findings also suggest possible interventions to decrease subjective influences on forensic decision-making. As a result, we believe that Herman's call for severe remedies to purge abuse assessment of subjective influences is premature. First, this study identified three forensic attitudes influencing case decisions and demonstrated that such subjective factors can be reliably and validly measured. Applying similar methodology, it is likely that additional subjective factors can be identified and quantified (e.g., a propensity to identify with the alleged victim vs. alleged perpetrator). As a result, it is feasible to test Jackson and Nuttall's (1997) speculation that self-awareness of biases is a key to curbing their influence. An analogy comes to mind: Just as a bathroom scale that reads +3 lbs before any weight is applied must be reset to zero for accurate measurement, so too an evaluator may need to make a conscious effort to "reset to zero" to minimize the effect of known predispositions. Second, the assessment of forensic attitudes can be used to design individualized training to address specific biases and misconceptions. For example, findings from preliminary analyses suggest that approximately 1 CPS worker in every 4 believes that the majority of adolescents who disclose sexual abuse are lying or making a false report. Third, results from the current study suggest that strong case facts may weaken the influence of subjective factors. Therefore, conducting comprehensive investigations may be the single most effective strategy for improving the reliability and validity of case decisions. Finally, a "team" approach to assessment that emphasizes diversity in professional position or discipline, gender, and experience level is likely to be useful in providing alternative perspectives to counterbalance individual biases.

Erroneous case decisions resulting from evaluator subjectivity and bias have the potential to devastate lives. Yet, progress in remedying this problem has been minimal since early warnings were sounded two decades ago (e.g., Berlinen & Conte, 1993; Corwin et al., 1987). If the field does not take the necessary steps to ensure the reliability of case decisions, there is little doubt that "drastic reforms" will ultimately be imposed.

References

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918.

Berlinen, L., & Conte, J. (1993). Sexual abuse evaluations: Conceptual and empirical obstacles. *Child Abuse & Neglect*, *19*(3), 371–384.

Boat, B. W., & Everson, M. D. (1988). Use of anatomical dolls among professionals in sexual abuse evaluations. *Child Abuse & Neglect*, *12*, 171–179.

Corwin, D., Berliner, L., Goodman, G., Goodwin, J., & White, S. (1987). Child sexual abuse and custody disputes: No easy answers. *Journal of Interpersonal Violence*, *2*(1), 91–105.

Everson, M. D., Boat, B. W., Bourg, S., & Robertson, K. R. (1996). Beliefs among professionals about rates of false allegations of child sexual abuse. *Journal of Interpersonal Violence*, *11*(4), 541–553.

Faller, K. C. (2003). *Understanding and assessing child sexual maltreatment* (2nd ed.). Thousand Oaks, CA: Sage.

Finlayson, L. M., & Koocher, G. P. (1991). Professional judgement and child abuse reporting in sexual abuse cases. *Professional Psychology: Research and Practice*, *22*, 464–472.

Fletcher, R., Fletcher, S., & Wagner, E. (1996). *Clinical epidemiology: The essentials* (3rd ed.). Baltimore, MD: Williams and Wilkins.

Herman, S. (2005). Improving decision making in forensic child sexual abuse evaluations. *Law and Human Behavior*, *29*(1), 87–120.

Herman, S. (2009). Forensic child sexual abuse evaluations: Accuracy, ethics and admissibility. In K. Kuehnle, & M. Connell (Eds.), *The evaluation of child sexual abuse allegations: A comprehensive guide to assessment and testing* (pp. 247–266). Hoboken, NJ: Wiley.

Hershkowitz, I., Fisher, S., Lamb, M. E., & Harowitz, D. (2007). Improving credibility assessment in child sexual abuse allegations: The role of the NICHD investigative interview protocol. *Child Abuse & Neglect*, *31*(8), 99–110.

Horner, T. M., Guyer, M. J., & Kalter, N. M. (1993a). Clinical expertise and the assessment of child sexual abuse. *Journal of the American Academy of Child and Adolescent Psychiatry*, *32*(5), 925–931.

Horner, T. M., Guyer, M. J., & Kalter, N. M. (1993b). The biases of child sexual abuse experts: Believing is seeing. *Bulletin of the American Academy of Psychiatry Law*, *21*(3), 281–292.

Jackson, H., & Nuttall, R. (1993). Clinician responses to sexual abuse allegations. *Child Abuse & Neglect*, *17*, 127–143.

Jackson, H., & Nuttall, R. (1997). *Childhood abuse: Effects on clinicians' personal and professional lives*. Thousand Oaks, CA: Sage.

Kraus, S. J. (1995). Attitude and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin*, *21*(1), 58–75.
McGraw, J. M., & Smith, H. A. (1992). Child sexual abuse allegations amidst divorce and custody proceedings: Refining the validation. *Journal of Child Sexual Abuse*, *1*, 49–62.
Ney, T. (1995). Assessing allegations in child sexual abuse: An overview. In T. Ney (Ed.), *True and false allegations of child sexual abuse* (pp. 3–20). New York, NY: Brunner/Mazel.
Nunnally, J. C., & Berstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.
Poole, A. A., & Lindsay, D. S. (1998). Assessing the accuracy of young children's reports: Lessons from the investigation of child sexual abuse. *Applied and Preventive Psychology*, *7*(1), 1–26.
Realmuto, G. M., Jensen, J., & Wescoe, S. (1990). Specificity and sensitivity of sexually anatomically correct dolls in substantiating abuse: A pilot study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *29*(5), 743–746.
Realmuto, G. M., & Wescoe, S. (1992). Agreement among professionals about a child's sexual abuse status: Interviews with sexually anatomically correct dolls as indicators of abuse. *Child Abuse and Neglect*, *126*(5), 719–725.
Rosnow, R. L., & Rosenthal, R. (2002). Contrasts and correlations in theory assessment. *Journal of Pediatric Psychology*, *27*(1), 59–66.
Runyan, D. K. (1998). Prevalence, risk, sensitivity and specificity: A commentary on the epidemiology of child sexual abuse. *Child Abuse & Neglect*, *22*(6), 493–498.
Saunders, E. J. (1988). A comparative study of attitudes toward child sexual abuse among social work and judicial system professionals. *Child Abuse & Neglect*, *12*, 83–90.
Shumaker, K. R. (2000). Measured professional competence between and among different mental health disciplines when evaluating and making recommendations in cases of suspected child sexual abuse. *Dissertation Abstracts International*, *60*(11), 5791B (UMI no. 9950748).