The usability and effectiveness of the Finnish investigative instrument of child sexual abuse in mock evaluations

Tadei, Alessandro; Santtila, P; Antfolk, Jan

Published in:
Nordic Psychology

DOI:
10.1080/19012276.2019.1662323

Published: 01/01/2019

Document Version
Accepted author manuscript

This is a post-print of an accepted manuscript to be published by Taylor & Francis in Nordic Psychology. Please reference the final publication. The final proofed version might differ from the current document.

The Usability and Effectiveness of the Finnish Investigative Instrument of Child Sexual Abuse in Mock Evaluations

Alessandro Tadei[1], Pekka Santtila[1,2], Jan Antfolk[1],

[1] Åbo Akademi University, Turku, Finland

[2] New York University, Shanghai, China

Abstract

The Finnish Investigative Instrument of Child Sexual Abuse (FICSA) is a computerized tool that uses Bayesian statistics to provide a base rate for an alleged child sexual abuse (CSA), using population-level information about correlates of CSA. FICSA can, thus, assist decision-making in investigations of CSA. Here, we compared forensic experts' and students' ability to use FICSA and whether its use affected the estimates of the probability of CSA in mock-scenarios. The use of FICSA was compared to only having access to the empirical information about CSA risk and protective factors, which FICSA is based on, and to unassisted decision-making. The 54 participants analyzed two scenarios of possible CSA and estimated the probability of the CSA allegation being true. The results show that participants using FICSA were prone to make technical mistakes that affect the correctness of the probability estimation. The performance of experts and students was equivalent in all the conditions, with the exception of the group using FICSA, where experts tended to deviate from the probability provided by FICSA more than students. Having only access to empirical information did not improve estimates compared to unassisted decision-making. Both students and experts tended to adjust the estimates provided by FICSA downwards, that is, to decrease the probability of abuse. We conclude that FICSA has the potential to assist investigators to correctly integrate evidence and calculate probabilities but that proper training is required.

Keywords: FICSA, child sexual abuse, forensic psychology, Bayesian statistics, decision-making

The Usability and Effectiveness of the Finnish Investigative Instrument of Child Sexual Abuse in Mock Evaluations

The Finnish Investigative Instrument of Child Sexual Abuse (FICSA, for a detailed description about the instrument and its limitations, see Tadei et al., 2017) is a decision-making tool based on Bayesian statistics that estimates the base probability of an alleged child sexual abuse (CSA) case being true at an early stage of the investigations. The FICSA does not require any proximate information about the alleged event. As in other methods that use information provided by the victim of an alleged crime to test the veracity of their case (De Zutter, Horselenberg, & van Koppen, 2017), the data FICSA uses to assess the probability of CSA having occurred is based on information about, and provided by, the alleged victim. This information includes the child's previous experiences with crime, alcohol, drugs and smoking, domestic violence, peer relationships, and life habits. FICSA uses, in total, 42 observations to calculate the probability of CSA. Of these observations, 14 are used only for boys, 25 only for girls, and 3 are used for both genders. Through a feature selection procedure, these 42 variables were selected from a pool of 903 variables from a stratified population-based victimization-study of 11,000 Finnish children (Ellonen, Fagerlund, Kääriäinen, Peltola, & Sariola, 2013). The FICSA employs naïve Bayes classifier where the 42 variables are considered independent from each other. The state of these variables has an impact on the probability of the class variable (i.e., whether the child was sexually abused or not). Even if having all the observations requested by FICSA increases the accuracy of the outcome estimate, FICSA can provide an outcome also when details are missing. According to a simulation study (Tadei et al., 2017), FICSA has a very high discriminatory effectiveness (*AUC* = .97 for boys and *AUC* = .88 for girls). Although these diagnostic utility values are very promising, it is yet unclear how well FICSA will perform in actual investigations of CSA

allegations. Figure 1 shows the logic FICSA is based on and how data from the case under investigation are entered.

[Figure 1]

Figure 1. Example of a simple Bayesian model that uses FICSA's logic.

The model starts from the child sexual abuse base rate for the population under examination and updates the probability after new evidence is added. Panel A shows the base rate of abuse (.03) and no other observation has yet been made. Panel B shows the updated probability after discovering that the child has been pushed or shaken by the father more than 12 months ago (the probability for this is.15 for CSA victims and .03 for non-victims). Panel C shows the final CSA probability (.46), which is based also on the additional discoveries that the child smokes daily and that the child has not experienced anyone unknown behaving in a rude way on the Internet during the last year.

FICSA (Tadei et al., 2017) was created to provide a possible solution to the multiple issues, theoretical and practical, that can negatively affect CSA investigations. The first problem the investigators have to deal with is the under-reporting of CSA (i.e., real cases that are never reported to the authorities). However, recently attention has been paid to the dangers of over-reporting (i.e., false cases reported to authorities as real ones) as well (Bolen & Scannapieco, 1999; Melkman, Hershkowitz, & Zur, 2017). It is complicated to estimate the extent of the under-reporting of CSA. However, a recent study estimated that approximately 20% of CSA victims do not disclose the event to anyone (Lahtinen, Laitila, Korkman, & Ellonen, 2018). In this study, the question whether cases reported to someone reached the authorities remained unanswered, as 48% of the disclosures were to friends, and only 12% were disclosures directly to authorities. The over-reporting of CSA is also difficult to estimate

with high precision. However, Melkman, Hershkowitz, and Zur (2017) estimated that about 58% of disclosures made by the alleged victims could be considered credible. Comparable results have been reached by Korkman, Antfolk, Fagerlund, and Santtila (*under review*), who report that 25-40% of CSA allegations in Finland might be unfounded.

A second problem that may lead to wrong conclusions in alleged CSA cases is the lack of witnesses and of corroborating evidence (Goodman, Batterman-Faunce, Schaaf, & Kenney, 2002; Herman, 2010). In these cases, the child's statements are what legal professionals usually have to base their decisions on. For this reason, it is of paramount importance that forensic interviews are conducted according to the highest standards (see La Rooy et al., 2015). However, research shows that the quality of expert interviews and assessments varies and may sometimes be low (Korkman, Santtila, Westeråker, & Sandnabba, 2008; Zajac, Garry, London, Goodyear-Smith, & Hayne, 2013). This is also evidenced by the low inter-rater reliability among purported experts (Jackson & Nuttall, 1993). Finally, in our own research, we have found that legal professionals are not always sufficiently equipped to evaluate expert testimony (*self citation B)*, nor to correctly match expert witness competencies to topics to be addressed (*self citation A)*.

FICSA was created to provide information that can improve and facilitate CSA investigations. Research already has demonstrated that in many fields actuarial decision-making processes lead to more accurate assessments than clinical decisions (Dawes, Faust, & Meehl, 1989; Janus & Prentky, 2003). One clear advantage of an automated decision-support tool is that it is in principle free from different kinds of cognitive biases that affect investigators. The risk of confirmation bias is especially relevant in this context. When using FICSA, available and modellable information is integrated objectively. Also, the information needed as input for FICSA can be gathered through a background questionnaire that the child answers in written form. This avoids many possible negative effects of verbal interaction

between an interviewer and the child, which risks the creation of false memories *in primis* (see Otgaar, de Ruiter, Howe, Hoetmer, & van Reekum, 2017). These negative effects are more likely to appear when the interviewers have already made up their mind about the case or ask suggestive questions (Korkman, Santtila, & Sandnabba, 2006; Korkman et al., 2008), and FICSA can potentially limit their problematic impact already in the early stage of investigations. It is, however, possible that the questionnaire will be misunderstood by the child, entering wrong information. In case the expert designated to supervise this procedure had doubts about the validity of the answers, an investigative interview that follows best-practice guidelines (Hershkowitz, Fisher, Lamb, & Horowitz, 2007; La Rooy et al., 2015) could be used to verify the information provided by the child.

FICSA gives a more accurate base rate to start the rest of the investigation from. Being based only on information that is not directly related to the CSA allegation, FICSA's result can and must, however, be integrated with other pieces of evidence discovered during the investigation. Furthermore, the use of FICSA at an early stage of an investigation permits the investigators to order the CSA allegations from the most to the least probable, and allow investigations to proceed following this order in case prioritization is needed. This would likely increase the number of cases that are correctly solved by facilitating effective use of investigative effort. It could also speed up the investigative process, which is important considering that the time between the alleged CSA event and the forensic interviews is one of the most crucial predictors for the reliability of the information collected as well as for a child's resistance to suggestive interviewing (Antfolk, Rönnholm, Salo, Rantamäki, & Santtila, n.d.; Pipe, Sutherland, Webster, Jones, & La Rooy, 2004; Quas et al., 2007).

The Role of CSA Investigative Units in Finland

In Finland, evaluations of CSA allegations are generally conducted by the police. However, in the case of very young (i.e. pre-school age) children or children with special

needs (e.g., developmental challenges) or otherwise particularly complex cases, the police can request the assistance of forensic expert units, which include psychiatrists, psychologists, and social workers (Duodecim, 2013a; Korkman et al., 2017). There are five units in Finland, with 5-15 professionals working in each unit. The role of these expert units is to produce evidence (i.e., videotaped forensic interviews with the child) that are used in a trial (Duodecim, 2013b; Korkman et al., 2017), should the prosecutor raise charges. The units often write expert statements listing the background information of the allegation, defining alternative hypotheses for the allegation, and weighting the information provided by the child in the light of these hypotheses. These statements are also considered in court, and albeit not binding for the court (Finnish Supreme Court, 2014), the statements may be influential for the outcome of a potential trial. Currently, experts are required to follow evidence-based guidelines and complete training on interviewing. There is no instrument or no guidelines regarding how to deal with background factors to the abuse, to set a baseline for the probability wherefore the estimation of the base rate is based on subjective integration of available empirical literature.

The Present Study

In the present study, we wanted to examine the usability of FICSA and whether its use would increase the accuracy of assessments by the Finnish CSA Investigative Unit experts, compared to having access to the empirical information about the risk and protective factors FICSA is based on, and compared to unassisted decision-making. We also compared forensic experts and psychology students. All the participants received two scenarios that briefly described the lives of a boy and a girl. Participants were either provided with scenarios only, with scenarios and empirical information about the relation between CSA and the variables included in the scenarios, or with scenarios and FICSA. Participants were asked to estimate the probability that each of the two described children had been victim of CSA.

We formulated a series of hypotheses. The first was that FICSA could be used properly after a very short training, that is, we expected that the participants would be able to accurately code the information in the cases they were presented into FICSA and then correctly interpret the probability calculated by the system. Therefore, we expected perfect accuracy in the participants' ability to use FICSA and, therefore, also no difference in the correct use of the tool between forensic experts and students. Our second hypothesis was that forensic experts would assess CSA cases more accurately than students both when provided with access to empirical information about the case under examination and when provided with the case only (i.e., using prior knowledge and clinical experience alone). This hypothesis was based on the fact that many experts have formal training in this specific context. We therefore also expected that experts would trust the probability calculated by FICSA less than students as we expected them to possess more relevant knowledge about other influential evidence that should be taken into consideration.

When comparing the use of FICSA with the other two conditions, we expected that participants using FICSA would give a more accurate[1] estimation of the CSA probability than the other participants. Similarly, we expected a better performance of the group provided with the empirical information, when compared with the group that could count only on their experience and knowledge. When asked to assess the probability of the alleged CSA under examination as either a substantiated case or not, we expected the participants to use the range 0-49% for the cases they considered unfounded, and 51-100% for the probable ones. We did not have any hypothesis about how participants would consider situations where the probability was exactly 50%.

Finally, when we, at a later stage, added details that could not be modeled in FICSA (i.e., we could not define the impact of these variables, if any, on the CSA probability) to the

[1] In the present study, for the probabilities estimated by the participants who used FICSA, we use the term "accuracy" to refer to the agreement between the participants and the outcome of FICSA itself.

cases, we expected all participants to update their estimation according to the type of evidence (i.e., increasing the probability of abuse for the girl scenario and decreasing the probability of abuse in the boy scenario).

The limitations of FICSA were not the object of this study, as they have been previously discussed in a different paper (Tadei et al., 2017).

Method

Participants

The sample consisted of 54 participants. Of these, 27 were medical doctors, psychologists and social workers employed in three of the five Child Sexual Abuse Investigative Units operating in Finland. The other 27 participants were students who had completed two non-compulsory courses in Forensic Psychology offered at Åbo Akademi University. In these two courses, among other things, the students learned the notion of base rate and some basics of Bayesian logic. They did not, however, receive any training in the use of FICSA or about the variables we identified as statistically related to CSA. Of the participants, 83% (n = 45) were women, and 17% (n = 9) were men. The mean age of the experts was 42 (SD = 10) and 24 (SD = 4) for the students. The experts had, on average, five years of work experience and had, on average, handled 116 (SD = 123) CSA cases each. All but one participant provided an answer to all the questions in the questionnaire. The participant who did not fill the questionnaire in its entirety was excluded from the analyses that required the missing information.

Ethical Permission

All participants signed an informed consent form that specified the topic of the research, policies about anonymity, and their right to withdraw from the study at any point. The consent form also explicitly stated that participants had to inform the experimenter in case of any familiarity with FICSA. Before data collection commenced, the Institutional

Review Board of the Faculty of Arts, Psychology and Theology, Åbo Akademi University, granted permission to carry out this study.

Measures and Procedures

Participants were given two scenarios (the order of scenarios was counterbalanced across participants) that briefly described the life of two 15-year-old Finnish children: A girl, *Milla* and a boy, *Otto*. Each scenario was created to contain information about 15 out of the 42 variables that FICSA uses to estimate the probability of a child being victim of CSA. This means that the base rate probability that Milla and Otto had been abused could be modeled based on the population-based information regarding the associations between these variables and the risk of sexual abuse.

Participants were randomly divided into three experimental conditions, making sure that each condition had the same number of participants, equally divided between students and experts. We also aimed to balance the number of psychologists, psychiatrists, and social workers. The three experimental conditions were: participants who received the scenarios only (hereafter, scenarios only group), participants who received the scenarios and a written summary containing all and only the empirical information FICSA uses to calculate the probability of abuse (empirical info group), and participants who received the scenarios and a laptop running FICSA (FICSA group).

To further clarify the material, here is an extract of the summary the empirical info group received together with the scenario about Otto: […] *51.3% of Finns under 16 years of age are female / The prevalence of CSA involving boys as victims is 0.7% / Seeing their parent(s) drunk about once a month is common both for non-victims (23%) and CSA victims (29%)* […].

Participants were also instructed about what events were to be considered CSA and what not, so to be consistent with the definition of CSA used for building FICSA. The

instructions were: *[…] In this study we define Child Sexual Abuse as anything from receiving a proposal to do something sexual to sexual penetration of a person below 17 years of age by an offender that is at least 5 years older than the victim. Please, keep this definition in mind when you perform the subsequent tasks.*

Training. FICSA group was instructed how to use FICSA through a short scenario about a non-sensitive topic: Under the supervision of the researchers, FICSA group members entered the data from this example scenario in a simple model created for training purposes and, finally, obtained a probability estimate. Only when participants learned how to correctly use the software, were they allowed to proceed to the actual experiment. For clarity, here is an example of how the correct information should be entered in FICSA. In the first lines of Milla's scenario we read "[her father] forbids her to have a cellphone and to make use of social networks. She says she has never been bullied, threatened, or sexually harassed by phone or phone messages". The respondents had a list of 15 questions that started with "Bullied or insulted through text messages during the last year?" and a drop down menu with the possible answers, "Yes" and "No" in this case. They had to click on the correct answer, "No", and to move on to the second question.

Reporting. After having read a scenario, all participants had to state, using only the limited information available, if they believed the child in question was a CSA victim or not and express the probability that CSA had occurred. The two questions were formulated as follows: *i)* "Would you conclude that the child in question has been victim of CSA or not? (Yes/No)", and *ii)* "What is for you the probability of the CSA? (0-100%)". FICSA group, also, had to report the probability calculated by the FICSA model. This last measurement was used to measure the impact of FICSA on the decisional process, as well as to verify that FICSA had been used properly. After having answered these questions, participants were presented with two additional pieces of information about the child's life. This information

could not be modeled in FICSA. After this, participants were asked to update their previous answers to the questions. Figure 2 summarizes the entire procedure.

[Figure 2]

Figure 2. Study design for one scenario.

The exact procedure was repeated for the second scenario.

Scenarios. For clarity, we report here the two scenarios, together with the non-modellable information.

Scenario concerning Milla: "Milla is a 15-year-old Finnish girl. Her father is a retired officer of the Finnish army. He has always been extremely strict with the daughter. He forbids her to have a cellphone and to make use of social networks. She says she has never been bullied, threatened, or sexually harassed by phone or phone messages. Also, no one has ever behaved in a rude way towards her on the web, and she is not aware of any bad gossip in social networks that would concern her person.

Milla is also not allowed out after 10pm, not even during weekends or when a party is organized by her classmates. At home, she has dinner with the family every day—it is considered as a sort of ritual. Two years ago Milla was studying for the entire afternoon at her best friend's place and came back one hour late for dinner. Her father got very angry, violently shook her by the shoulders and told her she was 'a stupid little girl if she thought she could do whatever she wanted in that house'.

On the other hand, Milla fights her father's authority when she is out of home. She admits to having tried ecstasy a couple of times and that she sometimes smokes cigarettes with her friends at school, not daily. However, she has never drunk alcohol.

Milla has a boyfriend that she keeps secret. They never had sex. Also no one among her peers has tried to have sex with her.

Milla says that in the last six months she certainly feels worried, but she does not say why."

Non-modellable information concerning Milla: *i*) The math teacher said that in the last months, she saw several times a group of dangerous looking guys waiting for Milla outside the school. She never looked happy to follow them. When the teacher asked her if they were her friends, she replied they were, but in a very elusive way. *ii*) A gynecologist has examined Milla. She had some concerns for partially damaged hymen she observed during the examination, but she could not draw a definite conclusion about abuse. The gynecologist did not find other signs of possible violence.

Scenario concerning Otto: "Otto is a 15-year-old Finnish boy. He is going through a difficult period. His father had an accident on the job and lost mobility in his legs. This happened one year ago. After this, the father started drinking heavily, and, maximum once a month, Otto finds his father completely drunk and has to take care of him.

Otto's classmates have heard about the father's problems and started to mock Otto. In particular, there are some boys his age calling him names every time they run into him on the public transport. Because of his father, the boys call Otto 'alcoholic' and 'junkie'—even if Otto never drank or smoked anything. It is not rare that, on the public transport, these boys also attack Otto physically, and once he ended up with a black eye. Recently, the bullying also happens through messages to his phone. No one has encouraged him to talk to some authority or to his teachers about the situation.

Even if Otto doesn't like his classmates, he feels the need to belong to his peer group and to stay out of his home as much as possible. Often he hangs out with peers in the park or at the main square. He stays out late, fairly often even until after midnight. During these

moments it sometimes happened that the peer group plays a game, in which someone chosen at random needs to touch someone else's genitalia. Otto recalls he has done this with about five of his peers. This is the only form of sexual contact he had with someone his age. Otto, in fact, disagrees to some extent when asked if the others consider him handsome. He has never had a girlfriend or a boyfriend, and no peers never showed any interest in having sex with him. Otto says no one ever asked him for anything sexual online.

Last month Otto was victim of a robbery where his wallet was taken and his phone was broken. Otto thinks he recognized the attacker as a classmate's 27-year-old brother."

Non-modellable information concerning Otto: i) The suspicion of CSA comes from Otto's worried grandmother. She said that her grandson had changed, was frequenting bad company, and she was sure someone was doing "bad things, sexual things" to him. The woman, in the past, already has reported possible CSA for other grandchildren. Some of these cases are still under investigation, the others have been classified as "unfounded". *ii*) Otto denies having been victim of any sexual violence.

Statistical Analyses

To test our hypothesis that FICSA can be correctly used after only short training, we first calculated the percentage of the FICSA group members who were able to use FICSA without making mistakes (i.e., FICSA calculated the correct probability for the scenario under examination). We used the Fisher's Exact Test to determine if being a student or an expert was associated with the capacity to use FICSA. Pearson correlation was used to determine if the number of mistakes (i.e., each question answered wrong counted as one mistake) in the two scenarios were correlated.

Independent samples *t*-tests were used to test whether the order in which the scenarios were presented influenced the number of errors in using FICSA, and investigate if, as we expected, forensic experts were more accurate in assessing the CSA probability compared to

students in scenarios only group and empirical info group. Independent samples *t*-tests were also used to investigate if the use of FICSA increases the accuracy of CSA assessments compared to scenarios only group and empirical info group, and if the use of empirical information helps the assessment rather than not.

Paired-samples *t*-tests were used to investigate if the FICSA group participants tended to decrease, increase, or keep the probability calculated by FICSA, and to investigate if they updated their estimates after being presented with the non-modellable information.

Finally, we identified the probability range participants used to assess a case they considered CSA rather than not. We investigated in particular the use they did of the value 50%.

For each *t*-test we used a bootstrapping procedure with 1000 re-samples because, as concerns Otto's scenario, the distribution of the probabilities estimated by scenarios only group, empirical info group and FICSA group presented unequal variances (Otto's scenario, all participants: $F(2, 50) = 5.60$, $p < .01$; Otto's scenario, wrong FICSA excluded: $F(2, 37) = 3.46$, $p < .05$; Milla's scenario, all participants: $F(2, 50) = .40$, $p = .67$; Milla's scenario, wrong FICSA excluded: $F(2, 44) = .28$, $p = .76$). For similarity, we decided to use bootstrapping also on the analyses about Milla's scenario.

Results

Correct Use of FICSA

Whereas both scenarios only group and empirical info group were asked to base their probability estimations on their own reasoning only, the probability estimations of FICSA group were potentially influenced by the FICSA outcome. This outcome could be correct or incorrect, depending on whether the software was used properly or not (i.e., entering correct or incorrect information into the FICSA). Therefore, we first calculated the proportion of participants who correctly entered all information in the FICSA. The analyses indicated that

only 35% of participants in FICSA group made no mistake in using FICSA in the scenario about the boy. The corresponding percentage was 70% in the scenario about the girl. The Fisher's exact test showed that being a student or a forensic expert was not associated with the probability of using FICSA correctly (χ^2 = 2.51, p = .21). Also, correct use of the FICSA was not influenced by the order in which the scenarios were presented to the participants (mistakes when starting with the boy's scenario, M = .85, SD = 1.14, and mistakes when starting with the girl's scenario, M = 1.25, SD = 1.89, and the difference between these means, -.40, BCa 95% CI [-1.47, 0.56] was not statistically significant, p = .42). On the other hand, the number of mistakes (M = 1.45, SD = 1.70 for Otto's scenario and M = .65, SD = 1.31 for Milla's scenario) in one scenario correlated positively with the number of mistakes in the other one, r = .74, p < .001.

Accuracy in Estimating CSA Probability

We also investigated whether, in the three different experimental conditions separately and altogether, experts were more accurate than students in assessing the CSA probability. For this analysis, we excluded the participants who used FICSA incorrectly. The CSA probability calculated by FICSA was 67.8% for Otto and 25.5% for Milla. Table 1 shows the differences between experts and students, investigated through independent samples t-tests with bootstrapping. Contrary to our expectation, we found that the only experimental condition in which experts differed statistically significantly from students was FICSA group. In the FICSA group, the students followed FICSA's results more closely than the experts.

[Table 1]

The Impact of FICSA and the Empirical Information on Probability Estimates

Through independent samples *t*-tests with bootstrapping, we investigated the differences between participants who used FICSA correctly (FICSA group) and participants without FICSA (scenarios only group and empirical info group). On average, participants using FICSA estimated a more correct CSA probability for both scenarios, but the difference from the other two groups was not statistically significant. Concerning the scenario about the boy (correct probability = 67.8%), the FICSA group estimated a higher probability (M = 36.79, SE = 12.06) than the scenarios only group and the empirical info group together (M = 15.60, SE = 3.06). The mean difference, 21.18, BCa 95% CI [-1.44, 46.17], was, however, not statistically significant, p = .10. In the scenario about the girl (correct probability = 25.5%), the FICSA group again estimated a higher probability (M = 15.82, SE = 3.45) than the scenarios only group and the empirical info group together (M = 10.30, SE = 2.19). The difference was 5.52, BCa 95% CI [-2.22, 12.75] was not, however, statistically significant, p = .18.

As we found no difference between the FICSA group and the two other groups, we investigated the correlation between the probability calculated by FICSA and the one estimated by its users. Hence, we investigated if the FICSA group was influenced by FICSA's estimations at all. Since here we did not investigate the correctness of the estimations, we included also participants who did not use FICSA correctly, and joined the two scenarios together for increased statistical power. We observed a moderate positive correlation between the probability calculated by FICSA and the one expressed by the participants, r = .43, p < .01. A higher probability calculated by FICSA thus tended to increase the value estimated by FICSA group members, meaning that it is likely they were influenced by FICSA's result.

We used paired samples *t*-tests with bootstrapping to investigate whether the FICSA group tended to decrease, increase, or keep the probability calculated by FICSA. We

considered the entire FICSA group also for this analysis. Results showed that in both scenarios the respondents tended to decrease the value offered by FICSA. As concerns Otto's scenario, participants decreased the estimation from M = 64.69 (SE = 7.05) to M = 25.89 (SE = 6.07) with a mean difference of 38.80, BCa 95% CI [23.77, 54.05], p = .001. For Milla's scenario, participants decreased the FICSA's result from M = 23.50 (SE = 2.70) to M = 14.35 (SE = 3.03) with a mean difference of 9.15, BCa 95% CI [4.18, 14.50], p < .01.

Using independent samples t-tests with bootstrapping, we also analyzed whether having empirical information necessary for the calculation (empirical info group) available increased the accuracy compared to only having only scenarios (scenarios only group). The analyses showed that scenarios only group and empirical info group did not differ in terms of CSA assessment accuracy. Concerning the scenario about the boy, the estimation of the empirical info group was 13.09 (SE = 3.64), whereas the estimation of scenarios only group was 17.97 (SE = 4.68). The mean difference was 4.88, BCa 95% CI [-6.34, 17.56], and not statistically significant, p = .44. Concerning the scenario about the girl, the estimation of the empirical info was 12.25 (SE = 3.25), and the scenarios only 8.46 (SE = 3.15). The mean difference was -3.79, BCa 95% CI [-12.11, 5.32], and not statistically significant, p = .44.

Non-modellable information

We then investigated whether participants updated their estimates after being presented with non-modellable information, and, if so, in which direction. For this analysis, we used paired samples t-tests with bootstrapping and also the participants who used FICSA incorrectly were included. Table 2 shows that the participants did not modify their estimation about Otto, while they increased it significantly for Milla, considering the experimental conditions both separately and together.

[Table 2]

Probability Threshold to Define Abuse or Non-abuse

We were finally interested in understanding which probability range was used by our sample to indicate situations of abuse and of non-abuse. Table 3 shows the percentages of participants who used a value that was lower, equal, or higher than 50% to indicate the two possible outcomes of their assessment.

[Table 3]

These results indicate that 50% is mostly used to indicate situations of non-abuse, while the most part of the sample used values between 0 and 49% to indicate non-abuse, and 51-100% for abuse.

Discussion

We investigated how the use of FICSA affects CSA assessments made by forensic expert and students of forensic psychology. Furthermore, we investigated how difficult it is to use FICSA after only short training. We also investigated if forensic experts perform better than students in assessing CSA cases with the tools we provided. Finally, we aimed to identify the probability ranges used to define a case as abuse or non-abuse.

Correct Use of FICSA

Contrary to our expectations, many participants made mistakes when coding FICSA variables from the case scenarios, with 65% of participants making at least one mistake in Otto's scenario and 30% in Milla's scenario. A post-hoc evaluation of the errors suggested that in many cases mistakes seemed to be clerical errors and from not double checking the information entered into the FICSA. For example, a sentence in Otto's scenario is "Last month Otto was victim of a robbery where his wallet was taken […]", to the question "Was Otto victim of a robbery during the last year?", a participant answered "No". Also, most of the

questions (12 out of 15 in the scenario concerning Otto, 8 out of 15 in the scenario concerning Milla) have been wrongly answered at least once, indicating that the mistakes were not related to a few problematic variables. Equally many experts and students made mistakes, and the number of mistakes was not affected by the order of the scenarios. Furthermore, we found a positive correlation between the number of mistakes in the two scenarios, indicating that people who made mistakes in one scenario were likely to do the same in the second scenario as well. Since even one mistake in entering data in FICSA can result in inaccurate probabilities calculated by the software, the goal is to bring the number of errors down to zero. Apart from more extensive training, one possibility is to delete the step in which the expert manually enters the collected information in FICSA. Instead, the alleged victim could answer a list of questions directly and these answers could automatically be entered into the FICSA. In case of questionable validity of the information entered in FICSA provided by the child, the investigative interview could also be designed to retrieve information about the FICSA variables. We currently advise that the data entered in FICSA are always double-checked by a second expert. Bringing the mistakes down to zero, together with an extensive training about the logic behind FICSA and the interpretation of its results, could also likely help improving the inter-rater reliability that is now undoubtedly negatively affected by wrong data entered in the tool.

The Effects of FICSA on Decision-Making

We hypothesized that forensic experts would have been more accurate than students in estimating the CSA probability when provided with the scenarios only or with the scenarios and the empirical information necessary to calculate a more precise probability. This hypothesis was not supported as the forensic experts' performance did not differ from that of the students. These two groups, instead, performed differently when they used FICSA. As predicted, students estimated a CSA probability closer to that calculated by FICSA compared

to the forensic experts. The experts, instead, adjusted the probability downwards. This does not mean that students are better than forensic experts in assessing CSA cases, but only that the experts were more conservative in their estimates. Neither scenario included details about the alleged sexual abuse, but only about the children's life habits, relations and personal experiences. Currently, forensic experts would not consider these pieces of information as informative in CSA investigations. Therefore, it is possible that when FICSA calculated a probability as high as 67.8%, experts did not consider this probability reliable and justifiable to police and other legal professionals. It should also, however, be considered that experts may be overly conservative. The theory behind the model, the choice of the questions asked to the children, and the interpretation of the results should be more clearly illustrated so that experts can make clearer choices regarding whether to adjust or not the probability created by FICSA. When we compared the performance of the group that used FICSA with the other two groups, we noticed that the average answers from FICSA group were closer to the ones provided by the software, suggesting that FICSA influenced estimates of the probabilities. The finding was, however, not statistically significant, perhaps as a result of high variation in the answers or limited statistical power. Hence, we cannot fully conclude that the use of FICSA affects the decision-making during CSA assessment. However, having observed a moderate positive correlation of .43 between FICSA results and FICSA group estimations, we can claim that FICSA slightly influences the reasoning of its users, even if both students and forensic experts tended to decrease the probability calculated by the model.

Interestingly, the estimations of the participants in all the experimental conditions were much closer to the estimate provided by FICSA in the scenario about the girl, than in the one about the boy. It is possible that perceptions about the increased likelihood of girls (vs. boys) being victims of abuse led to this finding. In fact, independently of the scenario and the experimental condition, all the participants tended to provide quite low probabilities (19.5%

for the boy and 11.8% for the girl). With the corresponding FICSA estimates being 25.5% for the girl and 67.8% for the boy, it is clear how consistently estimating low probabilities results in more similar results to the one provided by FICSA for Milla's scenario.

Our hypothesis about the advantage of using empirical information was disconfirmed. Even if, with proper calculus skills, it is possible to estimate the correct probability using only the provided empirical information, our data showed no difference between scenarios only group and empirical info group. We can therefore conclude that having access to empirical information had no impact in itself on the accuracy of the assessment. However, we had low statistical power to identify differences.

Non-Modellable Information and the Final Decision

Once the non-modellable information was added, we expected that all the participants would update their estimations, increasing it for Milla and decreasing it for Otto. This hypothesis was supported only in the scenario about the girl, where the participants increased their estimation. In the scenario about the boy, the estimations before and after non-modellable information did not differ. Not knowing exactly which impact the provided non-modellable information should have on the CSA probability, we cannot define adjustments as correct or incorrect. We encourage updating the FICSA's result in the light of new evidence, especially when there is sound empirical evidence of the evidence substantially impacting the likelihood of abuse (e.g., confession by the alleged perpetrator, DNA evidence, or images of the abuse).

Almost all participants used the range 0-49% to define a case as non-abuse, and the range 51-100% to define abuse. This is in line with our expectations. In line with our previous argument about conservatism in this decision-making, 89.5% of the sample used the probability 50% to define a case as non-abuse. However, it should be noted that the amount of cases defined as abuse or non-abuse was strongly unbalanced towards the latter (92.6%).

Limitations

Even if our sample can be considered small, which negatively impacts statistical power, it is highly representative of the population of interest. We collected data from more than half of the existing CSA Investigative Units in Finland and two out of the included units are among the largest in the country. The number of students involved was limited to equal the number of forensic experts we recruited.

Using only two scenarios, one with a high and one with a low CSA probability to estimate, might have limited the conclusions we could draw from our study. Including four scenarios would have allowed to orthogonally manipulate gender and low/high CSA probability. However, doubling the scenarios, and consequently the time required to the participants to answer all the questions, would have further decreased the sample size.

Furthermore, both scenarios included 15 arbitrarily chosen variables each. Being taken from FICSA, the amount of variables per scenarios could have reached a maximum of 17 for Otto and 28 for Milla. We chose to use only 15 variables to limit the time required to complete all the tasks. In addition, a different combination of variables might have led to different probabilities estimated by FICSA and, possibly, by the participants. We chose these particular sets of variables because they let us build one high and one low probability scenario (Otto: 67.8%, Milla: 25.5%). In practice, experts from CSA investigative units are seldom or never asked to evaluate allegations that present so little information, especially presenting only information not directly related to the CSA event itself, as was the case in the scenarios we used in the current study. Our choice, however, allowed us to create scenarios with a rate calculated by FICSA without statistically irrelevant details that could have distracted the participants. Using only FICSA variables in our scenarios also let us study our participants' capacity to integrate multiple complex probabilistic information compared to a software. The disadvantage was, instead, that we relied on a somewhat biased approach favoring FICSA, as

we asked professionals to perform a task that is different from their current daily routine. Nevertheless, the current task represents a situation for which FICSA was built.

We also want to point out that among the variables used by FICSA to estimate the CSA probability, there are some (e.g., use of drugs or alcohol) that reflects a risky behavior. It is important to note that reason for these variables to be included is not that the behaviors are potentially dangerous, but only that these behaviors are statistically related to the risk of being victim of CSA. Moreover, while also the child's behavior is relevant in statistical estimations of CSA risk, moral blame is always to be attributed to the perpetrator and not to the victim in the event of abuse. It is important that this is considered also in actual assessment so that the investigative procedure does not put blame on the victim.

Conclusions

The present study examined the use of the Finnish Investigative Instrument of Child Sexual Abuse, and if having access to empirical information about CSA improves the accuracy in CSA assessments. Only having access to empirical information did not improve decision-making, therefore a tool that can assist in properly using evidence is of paramount importance. On the other hand, the current study demonstrated that, without a proper training, participants had difficulties in using FICSA and interpreting its outcomes. More research is also required to explore the integration between information provided by FICSA and non-modellable evidence to updating the FICSA in the light of pertinent new evidence.

List of abbreviations

FICSA: Finnish Investigative Instrument of Child Sexual Abuse

CSA: child sexual abuse

AUC: area under the curve

References

Antfolk, J., Rönnholm, M., Salo, B., Rantamäki, S., & Santtila, P. (n.d.). Children's Suggestibility in Interviews – A Meta-analysis of the Effects of Child's Age, Post-Event Delay, Event-Type, and Question-Type.

Bolen, R. M., & Scannapieco, M. (1999). Prevalence of child sexual abuse: A corrective metanalysis. *Social Service Review*, *73*(3), 281–313. https://doi.org/10.1086/514425

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

De Zutter, A., Horselenberg, R., & van Koppen, P. J. (2017). Detecting the True Nature of Allegations of Rape. *Journal of Police and Criminal Psychology*, *32*(2), 114–127. https://doi.org/10.1007/s11896-016-9203-z

Duodecim (the Finnish Medical Society). (2013a). Current Care Guidelines: Investigating Suspicions of Child Sexual Abuse. Retrieved from http://www.kaypahoito.fi/web/kh/suositukset/suositus;jsessionid=7E3A7E56839CC04A143BED0DA7F6BD97?id=hoi34040

Duodecim (the Finnish Medical Society). (2013b). Current Care Guidelines: Investigating Suspicions of Child Sexual Abuse.

Ellonen, N., Fagerlund, M., Kääriäinen, J., Peltola, M., & Sariola, H. (2013). Child Victim Survey 2013 [dataset]. Version 1.0 (2015-03-16). Finnish Social Science Data Archive. Retrieved from http://urn.fi/urn:nbn:fi:fsd:T-FSD2943

Finnish Supreme Court. A Decision Concerning a Child Sexual Abuse Allegation. (2014). Retrieved from http://korkeinoikeus.fi/fi/index/ennakkopaatokset/precedent/1404215595714.html

Goodman, G. S., Batterman-Faunce, J. M., Schaaf, J. M., & Kenney, R. (2002). Nearly 4 years after an event: children's eyewitness memory and adults' perceptions of children's

accuracy. *Child Abuse & Neglect*, *26*(8), 849–884. https://doi.org/10.1016/S0145-2134(02)00354-X

Herman, S. (2010). The role of corroborative evidence in child sexual abuse evaluations. *Journal of Investigative Psychology and Offender Profiling*, *7*(3), 189–212. https://doi.org/10.1002/jip.122

Hershkowitz, I., Fisher, S., Lamb, M. E., & Horowitz, D. (2007). Improving credibility assessment in child sexual abuse allegations: The role of the NICHD investigative interview protocol. *Child Abuse & Neglect*, *31*(2), 99–110. https://doi.org/10.1016/J.CHIABU.2006.09.005

Jackson, H., & Nuttall, R. (1993). Clinician responses to sexual abuse allegations. *Child Abuse and Neglect*, *17*(1), 127–143. https://doi.org/10.1016/0145-2134(93)90013-U

Janus, E. S., & Prentky, R. A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility and accountability. *American Criminal Law Review*.

Korkman, J., Pakkanen, T., Laajasalo, T., Korkman, J., Pakkanen, T., & Laajasalo, T. (2017). Child Forensic Interviewing in Finland: Investigating Suspected Child Abuse at the Forensic Psychology Unit for Children and Adolescents. In *Collaborating Against Child Abuse: Exploring the Nordic Barnahus Model* (pp. 145–164). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-58388-4

Korkman, J., Santtila, P., & Sandnabba, N. K. (2006). Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scandinavian Journal of Psychology*. https://doi.org/10.1111/j.1467-9450.2006.00498.x

Korkman, J., Santtila, P., Westeråker, M., & Sandnabba, N. K. (2008). Interviewing techniques and follow-up questions in child sexual abuse interviews. *European Journal of Developmental Psychology*, *5*(1), 108–128. https://doi.org/10.1080/17405620701210460

La Rooy, D., Brubacher, S. P., Aromäki-Stratos, A., Cyr, M., Hershkowitz, I., Korkman, J., … Lamb, M. E. (2015). The NICHD protocol: a review of an internationally-used evidence-based tool for training child forensic interviewers. *Journal of Criminological Research, Policy and Practice*, *1*(2), 76–89. https://doi.org/10.1108/JCRPP-01-2015-0001

Lahtinen, H.-M., Laitila, A., Korkman, J., & Ellonen, N. (2018). Children's disclosures of sexual abuse in a population-based sample. *Child Abuse & Neglect*, *76*, 84–94. https://doi.org/10.1016/J.CHIABU.2017.10.011

Melkman, E. P., Hershkowitz, I., & Zur, R. (2017). Credibility assessment in child sexual abuse investigations: A descriptive analysis. *Child Abuse & Neglect*, *67*, 76–85. https://doi.org/10.1016/J.CHIABU.2017.01.027

Otgaar, H., de Ruiter, C., Howe, M. L., Hoetmer, L., & van Reekum, P. (2017). A Case Concerning Children's False Memories of Abuse: Recommendations Regarding Expert Witness Work. *Psychiatry, Psychology and Law*, *24*(3), 365–378. https://doi.org/10.1080/13218719.2016.1230924

Pipe, M. E., Sutherland, R., Webster, N., Jones, C., & La Rooy, D. (2004). Do early interviews affect children's long-term event recall? *Applied Cognitive Psychology*, *18*(7), 823–839. https://doi.org/10.1002/acp.1053

Quas, J. A., Malloy, L. C., Melinder, A., Goodman, G. S., D'Mello, M., & Schaaf, J. (2007). Developmental Differences in the Effects of Repeated Interviews and Interviewer Bias on Young Children's Event Memory and False Reports. *Developmental Psychology*, *43*(4), 823–837. https://doi.org/10.1037/0012-1649.43.4.823

Tadei, A., Pensar, J., Corander, J., Finnilä, K., Santtila, P., & Antfolk, J. (2019). A Bayesian Decision-Support Tool for Child Sexual Abuse Assessment and Investigation. *Sexual Abuse: A Journal of Research and Treatment*, *31*(4), 374-396.

https://doi.org/10.1177/1079063217732791

Zajac, R., Garry, M., London, K., Goodyear-Smith, F., & Hayne, H. (2013). Misconceptions about childhood sexual abuse and child witnesses: Implications for psychological experts in the courtroom. *Memory*. https://doi.org/10.1080/09658211.2013.778287