BMJ Paediatrics Open

Reliability of the Sexual Knowledge Picture Instrument: a potential diagnostic instrument for sexual abuse in young children

Kirsten van Ham ,[1] Shanti Bolt,[2] Mariska van Doesterling,[2] Sonja Brilleslijper-Kater,[1] Rian Teeuw,[1] Rick van Rijn,[3] Hans van Goudoever,[1] Hanneke van der Lee[4]

To cite: van Ham K, Bolt S, van Doesterling M, *et al*. Reliability of the Sexual Knowledge Picture Instrument: a potential diagnostic instrument for sexual abuse in young children. *BMJ Paediatrics Open* 2022;**6**:e001437. doi:10.1136/bmjpo-2022-001437

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjpo-2022-001437).

Received 5 February 2022
Accepted 6 May 2022

[1]Paediatrics, Amsterdam UMC Locatie AMC, Amsterdam, The Netherlands
[2]Social Paediatrics, Emma Children's Hospital, Amsterdam, The Netherlands
[3]Paediatric Radiology, Amsterdam UMC Locatie AMC, Amsterdam, The Netherlands
[4]Epidemiology, Kennisinstituut van Medisch Specialisten, Utrecht, The Netherlands

Correspondence to
Dr Sonja Brilleslijper-Kater; s.n.brilleslijper-kater@amsterdamumc.nl

ABSTRACT

Objective To determine the intra-rater and inter-rater reliability of the Sexual Knowledge Picture Instrument (SKPI), a potential diagnostic instrument for young suspected victims of sexual abuse containing three scoring forms, that is, verbal responses, non-verbal reactions and red flags.

Design Video-recorded SKPI interviews with children with and without suspicion of child sexual abuse were observed and scored by two trained, independent raters. The second rater repeated the assessment 6 weeks after initial rating to evaluate for intra-rater reliability.

Subjects 78 children aged 3–9 years old were included in the study. 39 of those included had known suspicion of sexual abuse and the other 39 had no suspicion.

Main outcome measures Intra-rater and inter-rater reliability of the scores per study group and in the total sample were assessed by Cohen's kappa and percentage of agreement (POA).

Results The median intra-rater Cohen's kappa exceeded 0.90 and the POA exceeded 95 for all three forms in both study groups, except for the red flag form (median Cohen's kappa 0.54 and POA 87 in the suspected group, and 0.84 and 92, respectively, in the total sample). For the verbal scoring form the median inter-rater Cohen's kappa and POA were 1.00 and 100, respectively, in both groups. For the non-verbal form the median inter-rater kappa and POA were 0.37 and 97, respectively, in the suspected group, and 0.47 and 100, respectively, in the control group. For the red flag form, they were 0.37 and 76, respectively, in the suspected group and 0.42 and 77, respectively, in the control group.

Conclusion The reliability of the SKPI verbal form was sufficient, but there is room for improvement in the non-verbal and red flag scoring forms. These forms may be improved by adjusting the manual and improving rater training.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Despite its major consequences, sexual abuse in young children often remains unrecognised by medical and psychological professionals.

WHAT THIS STUDY ADDS

⇒ The verbal scoring form of the Sexual Knowledge Picture Instrument has adequate intra-rater and inter-rater reliability.

⇒ The reliability of the non-verbal and red flag scoring forms is suboptimal, requiring improvement of the manual and interviewer training for these forms.

HOW THIS STUDY MIGHT AFFECT RESEARCH AND PRACTICE

⇒ This study is part of the validation of an instrument that can be used in the diagnosis of sexual abuse in young children.

INTRODUCTION

Child sexual abuse (CSA) is a worldwide problem with potentially detrimental consequences to the victims.[1–4] Short-term and long-term health effects that may arise as a result include depression, anxiety, post-traumatic stress disorder, eating disorders, substance abuse, and somatic syndromes such as sleeping disorders and heart and lung diseases.[4–7] Early detection of signs of CSA by medical or psychological professionals is crucial to provide specialist support to the victims and to protect possible future victims. However, as reported by adults who were victims of CSA, and supported by the gap between prevalence numbers reported by authorities and self-report studies, we know that timely diagnosis of CSA is uncommon.[8–14]

Professionals who see young children with a suspicion of CSA are challenged for several reasons. When a child is presented for healthcare due to a suspected CSA, the chance of finding physical evidence is very small.[15 16] Due to the nature of the abuse, there are usually no witnesses, although recording the abuse, either for personal use or to share on the dark web, does occur.[17] Victims may struggle with feelings of dependency on, and loyalty to, the

perpetrator, as well as feelings of shame and guilt or fear of being blamed if they disclose about sexual abuse. The limited verbal capacity of young children may hamper their ability to express their experiences, thoughts and feelings even more.[11 14] Unfortunately, lessons from the past make us aware that the use of developed tools to facilitate disclosure, such as dolls and diagrams, even by professionals, can lead to false positive results.[18–20] This can have major consequences, especially if such findings are used during the legal process, as was shown in notorious cases of false allegations of CSA.[21–24] The current lack of scientific substantiation and the risk of improper tool use emphasise the importance of developing reliable, structured, evidence-based and uniform methods to support the diagnosis of CSA in clinical practice.

A potential diagnostic instrument for medical and psychological professionals in cases of suspected CSA in young children (aged 3–9 years) is the Sexual Knowledge Picture Instrument (SKPI), based on previous work by Brilleslijper-Kater.[25] This instrument consists of a child-friendly picture book with 15 illustrations about family routines, gender differences and identity, genitals and their functions, reproduction, intimate and sexual behaviour in adults, and normal physical intimacy in children. A semistructured interview technique from a manual allows a trained interviewer to conduct an open conversation with the child about the topics in the pictures and to potentially overcome the burden of disclosure. Afterwards, video recordings of each interview can be scored according to three standardised scoring lists from the manual: one on the child's verbal responses, one on non-verbal behavioural reactions and one on overall impression and/or alarm signs (the so-called 'red flags'). The SKPI pictures and manual are presented in online supplemental appendices 1 and 2.

The aim of this study is to determine the intra-rater and inter-rater reliability of the SKPI. This is the first of two studies planned to validate the SKPI as a diagnostic instrument for CSA in children aged 3–9 years.[26] If the diagnostic accuracy is proven to be adequate, this tool could be a valuable addition to current medical and psychological diagnostic work-up in young children with a suspicion of CSA.

METHODS

Subject selection

In 2016, the Picture Instrument for Child Sexual Abuse Screening (PICAS) study started at Amsterdam University Medical Center. It included children aged 3–9 years with and without suspicion of CSA. During the study, trained interviewers used the SKPI with a sample of children from two different sources:

- First, a group consisting of suspected victims of CSA who had either been referred to the Department of Social Paediatrics in one of three participating Dutch university medical centres or who were investigated by a vice squad of the Dutch national police.
- Second, a control group consisting of children considered not to be victims of CSA.

For more details on the study procedures, we refer to the article on the protocol.[26]

As recommended by de Vet *et al*,[27] a minimum sample size of 50 subjects is required in validation studies of measurement instruments. To reach this number, all 39 children with suspicion of CSA who had been interviewed with the latest version of the scoring forms were included, as well as a selected sample of 39 children from the control group with equal age and gender distribution.

Data collection

Video-recorded interviews with the 78 children were scored three times: immediately by a first rater (who was one of the eight interviewers), a second time by the second rater (one forensic science master's student) and a third time by the same second rater after a minimum interval of 6 weeks, to preclude recollection. All raters were either physicians or master's students with medical or forensic background. They were individually trained by a specialised child psychologist (SB-K) and/or the main researcher (KvH) on how to conduct the semistructured interviews and how to work with the manual. All raters were blind to participants' medical and psychological background information, and only the first rater was aware of the study group to which each child belonged.

The verbal scoring form contained all 52 interview questions from the manual. By checking one of four (n=45) or five (n=7) answer options, each rater scored the answer given by the child. The non-verbal scoring form contained a table listing a total of 24 behavioural reactions. Each reaction could be checked for presence while observing each of the 15 pictures. The red flag scoring form consisted of three overarching questions with binary answer options to assess the interviewer's overall impression of the child's verbal and non-verbal behaviour during the interview.

Statistical analysis

The SKPI's intra-rater reliability was assessed by comparing the two scorings of the second rater at different time points. Inter-rater reliability was assessed by comparing the rater scores for each child between the first rater and the primary scoring of the second rater. Data analysis was performed using the IBM SPSS software package (IBM SPSS Statistics for Windows, V.26.0).

Descriptive statistics (percentages, median and IQR) were used to describe the demographic characteristics of the study population. For the verbal scoring, no, multiple answer options or 'other…' were considered a missing value. We calculated both Cohen's kappa and percentage of agreement (POA) to assess intra-rater and inter-rater reliability. By definition, POA is higher than Cohen's kappa, since kappa is adjusted for agreement by coincidence. For this reason, kappa is generally preferred over POA. However, in contrast to kappa, POA can always be calculated, even when some options have not been scored

Table 1 Baseline characteristics study population

Variables	Suspected CSA group (n=39)	Control group (n=39)	Total sample (n=78)
Male, n (%)	15 (39)	20 (51)	35 (45)
Age (years), median (IQR)	5 (3-7)	5 (4-7)	5 (4-7)
Age groups, n (%)			
3 years	10 (26)	7 (18)	17 (22)
4 years	8 (20)	7 (18)	15 (19)
5 years	5 (13)	7 (18)	12 (15)
6 years	6 (15)	6 (15)	12 (15)
7 years	1 (3)	6 (15)	7 (9)
8 years	9 (23)	6 (16)	15 (20)

by one of the raters, as was the case for many items, in particular on the non-verbal scoring form.[28]

For the interpretation of Cohen's kappa, Landis and Koch's[29] (arbitrary) grading system was applied on the median kappa per form, with a Cohen's kappa of <0 signifying poor agreement, 0.00–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement and 0.81–1.00 as almost perfect agreement. For the interpretation of POA, a median of ≥80% agreement between raters was considered acceptable.[28]

For each of the three separate scoring forms, Cohen's kappa and POA of all items and the median (IQR) per form were calculated in both study groups and in the total study sample.

Patient and public involvement

During the course of PICAS we received input from several adult CSA survivors who lived with the burden of the abuse throughout their childhood. The aim was to carefully assess and evaluate each step of the study with them. We intend to disseminate the main results to all parents and caregivers from the included subjects, as well as these CSA survivors, and will continue seeking their involvement in the development of a tool and appropriate methods of dissemination.

RESULTS

Baseline characteristics

The baseline characteristics of the study population are shown in table 1. The median age was 5 years (IQR: 4–7). Slightly more girls than boys were included (55% vs 45%) in the total sample and in particular in the suspected group (61% vs 39%).

Intra-rater and inter-rater reliability per group

Tables 2 and 3 present aggregated intra-rater and inter-rater reliability, respectively, on all items of the verbal, non-verbal and red flag scoring forms in the suspected CSA group, the control group and the total sample, represented by Cohen's kappa and POA.

Verbal scoring form

Intra-rater and inter-rater agreement on the verbal scoring form are almost perfect in both the suspected and control groups (both median Cohen's kappa 1.00, POA 100). For intra-rater and inter-rater agreement on each of the 52 questions on the verbal scoring form, divided per study group and for the total sample, we refer to online supplemental appendix 3.

Non-verbal scoring form

For the non-verbal form, the median intra-rater Cohen's kappa and POA were 0.91 and 100, respectively, in the suspected group and 0.92 and 100, respectively, in the control group. The median inter-rater Cohen's kappa and POA were 0.37 and 97, respectively, in the suspected

Table 2 Intra-rater reliability

Outcome measure	Suspected CSA group	Control group	Total sample
Verbal scoring form (52 items)			
Cohen's kappa, median (IQR)	1.00 (1.00-1.00)*	1.00 (1.00-1.00)†	1.00 (0.96-1.00)
POA, median (IQR)	100 (100-100)	100 (98-100)	100 (98-100)
Non-verbal scoring form (360 items)			
Cohen's kappa, median (IQR)	0.91 (0.79-1.00)‡	0.92 (0.84-1.00)§	0.90 (0.79-1.00)¶
POA, median (IQR)	100 (97-100)	100 (100-100)	100 (99-100)
Red flag scoring form (3 items)			
Cohen's kappa, median (min-max)	0.54 (0.52- 0.55)	0.95 (0.89-1.00)	0.84 (0.64-0.86)
POA, median (min-max)	87 (77-92)	97 (95-100)	92 (89-94)

*kappa could be calculated for 49 out of 52 questions.
†kappa could be calculated for 44 out of 52 questions.
‡kappa could be calculated for 204 out of 360 reactions.
§kappa could be calculated for 148 out of 360 reactions.
¶kappa could be calculated for 233 out of 360 reactions.
IQR, interquartile range; min-max, lowest and highest value; POA, percentage of agreement.

Table 3 Inter-rater reliability

Outcome measure	Suspected CSA group	Control group	Total sample
Verbal scoring form (52 items)			
Cohen's kappa, median (IQR)	1.00 (0.69-1.00)*	1.00 (0.76-1.00)†	0.91 (0.66-1.00)‡
POA, median (IQR)	100 (94-100)	100 (94-100)	98 (95-100)
Non-verbal scoring form (360 items)			
Cohen's kappa, median (IQR)	0.37 (-.03-0.55)§	0.47 (0.22-0.79)¶	0.36 (-0.01-0.53)**
POA, median (IQR)	97 (92-100)	100 (97-100)	97 (94-100)
Red flag scoring form (3 items)			
Cohen's kappa, median (min-max)	0.42 (0.27-0.47)	(0.38-0.52)††	0.51 (0.45-0.61)
POA, median (min-max)	74 (73-87)	77 (72-97)	82 (73-83)

*kappa could be calculated for 45 out of 52 questions.
†kappa could be calculated for 41 out of 52 questions.
‡kappa could be calculated for 48 out of 52 questions.
§kappa could be calculated for 183 out of 360 reactions.
¶kappa could be calculated for 87 out of 360 reactions.
**kappa could be calculated for 206 out of 360 reactions.
††Kappa could be calculated for 2 out of 3 questions; therefore, only minimum and maximum values given.
IQR, interquartile range; min-max, lowest and highest value.

group and 0.47 and 100, respectively, in the control group. Intra-rater and inter-rater agreement of the non-verbal scoring form on each possible reaction and for each of the 15 pictures per each study group and in the total sample are presented in online supplemental appendix 4.

Red flag scoring form

For the red flag form, the median intra-rater Cohen's kappa and POA were 0.54 and 87, respectively, in the suspected group and 0.95 and 97, respectively, in the control group. The median inter-rater Cohen's kappa and POA were 0.37 and 74, respectively, in the suspected group and 0.42 and 77, respectively, in the control group. For results per question divided per study group and in the total sample, we refer to online supplemental appendix 5.

DISCUSSION

The aim of this study was to evaluate the inter-rater and intra-rater reliability of the scoring method of the SKPI, consisting of a verbal, non-verbal and red flag scoring form, in a group of suspected CSA victims and a healthy control group. The intra-rater reliability of the verbal, non-verbal and red flag scoring forms is substantial to almost perfect, except for the red flag form in the suspected group, which is moderate. All median intra-rater POAs showed acceptable agreement for each of the three forms. The inter-rater reliability of the verbal scoring form is substantial to almost perfect, but the non-verbal and red flag forms show only fair to moderate reliability in both study groups. Inter-rater agreement is acceptable for the verbal and non-verbal forms, but the median POA was under the 80% threshold for the red flag form. The interpretation of Cohen's kappa is arbitrary, as stated in Landis and Koch's often-cited paper.[29] Moreover, Cohen's kappa depends on the distribution of the item scores, leading to lower kappa values with more skewed distributions, as is the case in many of the SKPI items. Therefore, the POA values may be preferable for determining SKPI reliability. Focusing on the results per item (online supplemental appendices 4 and 5), we notice that agreement varies widely between individual items in both the non-verbal and the red flag scoring forms.[30] Therefore, opportunities to improve the scoring method may be found at the level of individual items. For now, simply removing those items that lacked reliability does not seem the best solution, as it may decrease the face validity of the instrument. However, once the diagnostic accuracy of the instrument has been established, it is worth reconsidering this option. Another way to improve the reliability of non-verbal and red flag scoring may be to intensify rater training and to improve manual instructions, in particular with regard to less reliable scoring items.

On the verbal scoring form, raters were instructed to tick the box 'other…' if there was cause for doubt or, which was most often the case, if, despite the manual instructions, the interviewer was unable to ask the question during the interview. This led to a considerable amount of missing data during the analysis, as can be seen in online supplemental appendix 3.

Although the reliability in the CSA suspected group is slightly lower than in the control group for most verbal and non-verbal items, the intra-rater and inter-rater agreement for both forms are generally adequate. On the red flag form, however, the intra-rater reliability is remarkably lower in the suspected than in the control group. This may have been due to the fact that all scoring for this intra-rater analysis was performed by a single

rater who was trained once, before she first rated the video recordings. To improve both intra-rater and inter-rater agreement, in addition to one individual training, refresher courses and group training on how to work with the manual should be considered for all raters to ensure consistency in manual use and form scoring. During training at present, an example interview with a child from the control group is shown, and a single practice interview is conducted with a non-abused child. More extensive experience with use of the SKPI, including a practice interview with a child from the suspected group, should therefore also be included in training to improve interviewer and rater skills.

Strengths and limitations

A strength of the present study is its large sample size involving young children with suspected CSA. The study population consisted of a broad spectrum of children, including confirmed cases of CSA, children with high, moderate or low CSA suspicion in the suspected CSA group, and children with no suspicion in the control group. The study groups were analysed separately to evaluate the SKPI reliability in a group that is largely representative of the target population (suspected CSA group).

Another strength of this study is the blinding of the first and second rater. Only the first rater, who was also the interviewer, had some knowledge of the child's background and whether or not CSA was suspected. A study design with one suboptimally blinded rater and one fully blinded rater (as will be the case when the instrument is used in practice) enhances the validity of the results.

A limitation is that a single and relatively inexperienced second rater performed the repeated assessments, thus limiting the generalisability of the intra-rater reliability. A further limitation is that all interviewers and raters were female. This was not by design. Despite the use of a structured interview technique, children might have responded differently in interviews conducted by male interviewers.[31]

Recommendations for practice

When applied by experienced and trained professionals, the SKPI can be used to lower the threshold to start a conversation with a young child on sexually related topics. However, it is very important that video images of the interviews are analysed afterwards and, if necessary, that remarkable verbal and non-verbal reactions are discussed with another (independent) professional. Creating a balance between the preservation of privacy while enabling objective assessment remains a challenge. Taking into account the European General Data Protection Regulation, clear protocols must be developed and adhered to within each medical or psychological institution on how to deal with storage and/or sharing of data.[32]

Recommendations for research

The diagnostic accuracy of the SKPI will be investigated as a next step in our validation study. In addition, we recommend improving the manual and interviewer training.

CONCLUSION

The verbal scoring form of the SKPI has adequate intra-rater and inter-rater reliability. The reliability of the non-verbal and red flag scoring forms is suboptimal, requiring improvement of the manual and interviewer training for these forms. In its current form, the instrument can be used to open a conversation with a child suspected of being sexually abused. Due to its clear structure, the SKPI is a relevant additional tool for use in the medical, psychological and forensic field.

Acknowledgements We would like to thank all participating children and their parents.

Contributors RT, KvH, SB-K, RvR, HvG and HvdL conceived of the study and initiated the study design. KvH, SB and MvD performed the data collection, data analysis and interpretation. HvdL provided statistical expertise in clinical trial design and conducted the primary statistical analysis. KvH drafted the manuscript. All authors contributed to the refinement of and critical revision of the manuscript and approved the final version, and SB-K acts as guarantor of the study and manuscript.

Funding This study was sponsored by the Contribute Foundation (https://www.contribute.nl/), the Healthcare Insurers Innovation Foundation (2.969; 2016/020201) and the Janivo Foundation (2015.444).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study involves human participants and was approved and monitored by the Medical Ethical Board from the Amsterdam UMC, location AMC in Amsterdam and registered under 2015_173. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

ORCID iD

Kirsten van Ham http://orcid.org/0000-0001-6328-7123

REFERENCES

1 Singh MM, Parsekar SS, Nair SN. An epidemiological overview of child sexual abuse. *J Family Med Prim Care* 2014;3:430–5.
2 Stoltenborgh M, van Ijzendoorn MH, Euser EM, *et al*. A global perspective on child sexual abuse: meta-analysis of prevalence around the world. *Child Maltreat* 2011;16:79–101.

3 Barth J, Bermetz L, Heim E, *et al*. The current prevalence of child sexual abuse worldwide: a systematic review and meta-analysis. *Int J Public Health* 2013;58:469–83.
4 Wegman HL, Stetler C. A meta-analytic review of the effects of childhood abuse on medical outcomes in adulthood. *Psychosom Med* 2009;71:805–12.
5 Afari N, Ahumada SM, Wright LJ, *et al*. Psychological trauma and functional somatic syndromes: a systematic review and meta-analysis. *Psychosom Med* 2014;76:2–11.
6 Maniglio R. The impact of child sexual abuse on health: a systematic review of reviews. *Clin Psychol Rev* 2009;29:647–57.
7 Paras ML, Murad MH, Chen LP, *et al*. Sexual abuse and lifetime diagnosis of somatic disorders: a systematic review and meta-analysis. *JAMA* 2009;302:550–61.
8 Alaggia R, Collin-Vézina D, Lateef R. Facilitators and barriers to child sexual abuse (CsA) disclosures: a research update (2000-2016). *Trauma Violence Abuse* 2019;20:260–83.
9 Brattfjell ML, Flåm AM. "They were the ones that saw me and listened." From child sexual abuse to disclosure: Adults' recalls of the process towards final disclosure. *Child Abuse Negl* 2019;89:225–36.
10 Debelle G, Powell R. 'I just wanted someone to ask me': when to ask (about child sexual abuse). *Arch Dis Child* 2021;106:105–7.
11 Lemaigre C, Taylor EP, Gittoes C. Barriers and facilitators to disclosing sexual abuse in childhood and adolescence: a systematic review. *Child Abuse Negl* 2017;70:39–52.
12 McElvaney R. Disclosure of child sexual abuse: delays, Non-disclosure and partial disclosure. what the research tells US and implications for practice. *Child Abuse Review* 2013;24.
13 McElvaney R, Greene S, Hogan D. To tell or not to tell? factors influencing young people's informal disclosures of child sexual abuse. *J Interpers Violence* 2014;29:928–47.
14 Winters GM, Colombino N, Schaaf S, *et al*. Why do child sexual abuse victims not tell anyone about their abuse? an exploration of factors that prevent and promote disclosure. *Behav Sci Law* 2020;38:586–611.
15 Adams JA. Medical evaluation of suspected child sexual abuse: 2011 update. *J Child Sex Abus* 2011;20:588–605.
16 Gallion HR, Milam LJ, Littrell LL. Genital findings in cases of child sexual abuse: genital vs vaginal penetration. *J Pediatr Adolesc Gynecol* 2016;29:604–11.
17 Martin J, Alaggia R. Sexual abuse images in cyberspace: expanding the ecology of the child. *J Child Sex Abus* 2013;22:398–415.
18 Faller KC. Anatomical dolls: their use in assessment of children who may have been sexually abused. *J Child Sex Abus* 2005;14:1–21.
19 Domagalski K, Gongola J, Lyon TD, *et al*. Detecting children's true and false denials of wrongdoing: effects of question type and base rate knowledge. *Behav Sci Law* 2020;38:612–29.
20 Lyon TD. Twenty-five years of interviewing research and practice: Dolls, diagrams, and the dynamics of abuse disclosure. *APSAC (American Professional Society on the Abuse of Children) Advisor* 2012;24:14–19.
21 Bensussan P. Forensic psychiatry in France: the Outreau case and false allegations of child sexual abuse. *Child Adolesc Psychiatr Clin N Am* 2011;20:519–32.
22 Faller KC. The Witch-Hunt narrative: introduction and overview. *J Interpers Violence* 2017;32:784–804.
23 Gillies EH. The Witch-Hunt narrative: reflections on knowledge about child sexual abuse and the impact of McMartin in the state of California. *J Interpers Violence* 2017;32:956–66.
24 Lyon TD, Stolzenberg SN, McWilliams K. Wrongful Acquittals of sexual abuse. *J Interpers Violence* 2017;32:805–25.
25 Brilleslijper-Kater SN. *Beyond words: between-group differences in the ways sexually abused and nonabused preschool children reveal sexual knowledge*. Enschede: VU Vrije Universiteit, 2005.
26 van Ham K, Brilleslijper-Kater S, van der Lee H, *et al*. Validation of the sexual knowledge picture instrument as a diagnostic instrument for child sexual abuse: study protocol. *BMJ Paediatr Open* 2020;4:e000799.
27 de Vet HCW, Terwee CB, Mokkink LB, *et al*. *Measurement in medicine: a practical guide*. Cambridge: Cambridge University Press, 2011.
28 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276–82.
29 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
30 de Vet HCW, Mokkink LB, Terwee CB, *et al*. Clinicians are right not to like Cohen's κ. *BMJ* 2013;346:f2125.
31 Lamb ME, Garretson ME. The effects of interviewer gender and child gender on the informativeness of alleged child sexual abuse victims in forensic interviews. *Law Hum Behav* 2003;27:157–71.
32 The European parliamant and council of the European union. *General data protection regulation*. Official Journal of the European Union, 2018.