Journal of Child and Family Studies (2019) 28:3533–3544
https://doi.org/10.1007/s10826-019-01536-z

ORIGINAL PAPER

Safety and Risk Assessment in Child Welfare: A Reliability Study Using Multiple Measures

Annemiek Vial[1] · Mark Assink[1] · Geert Jan J. M. Stams[1] · Claudia van der Put[1]

Published online: 6 September 2019

Abstract

Objectives In the Netherlands, the Actuarial Risk Assessment Instrument for Youth Protection (ARIJ) is a widely used safety and risk assessment instrument in child welfare, although little is known about its reliability. Therefore, this study aimed to determine the reliability of the ARIJ by examining the inter- and intrarater reliability.

Methods For determining interrater reliability, professionals of two Dutch agencies (child and family support, $n = 39$ & child protection, $n = 24$) and master students ($n = 65$) each rated a random selection of 4 out of 24 vignettes. The vignettes were based on actual cases that were handled by the two agencies. For determining intrarater reliability, the professionals rated four vignettes twice with an interval of at least 3 months. Three reliability measures were calculated for each of the three samples: percent agreement, Krippendorff's alpha, and Gwet's Agreement Coefficient.

Results Overall, the items and outcome of the safety assessment instrument showed a moderate or higher than moderate interrater reliability, and a substantial to almost perfect intrarater reliability. In general, the risk assessment items showed a moderate interrater and a substantial-to-high intrarater reliability. The risk assessment outcome had a near perfect interrater reliability and a substantial to almost perfect intrarater reliability.

Conclusions The outcome of both the safety and risk assessment of the ARIJ proved to be reliable and justifies the use of the ARIJ in the Dutch child welfare by professionals with different levels of experience.

Keywords Child maltreatment · Interrater reliability · Intrarater reliability · Safety assessment · Risk assessment

Throughout the years, a large number of instruments for risk and safety assessment have been developed to improve decision making in child welfare. Nowadays, most child welfare workers use one or more of these instruments to guide their decisions on children's current and future safety. To draw conclusions on the quality of these instruments and the decisions made therewith, research examining the validity and reliability of instruments is needed. However, risk and safety assessment instruments are frequently implemented without proper empirical evaluation, both in the Netherlands (Ten Berge, 2008) and internationally (Knoke & Trocme, 2005), and thus limited knowledge is available on instrumental validity and reliability (Barlow, Fisher, & Jones, 2012). In the Netherlands, the Actuarial Risk Assessment Instrument for Youth Protection (ARIJ; Van der Put, Assink, & Stams, 2016) is a widely used instrument for safety and risk assessment, and the number of agencies using the ARIJ is increasing. However, there is very limited well conducted research available about the reliability of the ARIJ.

Since 2015, the ARIJ has been used in the Netherlands to assess the immediate and future safety of children, taking into account the different forms of child abuse and neglect. In the assessment process, a child welfare worker first determines a child's immediate safety, guided by the ARIJ safety assessment instrument. If immediate threats to the child's safety are assessed to be present, immediate measures to safeguard the child are taken. Examples of such measures include a court judge enforcing a (temporary) restraining order on an abusing caregiver, or placing a child in out-of-home care. If legal measures are not deemed

Supplementary information The online version of this article (https://doi.org/10.1007/s10826-019-01536-z) contains supplementary material, which is available to authorized users.

✉ Annemiek Vial
A.Vial@UvA.nl

[1] Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands

necessary, in-home protective measures could be taken to ensure the child's safety. If the immediate safety threats are mitigated or if no immediate safety threats are assessed to be present, the risk of future child maltreatment is assessed with the ARIJ risk assessment instrument. This instrument is actuarial in nature, meaning that the instrument calculates the risk using a fixed algorithm after a child welfare worker has scored all items (i.e., risk factors) of the risk assessment instrument. Based on this estimated risk and the risk factors assessed as present, interventions can be arranged to prevent future harm to the child. Both the safety and risk assessment instrument can be used in the initial assessment stage that follows directly after the registration of a family or child with a child welfare organization. In addition, the instruments can be used for monitoring purposes during treatment.

An instrument's validity and reliability needs to be assessed for evaluating its quality. Determining an instrument's validity involves examining whether an instrument truly measures what it is supposed to measure, whereas determining an instrument's reliability involves examining the consistency of the measurement. Reliability and validity are related psychometric properties in the sense that reliability is a condition for validity, but it is only one of a number of necessary conditions. It should be stressed that an instrument may be valid if it is reliable (Zhao, Feng, Liu, & Deng, 2018). Specifically, a low reliability negatively influences the validity by increasing measurement error. As a result, an instrument with a low reliability cannot properly distinguish between subjects (Kottner et al., 2011). Therefore, both the validity and reliability of an instrument need to be evaluated to determine and improve its quality.

An instrument's reliability can be evaluated by comparing ratings on the same cases by different raters (i.e., interrater reliability or consistency between raters) and by comparing ratings by the same rater on the same case at different times (i.e., intrarater reliability or self-consistency; Gwet, 2014; Koo & Li, 2016). While interrater reliability of safety and risk assessment instruments has been examined to a limited extent in the past, studies on the intrarater reliability of these instruments have not yet been performed. Previous research on the interrater reliability of risk assessment instruments showed mixed and inconclusive results that range from very low to very high, and are mostly expressed in a Kappa or correlation statistic (D'andrade, Austin, & Benton, 2008; Baird, Wagner, Healy, & Johnson, 1999; Barber, Shlonsky, Black, Goodman, & Trocmé, 2008; Bartelink, De Kwaadsteniet, Ten Berge, & Witteman, 2017; Cash 2001; Knoke & Trocme, 2005). Baird, Wagner, Healy, and Johnson (1999) specifically compared the interrater reliability of an actuarial risk assessment instrument to the interrater reliability of two consensus based instruments, and found the former to be more reliable. Risk assessment items that are more objective and concrete (e.g., the age of a child) tend to have a higher interrater reliability than more subjective items (e.g., determining whether a child was adequately supervised; D'andrade, Austin, & Benton, 2008; Knoke & Trocme, 2005).

To our knowledge, only two studies examined the interrater reliability of a child safety assessment instrument. One study performed in the Netherlands showed a slight to fair interrater reliability of the items of a Dutch safety assessment instrument, and a moderate interrater reliability of the safety outcome of that same instrument (Bartelink et al., 2017). Orsi et al. (2014) examined the interrater reliability of the items of several safety assessment instruments, and found mixed interrater reliability of the items, varying from slight to substantial. However, the results of both Bartelink et al. (2017) and Orsi et al. (2014) should be interpreted cautiously, as the results seemed to have been negatively influenced by the "prevalence problem". For this reason, Orsi et al. (2014) do no draw conclusions on which items of the safety assessment instrument they examined are most reliable. This problem (also known as "kappa paradox" or "paradox of high agreement but low reliability"), entails that reliability measures (e.g., Kappa, Krippendorff's alpha) are underestimated in case of a low or high item prevalence (Cicchetti & Feinstein, 1990; Hallgren, 2012; Lantz & Nebenzahl, 1996; Zhao et al., 2018). A low or high item prevalence usually goes hand in hand with high agreement between raters, because the variety in chosen response categories is low. However, most reliability measures handle low variety incorrectly, and as a result, the reliability measures falsely indicate a low reliability despite a high agreement. In other words, the calculations of these reliability measures do not adjust for low variance accordingly, which causes the prevalence problem. Studies on reliability of instruments used in child welfare may as well have been influenced by this problem (D'andrade, Austin, & Benton, 2008; Baird, Wagner, Healy, & Johnson, 1999; Barber, Shlonsky, Black, Goodman, & Trocmé, 2008; Cash 2001; Knoke & Trocme, 2005), since Cohen's kappa or Krippendorff's alpha were estimated in these studies. These statistics were also estimated in the studies of Bartelink et al. (2017) and Orsi et al. (2014).

The prevalence problem, and the statistics that may or may not be influenced by it, have been widely discussed in literature. Gwet (2002) formulated a possible solution to this problem and developed an agreement coefficient (Gwet's AC) that should be robust to item prevalence. Multiple studies in different research areas have indeed demonstrated that Gwet's AC is less influenced by item prevalence than Cohen's Kappa (see, for instance, Ait Lbacha et al., 2017; Ko et al., 2013; Wongpakaran, Wongpakaran, & Gwet, 2013; Zec, Soriani, Comoretto, &

Baldi, 2017). Ait Lbacha et al. (2017) and Ko et al. (2013) calculated Gwet's AC and Kappa in their studies on the agreement between different detection methods for an infection in ruminants and the agreement between clinicians observing pulse signs in stroke patients, respectively. The main purpose of these studies was not to compare Gwet's AC and Kappa, but the results did show a faulty low kappa in case of a low prevalence, whereas Gwet's AC is stable with varying prevalence rates. Wongpakaran et al. (2013) and Zec et al. (2017) specifically compared the performance of Gwet's AC and Kappa using data on personality disorder diagnoses and clinical trial quality assessments. Both studies revealed that Gwet's AC is a more stable measure than Kappa. Therefore, we decided to calculate Gwet's agreement coefficient to examine the reliability of the ARIJ safety and risk assessment instruments. To produce a more extensive overview of the reliability of both instruments, Krippendorff's alpha was also calculated, which is a frequently used statistic in research on reliability (Feng, 2014). This measure was proposed by Krippendorff as a "standard reliability measure" (Hayes & Krippendorff, 2007). By calculating both statistics, Krippendorff's alpha's robustness to item prevalence could be compared to Gwet's AC robustness to item prevalence. This way, it is possible to draw inferences on the reliability of the ARIJ using the statistic that is least influenced by item prevalence, and gives the most accurate depiction of reliability.

To examine the reliability of the two instruments in the broadest possible sense, the reliability was established at both the individual item and outcome level. Generally, the reliability of instruments is often only determined at the risk level and not at the item level (Orsi et al., 2014). However, for making improvements to the instrument's content, the reliability of each item is essential. Items showing a low reliability could be adjusted or may even be removed to improve the instrument's overall reliability, and possibly its validity.

Additionally, the reliability of the two instruments was determined for (vignettes of) cases that were handled by two different agencies, because the ARIJ is used by different agencies that all provide different types of child and/or family care. Sutherland et al. (2012) showed that case characteristics influenced the interrater reliability of risk assessment judgments. Cases with moderate levels of risk and of moderate complexity had a lower interrater reliability than cases with high or low levels of complexity and risk. By using cases from different agencies that differed in the levels of complexity and risk in the current study, we could explore the reliability of the ARIJ in different settings.

An unresolved issue is whether or not the reliability of a risk assessment is influenced by characteristics of the rater. Some studies on risk assessment instruments showed that a rater's experience can influence the rating (De Vogel & De Ruiter, 2006; Quesada, Calkins, & Jeglic, 2014; Penney, McMaster, & Wilkie, 2014), and the predictive validity of that rating (Webster et al., 2006; Teo, Holley, Leary, & McNiel, 2012). However, only one study showed that a rater's experience may influence the interrater reliability of risk assessment judgments (Sutherland et al., 2012). Specifically, Sutherland et al. showed that the interrater reliability was lower when professionals were less trained in conducting assessments. In all these studies, a professionals' experience with a specific instrument, experience with risk assessment instruments in general, the extent to which a professional has been trained in using these instruments, and clinical experience have been operationalized in different ways. As the literature suggests that rater experience may influence interrater reliability, the present study also examined the influence of experience on the interrater reliability of the ARIJ.

The central aim of this study was to examine the reliability (i.e., the interrater and intrarater reliability) of the items and the outcome of the ARIJ safety and risk assessment instruments. This was examined by asking professionals of two different child welfare agencies as well as master students to rate vignettes using the ARIJ. These vignettes were based on real cases that were handled by the two organizations. Besides this central aim, we compared the reliability of structured clinical judgments of risk to actuarially estimated risks, we examined the influence of rater type and vignette type on the interrater reliability of the safety and risk assessment outcome, and we examined Krippendorff's alpha and Gwet's AC's robustness to item prevalence.

Method

Participants

Child and family support agency participants

Initially, 59 professionals volunteered to participate. In total, 39 professionals (5 men, 34 women; $M_{age} = 38.90$ years, $SD = 11.39$; age range: 22–62 years) completed the questionnaire at time 1. Additionally, 5 professionals partly filled out the time 1 questionnaire and 15 professionals never opened the questionnaire. Although five questionnaires were incomplete, we chose to retain these questionnaires in analyzing the interrater reliability. In this way, we could determine the interrater reliability on as much ratings as possible. However, the demographic characteristics of these participants were not available. For the intrarater reliability measures, these ratings had to be excluded, because it is impossible to determine the reliability with one measurement only. At time 1, 13% of the

participants indicated that they had been trained in the ARIJ safety and risk assessment, and 87% had used the ARIJ between 0 and 5 times. The other 13% of the participants had more experience with the instrument. On average, the professionals had been working in their current position for 7 years ($SD = 7.11$; range: 1–25 years) and in child care in general for 12 years ($SD = 7.47$; range: 1–25 years).

At time 2 (at least 3 months after time 1), a total of 32 professionals participated. The mean test–retest interval was 18 weeks ($SD = 3.29$; time range: 12.99–26.14 weeks). Of the professionals, 3% indicated being trained in using the ARIJ safety and risk instruments during the test–retest interval, and as a result, 19% had been trained at time 2. Approximately one third of the professionals (31%) indicated that they had been using the instrument during the test–retest interval. For each vignette, the participants were asked on a 5-point Likert scale how much they remembered from their time 1 participation. The participants indicated that they partly recognized the vignettes from time 1 ($M = 2.06$; $SD = 1.08$; 1 = I recognize the cases from time 1; 5 = I do not recognize the cases from time 1), that they did not remember how they answered the items at time 1 ($M = 4.03$; $SD = 0.70$; 1 = I remember my answers on the items from time 1; 5 = I do not remember my answers on the items from time 1), and that they did not remember how they had assessed the risk of future maltreatment at time 1 ($M = 3.88$; $SD = 0.83$; 1 = I remember my answer from time 1; 5 = I do not remember my answer from time 1).

Child protection agency participants

In total, 24 participants (3 men; 21 women; $M_{age} = 39.92$ years; $SD = 11.23$; age range: 26–64 years), of the 36 professionals who were appointed for participation, completed the questionnaire at time 1. Additionally, 3 participants partly filled out the time 1 questionnaire. The answers of these 3 participants were included in the data analysis, but demographic characteristics of these participants were not available. In this way, the interrater reliability could be determined using as much ratings as possible. For the intrarater reliability measures, the five incomplete questionnaires were excluded from the analyses. At the start of this study, 88% of the professionals were trained in using the instrument. Additionally, 63% of the participants had been using the ARIJ safety assessment instrument for 2 years, whereas the other participants had used the instrument for (at least) 1 year. The professionals had been working in their current position for an average of 6 years ($SD = 6.07$; range: 1–20 years) and in child care for 9 years ($SD = 6.55$; range: 1–23 years).

A total of 19 participants completed the questionnaire at time 2. The mean test–retest interval was 20 weeks ($SD = 4.04$; time range: 13.14–27.01 weeks). During the test–retest interval, 5% of the professionals had been trained in using the instrument. As a result, 94% of the professionals had been trained at time 2. All participants had been using the instrument during the test–retest interval. The participants indicated that they partly recognized the vignettes from the first measurement ($M = 2.56$; $SD = 1.04$; 1 = I recognize the cases; 5 = I do not recognize the cases), that they did not remember how they had answered the items at time 1 ($M = 4.17$; $SD = 0.62$; 1 = I remember my answers on the items from time 1; 5 = I do not remember my answers on the items from time 1), and that they did not remember how they had assessed the risk of future maltreatment at time 1 ($M = 4.17$; $SD = 0.51$; 1 = I remember my answer from time 1; 5 = I do not remember my answer from time 1).

Master student participants

In total, 300 students were approached for research participation. Of this group, 65 students (3 men, 62 women, $M_{age} = 24.77$ years, $SD = 3.70$, age range: 21–46 years) completed the questionnaire. Additionally, 17 students partly filled out the questionnaire. The ARIJ instrument was never used by 89% of the students and 75% had never conducted a risk assessment. In total, 45% of the students had worked in child care before, and for most students this was in the form of an internship.

Procedure

Procedure child and family support agency

A call for research participation was placed on the child and family support (CFS) agency's intranet. The professionals willing to participate contacted the researchers and received a link to a digital online questionnaire. Three months after completing the first questionnaire, the participants received the link to the second questionnaire. In completing both questionnaires, it was possible to pause and save the questionnaire and continue at a later, more convenient moment. The average time to fill out each of the digital questionnaires was estimated to be one hour. The participants received a 20 Euro voucher after successfully completing both measurements.

Procedure child protection agency

The procedure was similar to the procedure described for the child and family support agency, with the exception of participant recruitment. Rather than letting professionals volunteer to participate, several teams of child protection (CP) workers were appointed to participate by the CP agency.

Procedure master students

Students enrolled in the Dutch Master's Program Forensic Child and Youth Care (2016-2017 and 2017-2018) at the University of Amsterdam were approached via email for a single research participation at time 1. The students who volunteered rated four random vignettes using an online questionnaire. Each student received a 10-euro voucher after having completed the questionnaire.

Measures

ARIJ safety assessment instrument

The ARIJ safety assessment instrument was developed to help determining immediate child safety (Van der Put, Assink, & Stams, 2016). The instrument consists of eight items, such as: 'The child is (in immediate danger of) being physically abused' and 'The child is (in immediate danger of) being sexually abused'. When an item is considered to be present, the child immediately needs to be safeguarded to prevent harm. Each of the eight items can be responded to with one of three categories: "Yes" (implying the threat described in the item is present), "No" (implying the threat described in the item is not present), and "Unknown" (implying there is insufficient information available at time of the assessment for a proper response). When at least one of the items is answered with "yes", the instrument concludes that a child should be safeguarded immediately. If at least one of the items is answered with 'unknown', the instrument concludes that further information about the child's safety needs to be obtained as soon as possible. When all items are answered with 'no', the instrument concludes that there are no concerns about the child's immediate safety.

ARIJ risk assessment instrument

The ARIJ risk assessment instrument is an actuarial risk assessment instrument that helps to determine the risk of future maltreatment, taking into account the different forms of child abuse and neglect (Van der Put, Assink, & Stams, 2016). The professional determines the presence or absence of each risk factor that is measured in each item of the instrument. Based on the responses to all items, the instrument calculates the risk for future child maltreatment. A "dynamic risk" is also calculated based on all the responses to the items in which a dynamic risk factor is measured. Both risks are expressed as low, medium, or high. The participants in this study were not aware of both actuarial risk outcomes, since these were calculated after data collection. In total, the instrument comprises 31 items, which are grouped in aspects of the current child safety situation (9 items), risk factors (14 items), and experimental items that are part of the instrument for research purposes (8 items). All items can be responded to with one of three categories: "Yes" (implying that the risk factor is present), "No" (implying that the risk factor is absent), and "Unknown" (implying there is insufficient information available at time of the assessment for a proper response). The ARIJ risk assessment instrument usually does not include a structured clinical judgment of risk. However, for the purpose of this study we included a question in which participants clinically assessed risk for future maltreatment based on how they assessed the risk factors.

Vignettes

Child and family support vignettes

Twelve short anonymous vignettes were used. The vignettes were clustered together in groups of four vignettes with different kinds of child maltreatment. Each participant received a randomly assigned group of four vignettes. Additionally, the order in which the four vignettes were presented to the participants was randomized. The vignettes had been created and used in previous research by Bartelink et al. (2017). The vignettes were based on real cases and described a variety of family compositions, social backgrounds, cultural backgrounds, problems (physical, sexual, emotional abuse, and neglect), and severity levels. The vignettes contained an average of 602 words ($SD = 94$; range: 453–724). A fictional vignette, which is similar to the vignettes used in this study, can be found in Supplementary Appendix A.

For each vignette, the participants were asked to judge its content on a 5-point Likert scale. Participants indicated that the vignettes were similar to their cases in daily practice ($M = 2.41$; $SD = 0.98$; 1 = strongly resemble cases that I handle in daily practice; 5 = do not resemble cases I handle in my daily practice), as severe as cases in their daily practice ($M = 2.85$; $SD = 0.73$; 1 = much less severe; 5 = much more severe), and included a similar amount of information ($M = 3.32$; $SD = 0.76$; 1 = much less information than in my daily practice; 5 = much more information than in my daily practice).

Child protection vignettes

Short vignettes were constructed by removing unnecessary information from twelve selected cases of the CP agency's official records. The vignettes contained an average of 609 words ($SD = 94$; range: 470–761). Each item of the safety and risk assessment instrument was represented in at least one vignette. The vignettes were read and checked by professionals of the CP agency (other than the participants

in this study) to assure they were representative of their daily practice. Since this agency usually handles more cases with children in immediate danger, the vignettes for this CP agency were expected to have a higher prevalence of the safety assessment items than the CFS vignettes. Participants indicated that the vignettes were similar to the cases in their daily practice ($M = 2.10$; $SD = 0.92$; 1 = strongly resemble; 5 = do not resemble), as severe as their cases in daily practice ($M = 3.04$; $SD = 0.34$; 1 = much less severe; 5 = much more severe), and included a similar amount of information ($M = 2.72$; $SD = 0.74$; 1 = much less information; 5 = much more information).

Student vignettes

The students received four randomly assigned vignettes out of the total of 24 vignettes that were used in this study. The CP and CFS vignettes were equally distributed among participants.

Qualtrics questionnaire

The questionnaire contained 4 out of 24 vignettes, and all items of the ARIJ safety and risk assessment part of the instrument per vignette. In a final question, all participants were asked to clinically assess the overall risk of future maltreatment for each vignette. At the first measurement, a number of control questions were asked to the participating professionals, for example about the similarity of the vignettes to the cases in their daily practice. Items about demographic characteristics and participants' clinical experience were only part of the questionnaire at the first measurement. In the time 2 questionnaire, several control questions were asked, such as whether participants remembered their answers to the questions of the first measurement. This question was asked to determine whether the results may have been influenced by recall bias. If participants state that they clearly recall how they answered the questionnaire at time 1, the intrarater reliability results must be interpreted carefully.

Data Analysis

Interrater reliability

To determine the interrater reliability of the safety and risk assessment items, three different statistics were calculated: the percent agreement, Krippendorff's α (with a bootstrap confidence interval) and Gwet's *AC1* (with a 95% confidence interval). First, the percent agreement was calculated to measure the actual agreement without corrections. The percent agreement was determined by calculating the mean of the percent agreement for the most frequently given answer per vignette, which is in accordance with the method of McHugh (2012). Since percent agreement does not correct for chance, this statistic usually is higher than the true agreement (Feng, 2015; Hallgren, 2012). It can be assumed that responding to the items is rather difficult for the participants, as they need to decide on the presence or absence of complex risk factors. Therefore, it is likely that participants sometimes guess the correct answer in responding to items. Krippendorff's α and Gwet's *AC1* were calculated to adjust for this guessing. These statistics are suitable for nominal data with multiple coders (Feng, 2015). Fleiss' Kappa is also a frequently used statistic for nominal data with multiple coders. However, the current study contains data missing by design, as none of the vignettes was rated by all the raters. It is therefore undesirable to calculate Fleiss' Kappa (Hallgren, 2012; Zapf, Castell, Morawietz, & Karch, 2016).

There has been a great deal of discussion about reliability statistics and their susceptibility to the prevalence problem (Cicchetti & Feinstein, 1990; Gwet, 2002/2008; Feng, 2015). For this reason, the prevalence of each response category of each item was calculated to examine its influence on Krippendorff's α and Gwet's *AC1*. With an item prevalence around 50% (for an item with two response categories), all statistics should perform alike (Gwet, 2008). By calculating the four statistics (prevalence, percent agreement, Krippendorff's α, and Gwet's *AC1*), the influence of the item's prevalence on reliability estimates was examined, and a comprehensive depiction of the interrater reliability could be obtained. To determine the interrater reliability of the risk assessment outcome, Gwet's *AC2* was calculated instead of Gwet's *AC1*, because the risk assessment outcome is ordinal. The *AC2* statistic corrects for partial agreement, which occurs when comparing ordinal variables (Gwet, 2008).

All interrater reliability analyses described above were conducted separately for four groups; child and family support professionals (CFS professionals), child protection professionals (CP professionals), students who rated the child and family support vignettes (CFS students), and students who rated the child protection vignettes (CP students). The ratings of the CP professionals, CFS professionals, and students were analyzed separately to explore the influence of rater type on the interrater reliability. Additionally, the vignettes for the two agencies were analyzed separately for the students to examine the influence of vignette type on the interrater reliability.

Intrarater reliability

The intrarater reliability was determined for each item of the safety and risk assessment by calculating percent agreement, Krippendorff's α (with a bootstrap confidence

interval), and Gwet's *AC1* (with a 95% confidence interval). The two ratings per participant (at time 1 and time 2) were paired and analyzed as if there were two raters for a vignette's item (Gwet, 2014). Additionally, Gwet's *AC2* (with a 95% confidence interval) was calculated for the ordinal risk assessment outcome. Finally, the prevalence (i.e., the prevalence of each response category) was also calculated for each item. These statistical analyses were conducted for the CFS professionals and the CP professionals separately.

Overview reliability analysis

The statistical programs R (version 1.0.153) and SPSS (version 24) were used to conduct the analyses. The R-package irr (Gamer, Lemon, Fellows, & Singh, 2015) and the kripp.boot package (Proutskova & Gruszczynski, 2017) were used to calculate Krippendorff's α and its confidence interval. To calculate Krippendorff's α's confidence interval, 1000 bootstraps were performed. Gwet's R-script was used to calculate Gwet's *AC1* and *AC2* (Gwet, 2017). The following guidelines were used for interpreting the strength of Krippendorff's α and Gwet's AC: 0.00–0.20 = slight reliability, 0.21–0.40 = fair reliability, 0.41–0.60 = moderate reliability, 0.61–0.80 = substantial reliability, 0.81–1.00 = almost perfect reliability (Landis & Koch, 1977).

The reliability of structured clinical judgments of risk versus actuarial estimated risks

The interrater reliability measures on the three risk assessment outcomes (actuarial risk based on all risk factors, dynamic actuarial risk based on the dynamic risk factors, and risk based on structured clinical judgment) were compared with each other by conducting t-tests on the risk outcomes within each group. Additionally, the intrarater reliability of the three different risk assessment outcomes were compared with each other for both professional groups. The R-script paired t-test for agreement coefficients (Gwet, 2016) was used for all these tests.

The influence of rater type on interrater reliability of safety and risk assessment outcomes

T-tests were conducted to compare the ratings by students and professionals. By doing this, we were able to examine the influence of rater experience on the interrater reliability of the safety and risk assessment outcomes. The CFS students were compared to the CFS professionals, and the CP students were compared to the CP professionals. Similarly, the R-script paired *t*-test for agreement coefficients (Gwet, 2016) was used for these tests.

The influence of vignette type on the interrater reliability of the ARIJ

The comparison of the ratings by the CFS students to the ratings of the CP students allowed us to examine the effect of vignette type on the interrater reliability of the risk and safety assessment outcomes. Again, the R-script paired t-test for agreement coefficients (Gwet, 2016) was used.

Results

The main aim of our study was to determine the reliability of the safety and risk assessment of the ARIJ. However, before interpreting reliability estimates, the robustness of Gwet's AC and Krippendorff's α to a low item prevalence was examined to determine the most stable measure of reliability. Supplementary Appendix B shows how Gwet's AC and Krippendorff's α relate to each other across different percentages of agreements. Each figure of Supplementary Appendix B represents a different fixed percent agreement. All the interrater reliability outcomes from the current study were included in these figures. In general, the figures reveal that Gwet's AC and Krippendorff's α can be quite different even though the percent agreement is very similar, and thus, Gwet's AC and Krippendorff's α should also be very similar. The figures with a fixed percent agreement between 60% and 90% show that Krippendorff's α decreased when item prevalence dropped below 20% or increased when item prevalence reached above 80%. This indicates that Krippendorff's α is underestimated in case of a low or high item prevalence. In contrast, Gwet's AC shows only a slight increase when item prevalence decreased or increased. An example of the influence of item prevalence on Krippendorff's α and Gwet's AC can be seen in item 4 of the safety assessment instrument for the child and family support vignettes. This item has a percent agreement of 94%, whereas Krippendorff's α is 0.02. Gwet's AC1 does seem to be in line with the percent agreement given that its value is 0.88 (see Supplementary Appendix C). Overall, Gwet's AC appeared to be a more stable measure than Krippendorff's α. For that reason, only Gwet's AC was used when determining the inter- and intrarater reliability of the ARIJ instruments. Refer to Supplementary Appendix C-H for a complete overview of all the measures.

Interrater Reliability of the ARIJ Safety Assessment Instrument

The interrater reliability of most safety assessment items varied between moderate and substantial (*AC1* = 0.41–0.80; see Supplementary Appendix C and D for all the interrater reliability measures of the ARIJ safety assessment

instrument for the four different groups). The items with the highest interrater reliability were about child abduction and honor-related violence ($AC1 = 0.70$–0.94; item 4) as well as domestic violence ($AC1 = 0.68$–0.94; item 7), which had a substantial to almost perfect reliability. The parental availability item (item 8) showed the lowest interrater reliability ($AC1 = 0.12$–0.46), which was slight to moderate. The interrater reliability of the safety assessment outcome was moderate to substantial ($AC1 = 0.40$–0.68).

Interrater Reliability of the ARIJ Risk Assessment Instrument

The interrater reliability of most risk assessment items varied between fair and substantial ($AC1 = 0.21$–0.80; see Supplementary Appendix E and F for all the interrater reliability measures of the risk assessment instrument for the four groups). The item 'the child is younger than 5 years old' (item 30) showed the best interrater reliability ($AC1 = 0.91$–0.97), which was almost perfect. The item 'caregiver has a history of abusing a child' (item 17) had the lowest reliability, which was a slight interrater reliability ($AC1 = 0.12$–0.18). The actuarial risk outcome based on all risk factors showed a substantial to almost perfect interrater reliability ($AC2 = 0.80$–0.96). In contrast, the actuarial risk outcome based on the dynamic risk factors showed a mixed interrater reliability ($AC2 = 0.39$–0.84), which was fair to almost perfect. Finally, the clinical risk outcome displayed a moderate to substantial interrater reliability ($AC2 = 0.56$–0.62).

Intrarater Reliability of the ARIJ Safety Assessment Instrument

Almost all safety assessment items showed substantial intrarater reliabilities ($AC1 = 0.61$–0.80; see Supplementary Appendix G and H for the intrarater reliability measures of the safety assessment instrument of the CFS vignettes and CP vignettes). The items related to child abduction and honor-related violence ($AC1 = 0.87$ and 0.88; item 4) as well as the domestic violence item ($AC1 = 0.84$ and 0.85; item 7) showed the highest intrarater reliability, which was almost perfect. Finally, the items about physical abuse (item 1) and parental availability (item 8) showed moderate to substantial ($AC1 = 0.41$ and 0.61) and fair to moderate intrarater reliability ($AC1 = 0.25$ and 0.68). The intrarater reliability of the safety assessment outcome was moderate to substantial ($AC1 = 0.57$ and 0.75).

Intrarater Reliability of the ARIJ Risk Assessment Instrument

Almost all risk assessment items had a moderate to substantial intrarater reliability ($AC1 = 0.41$–0.80; see Supplementary Appendix G and H for the intrarater reliability measures of the risk assessment instrument of the CFS vignettes and CP vignettes). One item showed almost perfect intrarater reliability ($AC1 = 0.92$ and 0.94) and asked whether or not a child is younger than 5 years old (item 30). Both the actuarial risk outcome based on all risk factors and the actuarial risk outcome based on the dynamic risk factors showed a substantial to almost perfect intrarater reliability ($AC2_{act} = 0.80$ and 0.93; $AC2_{dynact} = 0.62$ and 0.82). Finally, the intrarater reliability of the clinical risk outcome was substantial ($AC2 = 0.66$ and 0.79).

The Interrater Reliability of the Structured Clinical Risk Outcome vs Actuarial Risk Outcomes

The interrater reliability of the actuarial risk including all risk factors was higher than the clinical risk for three of the four participant groups (CFS professionals $AC2_{act} = 0.84$, $AC2_{clin} = 0.56$, $t = -3.50$, $p = 0.005$; CP professionals $AC2_{act} = 0.95$, $AC2_{clin} = 0.59$, $t = -6.00$, $p < 0.001$; CP students $AC2_{act} = 0.96$, $AC2_{clin} = 0.62$, $t = -4.53$, $p = 0.001$). Only the interrater reliability of the actuarial risk of the CFS students ($AC2 = 0.80$) was not significantly higher than the clinical risk ($AC2 = 0.59$; $t = -2.02$, $p = 0.07$). Similarly, in three groups, the actuarial risk based on all risk factors had a higher interrater reliability than the actuarial dynamic risk (CFS professionals $AC2_{act} = 0.84$, $AC2_{dynact} = 0.39$, $t = -4.17$, $p = 0.002$; CFS students $AC2_{act} = 0.80$, $AC2_{dynact} = 0.45$, $t = -2.1$, $p = 0.03$; CP students $AC2_{act} = 0.96$, $AC2_{dynact} = 0.72$, $t = -3.43$, $p = 0.006$). For the CP professionals, the interrater reliability of the actuarial risk based on all risk factors ($AC2 = 0.95$) was similar to the dynamic actuarial risk ($AC2 = 84$, $t = -1.55$, $p = 0.15$). The dynamic actuarial risk has the same interrater reliability as the structured clinical risk for most groups (CFS professionals $AC2_{dynact} = 0.39$, $AC2_{clin} = 0.46$, $t = -1.42$, $p = 0.18$; CFS students $AC2_{dynact} = 0.45$, $AC2_{clin} = 0.59$, $t = -0.72$, $p = 0.48$; CP students $AC2_{dynact} = 0.72$, $AC2_{clin} = 0.62$, $t = 1.45$, $p = 0.17$). Only for the CP professionals, the interrater reliability of the dynamic actuarial risk ($AC2 = 0.84$) was higher than the structured clinical risk ($AC2 = 0.59$, $t = 2.35$, $p = 0.04$).

The Intrarater Reliability of the Structured Clinical Risk Outcome vs Actuarial Risk Outcomes

The intrarater reliability of the actuarial risk based on all risk factors was higher than the intrarater reliability of clinical risk for both the CFS professionals ($AC2_{act} = 0.80$, $AC2_{clin} = 0.66$, $t = -2.21$, $p = 0.03$) and the CP professionals ($AC2_{act} = 0.93$, $AC2_{clin} = 0.79$, $t = -2.35$, $p = 0.02$). Similarly, the intrarater reliability of the actuarial risk was higher than the intrarater reliability of dynamic

actuarial risk for the CFS professionals ($AC2_{act} = 0.80$, $AC2_{dynact} = 0.60$, $t = -2.73$, $p = 0.01$) and the CP professionals ($AC2_{act} = 0.93$, $AC2_{dynact} = 0.82$, $t = -2.04$, $p < 0.05$). The intrarater reliability of the dynamic actuarial risk was similar to the intrarater reliability of clinical risk for the CFS professionals ($AC2_{dynact} = 0.60$, $AC2_{clin} = 0.66$, $t = -0.49$, $p = 0.62$) and CP professionals ($AC2_{dynact} = 0.82$, $AC2_{clin} = 0.79$, $t = 0.40$, $p = 0.69$).

Influence of Rater Experience on Interrater Reliability of Safety and Risk Assessment Instrument Outcomes

Overall, the reliability of ratings by students and professionals were comparable. Specifically, the interrater reliability of the safety assessment outcome was similar for the CFS vignettes ($AC1_{stud} = 0.49$, $AC1_{prof} = 0.40$, $t = 1.45$, $p = 0.17$) and the CP vignettes ($AC1_{stud} = 0.65$, $AC1_{prof} = 0.68$, $t = -0.25$, $p = 0.80$). Additionally, the interrater reliability of the actuarial risk was similar for students and professionals rating both types of vignettes: CFS ($AC2_{stud} = 0.80$, $AC2_{prof} = 0.84$, $t = -1.48$, $p = 0.17$) and the CP ($AC2_{stud} = 0.96$, $AC2_{prof} = 0.95$, $t = 0.60$, $p = 0.56$). Similarly, the interrater reliability of the dynamic actuarial risk was similar for students and professionals who rated the CFS vignettes ($AC2_{stud} = 0.45$, $AC2_{prof} = 0.39$, $t = 0.59$, $p = 0.56$). However, the interrater reliability of dynamic actuarial risk for the CP professionals was higher than for the CP students ($AC2_{stud} = 0.72$, $AC2_{prof} = 0.84$, $t = -2.64$, $p = 0.02$). The interrater reliability of the clinical risk was similar for the students and the professionals who rated the CFS vignettes ($AC2_{stud} = 0.59$, $AC2_{prof} = 0.56$, $t = 0.43$, $p = 0.67$) and the CP vignettes ($AC2_{stud} = 0.62$, $AC2_{prof} = 0.59$, $t = 0.28$, $p = 0.78$).

Influence of Cases from Different Agencies on Interrater Reliability of Safety and Risk Assessment Instrument Outcomes

The interrater reliability of the safety assessment outcome did not differ between the students who rated CFS vignettes or CP vignettes ($AC2_{cfs} = 0.49$, $AC2_{CP} = 0.65$, $t = 0.94$, $p = 0.37$). Similarly, the interrater reliability of all risk assessment outcomes was similar for the CFS students and the CP students: actuarial risk ($AC2_{cfs} = 0.80$, $AC2_{cp} = 0.96$, $t = -1.76$, $p = 0.11$), dynamic actuarial risk ($AC2_{cfs} = 0.45$, $AC2_{cp} = 0.72$, $t = -1.16$, $p = 0.27$), and clinical risk ($AC2_{cfs} = 0.59$, $AC2_{cp} = 0.62$, $t = -0.18$, $p = 0.86$).

Discussion

The results did not allow to draw a single firm conclusion about the reliability of the safety and risk assessment instruments of the ARIJ. First, the safety assessment items showed in general a reasonable reliability. However, intrarater reliability of the items was higher than the interrater reliability of the items. The reliability of the safety assessment items was mostly moderate (50%), substantial (25%), or almost perfect (16%) between the raters, but within the raters the reliability was mostly substantial (56%) or almost perfect (31%). Both the intrarater and interrater reliability of the safety assessment outcome was moderate to substantial. Second, the interrater reliability of the risk assessment items was lower than the interrater reliability of the safety assessment items; only 43% of the measures on the risk assessment items had a moderate interrater reliability, 11% had a substantial interrater reliability, and 4% had an almost perfect interrater reliability. Similar to the safety assessment items, the reliability of the risk assessment items was higher within the raters than the reliability between the raters, since the intrarater reliability was mainly moderate (50%) or substantial (35%). The interrater reliability of the actuarial risk outcome was almost perfect, and its intrarater reliability was substantial to almost perfect.

The reliability of the actuarial risk outcome was higher than the reliability of structured clinical risk judgment. This result supports the finding of Baird et al. (1999), who showed that the actuarial risk assessment instruments had the highest interrater reliability. The current findings expand on prior work by showing that the intrarater reliability was also higher for the actuarial risk than for the structured clinical risk. The high reliability of the actuarial outcome is likely due to the fact that the actuarial risk is based on the total number of risk factors that are present in a case, and therefore, differences on ratings of single risk factor do not impact the actuarial risk to a substantial extent. However, this advantage of an actuarial risk outcome appears not to hold for the actuarial dynamic risk. The reliability of the dynamic actuarial risk outcome was mostly lower than the reliability of the actuarial risk. This difference in reliability may be caused by three differences between the risk outcomes. First, the dynamic actuarial risk is based on less risk factors (13 risk factors) than the actuarial risk (23 risk factors). Therefore, the differences in ratings of single risk factors could make a greater impact on the dynamic actuarial risk. Second, the dynamic risk factors are less factual and may be less straightforward to answer (D'andrade, Austin, & Benton, 2008; Knoke & Trocme, 2005), and therefore less reliable. In line with this, the highest reliability was found for a static risk factor (i.e., the child is younger than 5 years old). Finally, the distribution of low, medium, and high risk in the actuarial risk outcome was uneven; most ratings were high risk. Since most ratings were high risk, it follows logically that the reliability is high due to a low variety. Conversely, the dynamic actuarial risk showed more variety in risk level, which may have caused a

lower reliability. Future research should examine what causes this difference in reliability of the actuarial and the actuarial dynamic risk. By knowing what causes this difference it may be possible to develop a reliable actuarial risk based on only dynamic risk factors. A dynamic actuarial risk could be particularly useful for child welfare services in assessing changes in risk.

Both rater experience and cases from different agencies did not affect the interrater reliability of the safety and risk assessment outcomes. It is promising that the reliability of the instruments was similar for cases from both agencies and for professionals with varying levels of experience. This may imply that the tools are usable in these different circumstances. However, given the design of our study, we should be careful with generalizing these results to the influence of specific types of rater experience on the interrater reliability. The two groups of students and professionals that we compared to one another in the current study differ in their clinical experience. The professionals have far more work experience in child welfare. The two groups of professionals also varied in their experience, for example in their experience with the ARIJ and the degree to which professionals were trained in using the ARIJ. To specifically determine the effect of types of experience on reliability, further research should examine this in more detail. Although we did not examine the effect of specific types of experience on reliability, these results seem to be in contrast with Sutherland et al. (2012), who found that lower levels of specialist training negatively affected interrater reliability.

The influence of agency type or setting on the reliability should also be considered in more detail in further research. The fact that the reliability was similar in two different settings may indicate that the instruments can be utilized in both settings. However, it is essential to first evaluate the validity of the instruments in these settings, before it can be stated that the instruments are applicable in both child and family support and child protection.

Gwet's AC proved to be a more stable reliability measure than Krippendorff's alpha. Krippendorff's alpha was inaccurately low if the item prevalence was high or low, whereas Gwet's AC slightly increased when item prevalence was high or low. This result is in line with previous studies showing that Gwet's AC is the most stable measure (Ait Lbacha et al., 2017; Ko et al., 2013; Wongpakaran, Wongpakaran, & Gwet, 2013; Zec., Soriani, Comoretto, & Baldi, 2017). Future research on the reliability of safety and risk assessment instruments should carefully consider which reliability measure(s) should be used, especially in case of a low item prevalence. In specifically the latter case, Gwet's AC should be considered as a reliability measure, so that the effect of item prevalence on reliability estimates can be minimized.

Limitations and Future Research

Some shortcomings need to be mentioned. Although risk levels varied in the original cases, the participants rated the vignettes, especially the child protection vignettes, with a high risk. As a result, there is insufficient variety in risk levels, which may have caused an inflated reliability of the actuarial risk. Further research should consider including more cases with a low or medium risk. However, this may entail that the cases will be more artificial, and it is important to keep in mind that the eventual risk levels still depend on how participants rate the risk factors.

Another important point is that a participant responding to an item with 'yes' may differ more from a participant responding to that same item with 'no' than from a participant responding with 'unknown'. Each item of the ARIJ can be answered with the following response categories: 'yes', 'no' and 'unknown'. Participants deeming a threat to be present and participants deeming a threat to be absent clearly disagree with each other, whereas a participant responding with "unknown" may in reality have a tendency towards 'yes' or 'no', but has reason to believe that insufficient evidence or information is presented for a 'yes' or a 'no'. Since the response category 'unknown' is essential in these instruments, there is, to our knowledge, no alternative for the currently used method. As a result, the reliability may have been underestimated in the current study.

Considering the statistical analyses, there are two primary limitations. First, the ratings, for both the inter-and intrarater reliability, were clustered. For the interrater reliability, each participant rated four vignettes, and for the intrarater reliability, each participant rated four cases twice. However, it was impossible to conduct multilevel analyses, as our data consisted of nominal and ordinal variables (intraclass correlations can only be calculated for continuous variables). Second, this study contained data missing by design for which it is not possible to correct (yet).

Another limitation is that the professionals only rated vignettes of their own agency. Therefore, it remains unknown how reliable the professionals rate vignettes of an other agency. For this reason, we were also not able to compare the ratings given by the professionals of both agencies. Since this is a vignette study, we were concerned about the ecological validity of our study. In an attempt to make the safety and risk assessments as realistic as possible in the sense that the assessments would closely resemble assessments in clinical practice, the professionals only assessed vignettes of their own company. If the professionals also rated vignettes of the other agency, much more ratings, and thus participants are needed. In planning this study, we assumed that it would not be possible to recruit sufficient participants for doing this. Therefore, we decided

to design the study with the best possible ecological validity.

Despite these limitations, this study has some important strengths. This is the first study that examined the intrarater reliability of a safety and risk assessment instrument. Additionally, multiple participant groups with varying experience levels rated vignettes from two different agencies. As a result, the reliability of the instrument was examined in a variety of circumstances. Lastly, this is the first study on the reliability of a safety and risk assessment instrument using Gwet's AC, and therefore avoiding the prevalence problem.

The current findings on the reliability of the ARIJ combined with findings of future studies on the ARIJ validity should give further practical guidance on how the safety and risk instruments can be improved. For example, by adjusting or excluding items that negatively affect the overall reliability and validity of the instrument. Based on their reliability, the following items need improvement: the safety assessment item related to parental availability (item 8), the risk assessment item related to a caregiver's history of abusing a child (item 17) and a caregiver's perception of the child as a problem (item 35).

In terms of reliability, the ARIJ safety assessment instrument compares favorably to other child safety assessment instruments, as a higher interrater reliability was found for both the items (Orsi et al., 2014; Bartelink et al., 2017) and the safety outcome of the ARIJ (Bartelink et al., 2017). This justifies the usage of the ARIJ safety assessment in practice. However, it is important to keep in mind that the previous studies seem to be negatively influenced by the prevalence problem, since they used Cohen's kappa and Krippendorff's alpha.

The ARIJ risk outcome proved to be reliable in a variety of circumstances, which justifies its use in practice and holds promises for the future of risk assessment in child welfare. On the other hand, as the items displayed a mixed reliability, it is important to be cautious with the use of ratings on risk factors in practice. When using the ARIJ risk assessment instrument in practice, the focus should be on the risk outcome. This is important as interventions should be in line with the risk outcome, which is prescribed by the risk-need-responsivity model (Bonta & Andrews, 2016).

Interestingly, the intrarater reliability of both instruments was higher than its interrater reliability. In other words, the instruments are rated more consistent within professionals than between professionals. This result may imply that professionals have their own consistent interpretation of the items and the instrument, but that these interpretations differ between professionals. Child welfare agencies should do more to increase the consistency in judgments of professionals in their agency. One way to do this, besides improving instruments, is by offering specialized training to professionals. After all, Sutherland et al. (2012) already found that specialist training improved the interrater reliability.

Acknowledgements We thank the professionals of Jeugdbescherming regio Amsterdam (Child Welfare Agency in Amsterdam) and Spirit (Child and Family Support Agency) for their valuable contributions, and in particular Inge Busschers and Carolien Konijn. We also thank all master's students that participated.

Authors' Contributions A.V. participated in the design of the study, collected the data, conducted all statistical analyses, and drafted the manuscript. M.A. and C.V.D.P. participated in the design of the study and critically reviewed the manuscript. GJS critically reviewed the manuscript. All authors contributed to and approved the final version of the manuscript.

Funding This work was funded by the Dutch organization for Health research and Development (ZonMW), grant number: 729300108. The funding organization was not involved in the data collection, data analysis, interpretation, and writing of the manuscript.

Compliance with Ethical Standards

Conflict of Interest A.V. declares that she has no conflict of interest. M.A., G.J.S. and C.v.d.P were involved in the development of the ARIJ.

Ethics Approval and Consent to Participate This study was conducted with approval of the Faculty Ethics Review Board (FMG–UvA) of the University of Amsterdam, the Netherlands. Implied consent procedures were considered appropriate by this Board. Participants were informed on the study procedure and voluntarily started the online questionnaire. No sensitive data were collected.

References

Ait Lbacha, H., Alali, S., Zouagui, Z., El Mamoun, L., Rhalem, A., Petit, E., Haddad, N., Gandoin, C., Boulios, H.-J., & Maillard, R. (2017). High prevalence of Anaplasma spp. in small ruminants in Morocco. *Transboundary and Emerging Diseases*, *64*(1), 250–263.

Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk assessment in child protective services: consensus and actuarial model reliability. *Child Welfare*, *78*(6), 723–748.

Barber, J. G., Shlonsky, A., Black, T., Goodman, D., & Trocmé, N. (2008). Reliability and predictive validity of a consensus-based risk assessment tool. *Journal of Public Child Welfare*, *2*(2), 173–195.

Barlow, J., Fisher, J. D., & Jones, D. (2012). Systematic review of models of analyzing significant harm. Oxford University. https://dera.ioe.ac.uk/14070/1/DFE-RR199.pdf

Bartelink, C., De Kwaadsteniet, L., Ten Berge, I. J., & Witteman, C. L. M. (2017). Is it safe? Reliability and validity of structured versus

unstructured child safety judgments. *Child & Youth Care Forum*, *46*(5), 745–768.
Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct*. New York, NY: Routledge.
Cash, S. J. (2001). Risk assessment in child welfare: the art and science. *Children and Youth Services Review*, *23*(11), 811–830.
Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558.
D'andrade, A., Austin, M. J., & Benton, A. (2008). Risk and safety assessment in child welfare: instrument comparisons. *Journal of Evidence-Based Social Work*, *5*(1–2), 31–56.
De Vogel, V., & Ruiter, Cde (2006). Structured professional judgment of violence risk in forensic clinical practice: a prospective study into the predictive validity of the Dutch HCR-2. *Psychology, Crime & Law*, *12*(3), 321–336.
Feng, G. C. (2014). Intercoder reliability indices: disuse, misuse, and abuse. *Quality & Quantity*, *48*(3), 1803–1815.
Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*, *11*(1), 13–22.
Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2015). *R-package 'irr' (version 0.84). Various coefficients of interrater reliability and agreement.* Retrieved from https://cran.r-project.org/web/packages/irr/irr.pdf.
Gwet, K. L. (2002). Interrater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Interrater Reliability Assessment Series*, *2*, 1–9.
Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.
Gwet, K. L. (2014). Intrarater reliability. In R. B. D'Agostino, L. Sullivan & J. Massaro (eds.), *Wiley Encyclopedia of Clinical Trials*. New Jersey: John Wiley & Sons.
Gwet, K. L. (2016). Paired t-test for agreement coefficients.r (r function). http://www.agreestat.com/r_functions.html.
Gwet, K. L. (2017). Agree.coeff3. raw.r (r function). http://www.agreestat.com/r_functions.html.
Hallgren, K. A. (2012). Computing interrater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.
Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.
Knoke, D., & Trocme, N. (2005). Reviewing the evidence on assessing risk for child abuse and neglect. *Brief Treatment and Crisis Intervention*, *5*(3), 310.
Ko, M. M., Park, T. Y., Lee, J. A., Choi, T. Y., Kang, B. K., & Lee, M. S. (2013). Interobserver reliability of pulse diagnosis using traditional Korean medicine for stroke patients. *The Journal of Alternative and Complementary Medicine*, *19*(1), 29–34.
Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.
Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies*, *48*(6), 661–671.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology*, *49*(4), 431–434.
McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.
Orsi, R., Drury, I. J., & Mackert, M. J. (2014). Reliable and valid: a procedure for establishing item-level interrater reliability for child maltreatment risk and safety assessments. *Children and Youth Services Review*, *43*, 58–66.
Penney, S. R., McMaster, R., & Wilkie, T. (2014). Multirater reliability of the historical, clinical, and risk management-2. *Assessment*, *21*(1), 15–27.
Proutskova, P., & Gruszczynski, M. (2017). *An r package for performing bootstrap replicates of krippendorff's alpha on intercoder reliability data*. R-package 'Kripp.boot'. https://github.com/MikeGruz/kripp.boot.
Quesada, S. P., Calkins, C., & Jeglic, E. L. (2014). An examination of the interrater reliability between practitioners and researchers on the Static-99. *International Journal of Offender Therapy and Comparative Criminology*, *58*(11), 1364–1375.
Sutherland, A. A., Johnstone, L., Davidson, K. M., Hart, S. D., Cooke, D. J., Kropp, P. R., & Stocks, R. (2012). Sexual violence risk assessment: an investigation of the interrater reliability of professional judgments made using the risk for sexual violence protocol. *International Journal of Forensic Mental Health*, *11*(2), 119–133.
Ten Berge, I. J. (2008). *Instrumenten voor risicotaxatie in situaties van (vermoedelijke) kindermishandeling*. Utrecht, Netherlands: Jeugdinstituut.
Teo, A. R., Holley, S. R., Leary, M., & McNiel, D. E. (2012). The relationship between level of training and accuracy of violence risk assessment. *Psychiatric Services*, *63*(11), 1089–1094.
Van der Put, C. E., Assink, M., & Stams, G. J. J. M. (2016). Predicting relapse of problematic child-rearing situations. *Children and Youth Services Review*, *61*, 288–295.
Webster, S. D., Mann, R. E., Carter, A. J., Long, J., Milner, R. J., O'Brien, M. D., Wakeling, H. C., & Ray, N. L. (2006). Interrater reliability of dynamic risk assessment with sexual offenders. *Psychology, Crime & Law*, *12*(4), 439–452.
Wongpakaran, N., Wongpakaran, T., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating interrater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 61.
Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring interrater reliability for nominal data—which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, *16*, 93.
Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). Suppl-1, M5: high agreement and high prevalence: the paradox of Cohen's Kappa. *The Open Nursing Journal*, *11*, 211–218.
Zhao, X., Feng, G. C., Liu, J. S., & Deng, K. (2018). We agreed to measure agreement-redefining reliability de-justifies Krippendorff's Alpha. *China Media Research*, *14*(2), 1–15.