Statement validity assessment: Inter-rater reliability of criteria-based content analysis in the mock-crime paradigm

Heinz Werner Gödert*, Matthias Gamer, Hans-Georg Rill and Gerhard Vossel
Johannes Gutenberg-University, Mainz, Germany

Purpose. The inter-rater reliability of criteria-based content analysis (CBCA) – the main component of statement validity assessment (SVA) – was investigated in a mock-crime study. The study also addressed the issue of the adequacy of diverse statistical indices of reliability. Furthermore, CBCA’s effectiveness in distinguishing between true and false statements was analysed.

Methods. Three raters were trained in CBCA. Subsequently, they analysed transcripts of 102 statements referring to a simulated theft of money. Some of the statements were based on experience and some were confabulated. The raters used 4-point scales, respectively, to judge the degree to which 18 of the 19 CBCA criteria were fulfilled in each statement.

Results. The analysis of rater judgment distributions revealed that, with judgments of individual raters varying only slightly across transcripts, the weighted kappa coefficient, the product-moment correlation, and the intra-class correlation were inadequate indices of reliability. The Finn-coefficient and percentage agreement, which were calculated as indices independent of rater judgment distributions, were sufficiently high with respect to 17 of the 18 assessed criteria. CBCA differentiated significantly between truthful and fabricated accounts.

Conclusions. The inter-rater reliability of CBCA achieved in the present study was satisfactory both, if considered absolutely, and as compared with other empirical findings. This suggests that CBCA can be utilized in the mock-crime paradigm with a sufficient degree of reliability.

At present, statement validity assessment (SVA) is the most common technique for measuring the credibility of verbal statements (Vrij, 2000). SVA was developed by German psychologists (see Undeutsch, 1984, 1989) originally in order to determine the credibility of child witnesses’ testimonies in trials for sexual abuse. However, SVA is, in

* Correspondence should be addressed to Dr. Heinz Werner Gödert, Department of Psychology, Johannes Gutenberg-University Mainz, Staudingerweg 9, D-55099 Mainz, Germany (e-mail: goedert@uni-mainz.de).

DOI:10.1348/13552505SX52680
principle, also capable of measuring the credibility of adults’ statements and of statements relating to topics other than sexual offenses (Arntzen, 1993; Porter & Yuille, 1995; Steller & Köhnken, 1989; Yuille & Cutshall, 1989). Accordingly, recent research has addressed the usefulness of SVA for the assessment of statements given by adults and defendants in cases of non-sexual crimes (e.g. Porter & Yuille, 1996).

SVA consists of three major elements (cf. Raskin & Esplin, 1991): (1) a structured interview; (2) the criteria-based content analysis (CBCA), which systematically assesses the content quality of the statement; (3) and the Validity Checklist, which relates the CBCA outcome to other evidence and to factors associated with the interview. The core component of SVA is the assessment of the statement’s content quality by means of CBCA. This diagnostic tool, which was developed by Steller and Köhnken (1989), comprises 19 criteria (cf. Table 1 plus Criterion 19: details characteristic of the offence) that are assumed to reflect a statement’s content quality (in terms of vividness, concreteness, originality, psychological coherence etc.; cf. Undeutsch, 1967). The more numerous and/or the more explicitly these criteria are met by a statement, the more it is probable that this statement is credible; that is, rests on real experience.

The validity of CBCA with respect to the differentiation of truthful and fabricated accounts has already been addressed by several field and experimental studies (e.g. Akehurst, Köhnken, & Höfer, 2001; Lamb et al., 1997; Parker & Brown, 2000; Vrij, Akehurst, Soukara, & Bull, 2002; for an overview see Ruby & Brigham, 1997; Vrij, 2000; Vrij & Akehurst, 1998). In contrast, relatively little research has been done on the issue of inter-rater reliability of CBCA, which is astonishing since sufficient reliability is an essential precondition for good validity. Only two studies have been published that focused exclusively and in detail on inter-rater reliability of CBCA (Anson, Golding, & Gully, 1993; Horowitz et al., 1997). In some other studies inter-rater agreement was routinely computed mainly to justify the use of mean ratings (averaged across raters) for subsequent validity analyses (Akehurst et al., 2001; Porter & Yuille, 1996; Sporer, 1997; Vrij et al., 2002; Vrij, Edward, & Bull, 2001a, 2001b; Vrij, Edward, Roberts, & Bull, 2000; Vrij, Kneller, & Mann, 2000; Zaparniuk, Yuille, & Taylor, 1995).

The two extensive reliability studies by Anson et al. (1993) and Horowitz et al. (1997) were field studies; that is, the statements analysed were children’s testimonies given in real forensic cases (alleged sexual abuse). Overall percentage inter-rater agreement was 75% and 82%, respectively but in both studies inter-rater agreement varied substantially between individual CBCA criteria.

However, with regard to the systematic investigation and further development of CBCA, experimental paradigms, especially the mock-crime paradigm, may prove useful, too. In the mock-crime paradigm, participants are involved in a simulated crime as either perpetrators or witnesses. Subsequently, they are subjected to an interrogation bearing on the mock-crime. Whereas the research on psychophysiological detection of deception (cf. Ben-Shakhar & Furedy, 1990; Lykken, 1998) has made extensive use of the mock-crime paradigm (cf. Kleiner, 2002; MacLaren, 2001), only three validation studies utilizing the mock-crime paradigm have been published in research on content-oriented SVA (Porter & Yuille, 1996; Tye, Amato, Honts, Devitt, & Peters, 1999; Vrij et al., 2002).

1 Whereas the mock-crimes in the studies by Porter and Yuille (1996) and Tye et al. (1999) involved simulated thefts of money and of a textbook, respectively, the statements in the Vrij et al. (2002) study referred to a ‘rubbing the blackboard’ event; that is, the simulated ‘crime’ consisted in wiping important information off a blackboard without being authorized to do so.
Compared with field research, the mock-crime paradigm has the disadvantage of limited external validity, but the advantage of higher internal validity. In particular, the validation criterion (real truth status of the relevant account/guilt status of the participant) can be controlled satisfactorily only in experimental studies. Inter-rater agreement was computed in two mock-crime studies. Porter and Yuille (1996) reported high percentage agreement (> 80%) on the basis of eight randomly selected interviews, while Vrij and his colleagues (2002) reported correlation coefficients between two raters and found acceptable consistency for most CBCA criteria, with substantial variability, however.

Another experimental approach that provides perfect control of the validation criterion as well is the film paradigm. Here, the relevant accounts refer to the content of a film. Prior to giving the account, participants in the truthful condition have actually watched the film, whereas participants in the untruthful condition have only been roughly informed about the film content (e.g. Akehurst et al., 2001; Köhnken, Schimossek, Aschermann, & Höfer, 1995; Vrij, Kneller & Mann, 2000). However, this paradigm has the major disadvantage that the independent variable ‘watching vs. not watching a film’ is not an adequate variation of the factor ‘experiencing vs. not experiencing a personally significant event’ (cf. Steller, Wellershaus, & Wolf, 1992). Thus, the external validity has to be considered lower for the film paradigm than for the mock-crime paradigm in which participants are personally involved in the relevant event. Inter-rater reliability for these two paradigms was found to be largely similar: overall observer agreement ranged from 74% to 92% in both paradigms and correlations between raters showed acceptable consistency for most CBCA criteria (Vrij et al., 2001a, 2001b; Vrij, Edward, et al., 2000; Zaparniuk et al., 1995). Only a study by Akehurst and colleagues (2001) reported a somewhat lower agreement between raters. Across studies, the lowest reliability in terms of inter-rater agreement was found for the general characteristics of the statement, especially for Criteria 1 (logical structure) and 2 (unstructured production).

A better solution for the problem of external validity is provided by a third experimental approach, the autobiographical paradigm. Here, the accounts in the truthful condition refer to autobiographical experiences ideally characterized by personal involvement, loss of control, and a negative emotional tone; that is, characteristics that are also typical for the situation of crime witnesses (victims). Examples of such autobiographical experiences are: being bitten by a dog (Steller et al., 1992; see also Steller, Wellershaus, & Wolf, 1988), bearing a child (Wolf & Steller, 1997), or being dumped by a girlfriend (Ruby & Brigham, 1998). A study by Sporer (1997), utilizing the autobiographical paradigm, found comparably low agreement (overall: 60%) and consistency between two raters, especially for criteria with very low or very high base rates.

Compared with field research and with the autobiographical paradigm, the mock-crime paradigm has the major advantage that the relevant event addressed by the statements can be manipulated systematically. Thus, the mock-crime paradigm may contribute to answering the question of which factors of the relevant event affect the validity of CBCA. However, the application of the mock-crime paradigm depends on

---

2 In the study by Akehurst et al. (2001) there were two truthful conditions. In one of these truthful conditions, subjects gave an account of a photography session they had watched on a video-recording. In the other truthful condition, the statements referred to a photography session the subjects had actually taken part in.
the premise that the CBCA criteria can be assessed reliably in accounts that refer to simulated crimes. Accordingly, the aim of the present study was to examine the inter-rater reliability and validity of CBCA in a mock-crime context. In addition, the issue of the adequacy of diverse statistical reliability indices was addressed.

Methods

Mock-crime

\( N = 102 \) women aged 17–67 years (\( M = 26, SD = 9, Md = 23 \)) participated in a mock-crime experiment either as perpetrators, as witnesses, or as uninvolved subjects (34 participants in each group).

Participants in the perpetrators group were instructed to commit a mock-crime. They were required to steal money from Professor Kunze’s office (Professor Kunze was a fictive person). The money was deposited in a hidden box, locked by a 10-digit combination lock. Since Professor Kunze was not very good at retaining digits in memory, he had written down the combination-digits on 10 slips of paper, each slip containing one digit, and hidden these slips of paper at different places of his office. The slips of paper did not only contain information about the digits but also about the place where the slip of paper with the next digit could be found. For example, on the slip of paper containing the first digit Professor Kunze also had noted the place, where he had hidden the slip of paper with the second digit of the digit combination, and so forth. The 10 slips of paper were located (1) in a desk drawer with a Germany sticker on it, (2) under a cactus, (3) under a porcelain dog, (4) in the saddlebag of a yellow bicycle parked in the office, (5) behind a picture of cows, (6) under a box containing water bottles, (7) in the pocket of a leather jacket, (8) under a bowl containing apples, (9) behind a darts target, (10) under a red carpet. The slip of paper with the last (10th) digit also contained information about the place where the money box was hidden. Incidentally, the participants in the perpetrators group had come to know the place where Professor Kunze had hidden the slip of paper containing the first digit of the digit combination. Thus, the perpetrators just had to go to Professor Kunze’s office, sequentially trace the 10 slips of paper, write down the 10 digits for the digit combination lock, open the lock, and take all the money deposited inside the box (a 100 DM note, about EUR 51).

When the perpetrators committed the theft another woman was also present in Professor Kunze’s office. Participants in the perpetrators group were told in advance that there would be another person in the office who allegedly took part in the experiment as a potential witness of the mock-crime. Participants in the perpetrators group were instructed to try to mask the theft and to give an impression as innocent as possible. Actually, the alleged witness was a confederate of the experimenter who pretended to be a naive participant in the witnesses group. For this purpose, the confederate performed the same cleaning tasks as the participants in the witnesses group (see below).

Participants in the witnesses group were instructed to clean up Professor Kunze’s office and, thereby, become witnesses of the money theft. They were told to go to Professor Kunze’s office where they would find a list of 10 cleaning tasks Professor Kunze asked them to dispatch. They were to (1) tighten a loose screw of the desk drawer with the Germany sticker on it, (2) water the cactus, (3) dust off the picture of cows, (4) wipe the porcelain dog, (5) take down the darts from the darts target, (6) put the leather jacket on the hanger, (7) clean the red carpet with a vacuum cleaner, (8) wipe the bowl containing apples, (9) dust behind the darts target, (10) water the water bottles.
cleaner, (8) put the yellow bicycle on the kickstand, (9) put the water bottles back into the box, (10) put the apples back into the bowl. Furthermore, the participants in the witnesses group were instructed that another person would enter the office and commit a theft while they had to carry out the cleaning tasks. The thief was alleged to be a genuine experimental participant that had been assigned to the perpetrators group. Actually, she was a confederate of the experimenter who committed the theft in the same manner as did the actual naive participants in the perpetrators condition. Thus, the participants in the witnesses group observed the theft and, thereby, became aware of the crime details. Furthermore, some of the crime details also were integral parts of the witnesses' cleaning job.

After committing or witnessing the mock-theft, the perpetrators and witnesses were introduced to a female investigator, who was blind with regard to their experimental condition. They were told that they were suspected of having committed the theft. In order to exculpate themselves from the accusation, participants were asked to give a testimony in which they blamed another person for having committed the theft. They were promised a reward of DM 15 (about EUR 8) if their testimony was considered credible by the investigator. Thus, the witnesses could make a truthful account, since they had personally observed the theft. The perpetrators had to invent a story in which they alleged that they had observed the true witness committing the theft.

Participants in the uninvolved subjects group neither committed nor witnessed the mock-theft. They were only roughly informed about the main aspects of the theft. They were told that on the same day DM 100 had been stolen from Professor Kunze’s office. They were further informed that it was known that during the theft, not only the thief, but also another person had been present in Professor Kunze’s office and that this second person had possibly witnessed the theft and possibly even had communicated with the thief. Furthermore, they were told that a reward of DM 15 was offered for a credible testimony concerning the theft. They were instructed to try to get the reward by pretending to be the person who had witnessed the theft. They were told in advance, that they could get the reward only if their testimonial account was considered credible by a female investigator, who was blind with respect to their experimental condition. Accordingly, they were instructed to give an impression as credible as possible in order to convince the investigator.

Prior to being introduced to the female investigator participants in each experimental group were given 15 minutes to prepare their testimonial accounts.

**Interview**

The testimonies were each given in a semi-standardized interview. At the beginning of the interview participants were asked to give a free narrative about the relevant event as detailed and thorough as possible. In this part of the interview further questions were asked by the investigator only if the participants’ accounts contained ambiguities or if it was difficult for the participants to give a free account. Subsequently, the investigator asked open-ended questions concerning a precise description of the thief and of the scene of crime. When the participants had answered these questions they were explicitly asked and given time to reconsider the completeness and accuracy of their statements. If the participants had nothing to add or to revise, the interview was finished.

The overall mean number of words per statement was 686 (SD = 314). For the perpetrators, witnesses, and uninvolved subjects, the average number of words was 694 (SD = 349), 825 (SD = 305), and 540 (SD = 215), respectively.
Scoring

The interviews were recorded and transcribed. Subsequently, the transcripts were scored by three independent raters (advanced female students of psychology) who had received training in CBCA (see below). The raters had not been informed about the details of the mock-crime or the truth status of the respective accounts or the base rates of truthful and confabulated accounts. Each of the three raters assessed, for each of the 102 accounts, the degrees of presence of the first 18 CBCA criteria (cf. Table 1). Ratings for each criterion were done on 4-point scales \(0 = \text{not present}, 1 = \text{slightly present}, 2 = \text{moderately present}, 3 = \text{strongly present}\). Criterion 19 (details characteristic of the offence) was not considered because it exclusively refers to sexual offences. The use of 4-point scales was adopted from Steller et al. (1992; see also Steller et al., 1988) and was best compatible with the scoring rules implemented in the rater training program (see below), which was co-developed by Köhnken (see Höfer, Krause, Petersen, Sievers, & Köhnken, 1999). In order to control for potential order effects, the sequence of the transcripts to be scored was varied from rater to rater. For rater A the order of the transcripts was 1, 2, 3, \ldots, 102. For rater B the order was 102, 101, 100, \ldots, 1. For rater C it was 51, 52, 50, 53, 49, 54, \ldots, 1, 102.

Rater training

Raters were trained according to a special CBCA training program developed by Krause (1997) and Petersen (1997; see also Höfer et al., 1999). For organizational reasons, the training was not as extensive as the original timetable. It consisted of seven sessions, each session taking between 2 and 3 hours. The training sessions are described in more detail in the Appendix. The transcripts utilized for training purposes were taken from Krause (1997), Petersen (1997), and Scheinberger (1993). Prior to the first training session, the trainees prepared themselves for the training by studying scientific CBCA literature in the order listed as follows: (1) Köhnken (1990, pp. 82–117); (2) Steller and Köhnken (1989); (3) Steller et al. (1992); (4) Wolf and Steller (1997); (5) Krahé and Kundrotas (1992); (6) Littmann and Szewczyk (1983); (7) Wellershaus (1992).

Results

For each of the CBCA criteria, the percentage agreement between the three raters, the weighted kappa coefficient (Cohen, 1968), the product-moment correlation (Crocker & Algina, 1986), the intra-class correlation (Bartko, 1966; Shrout & Fleiss, 1979), and the Finn-coefficient (Finn, 1970) were each calculated as indices of inter-rater reliability. All coefficients were calculated using the extension package \texttt{irr}, version 0.5 (Gamer, 2005) for the freely available statistical software \texttt{R}, version 2.0.0 (R Development Core Team, 2004).

Simple and extended percentage agreement

For each CBCA criterion, the relative frequency of transcripts for which all three raters had chosen the same degree on the 4-point scale was calculated (simple percentage agreement). Results are depicted in Table 1. Percentage agreement of 70% or more is considered as indicating high consistency among raters (Wellershaus & Wolf, 1989).
According to this, only six criteria (1, 7, 10, 11, 16, 17) achieved high inter-rater reliability. Another criterion (15) fell only slightly short of the 70% threshold. Nevertheless, with a 4-point scale it is acceptable if the ratings of different raters differ by only one scale point. Accordingly, for each criterion the extended percentage agreement was calculated; that is, it was considered as an agreement if the three raters chose two adjacent scale degrees (e.g. rater A chose scale degree 1; rater B and rater C chose scale degree 2). Fifteen criteria achieved a high extended percentage agreement (≥70%; see Table 1). The Criteria 3 and 14, respectively, fell only barely short of the 70% threshold. Only for Criterion 9 was the extended agreement considerably lower than 70%.

A major shortcoming of both the simple and extended percentage agreement is that the probability of chance agreement between raters is neglected (Frick & Semmel, 1978). For this reason, alternative indices of inter-rater reliability were calculated.

Weighted kappa coefficients, product-moment correlations, and intra-class correlations
For each CBCA criterion and each possible two-party combination of raters (A–B, A–C, B–C), respectively, the weighted kappa coefficient (κw) was calculated (see Table 2). Kappa reflects the chance-corrected proportion of inter-rater agreements. Furthermore,
with the weighted kappa coefficient, disagreements between raters are weighted depending on the discrepancy between the scale degrees chosen by the respective raters. By this method, the degree of disagreement between raters is taken into account. Since the 4-point scale represented a unidimensional continuum with constant distances between adjacent scale degrees (interval scale), the differences between chosen scale degrees were weighted quadratically (Cohen, 1968). In order to obtain one index of the overall inter-rater agreement for each CBCA criterion, the three pairwise kappa coefficients were averaged (Conger, 1980). The mean weighted kappa coefficients are depicted in Table 1.

Table 2. Pairwise weighted kappa coefficients ($\kappa_W$) and product-moment correlation coefficients ($r$) for the three raters (A, B, C)

<table>
<thead>
<tr>
<th>CBCA criterion</th>
<th>A–B</th>
<th>A–C</th>
<th>B–C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Logical structure</td>
<td>$\kappa_W$</td>
<td>$r$</td>
<td>$\kappa_W$</td>
</tr>
<tr>
<td>3. Quantity of details</td>
<td>.60</td>
<td>.703</td>
<td>.31</td>
</tr>
<tr>
<td>4. Contextual embedding</td>
<td>.39</td>
<td>.433</td>
<td>.63</td>
</tr>
<tr>
<td>5. Descript. of interact.</td>
<td>.12</td>
<td>.152</td>
<td>.36</td>
</tr>
<tr>
<td>6. Reprod. of convers.</td>
<td>.77</td>
<td>.818</td>
<td>.66</td>
</tr>
<tr>
<td>7. Unexp. complications</td>
<td>−.07</td>
<td>−.077</td>
<td>.04</td>
</tr>
<tr>
<td>8. Unusual details</td>
<td>−.01</td>
<td>−.026</td>
<td>.09</td>
</tr>
<tr>
<td>9. Superfluous details</td>
<td>.14</td>
<td>.209</td>
<td>.08</td>
</tr>
<tr>
<td>10. Details misunderstood</td>
<td>−.02</td>
<td>−.020</td>
<td>−.01</td>
</tr>
<tr>
<td>11. Rel. ext. associations</td>
<td>a</td>
<td>b</td>
<td>.00</td>
</tr>
<tr>
<td>12. Subj. mental state</td>
<td>.46</td>
<td>.600</td>
<td>.59</td>
</tr>
<tr>
<td>13. Perp. mental state</td>
<td>.52</td>
<td>.591</td>
<td>.53</td>
</tr>
<tr>
<td>15. Lack of memory</td>
<td>.09</td>
<td>.127</td>
<td>.16</td>
</tr>
<tr>
<td>16. Self-doubts</td>
<td>−.01</td>
<td>−.010</td>
<td>−.02</td>
</tr>
<tr>
<td>17. Self-deprecation</td>
<td>.00</td>
<td>b</td>
<td>−.01</td>
</tr>
<tr>
<td>18. Pard. the perpetrator</td>
<td>−.04</td>
<td>−.040</td>
<td>.52</td>
</tr>
</tbody>
</table>

Note: (a) $\kappa_w$ could not be calculated since both raters consistently chose scale degree 0. (b) could not be calculated since at least one rater consistently chose the same scale degree.

For each CBCA criterion and each possible two-party combination of raters (A–B, A–C, B–C), respectively, the product-moment correlation ($r$) was calculated (see Table 2). As an index of the overall inter-rater agreement, the three pairwise product-moment correlation coefficients were averaged for each CBCA criterion. Fisher’s Z-transformation was utilized for the averaging. The average product-moment correlations are depicted in Table 1.

As an alternative to averaging the three pairwise kappa coefficients in order to obtain an index for the overall agreement between the three raters, it is also possible to calculate one single kappa coefficient across all three raters. Nevertheless, the calculation of this index originally proposed by Cohen (1972) and referred to as weighted chi square is very complex. So, in view of the rating distribution problems described below, the calculation of weighted chi square values was abstained from.
For each CBCA criterion an ANOVA with the factors ‘transcript’ (1, 2, ..., 102) and ‘rater’ (A, B, C) was conducted (ANOVA model of chance effects, one observation per cell). Thus, the proportion of the overall variability that was due to the factor ‘transcript’ (i.e. not due to the factor ‘rater’ or to a ‘transcript’ × ‘rater’-interaction) could be calculated for each CBCA criterion. These proportions (generally referred to as intra-class correlations \(r_{AA}\)) are depicted in Table 1.

Weighted kappa coefficients from .40 to .74 are considered to indicate fair to good inter-rater agreement beyond chance; coefficients of .75 or more are taken to represent excellent agreement beyond chance (Fleiss, 1981). These specifications also apply to the product-moment correlation and to the intra-class correlation, because, given ideal conditions (equal marginal distributions in the contingency table, sufficient number of cases [transcripts]), the quadratically weighted kappa is equivalent to the product-moment correlation and intra-class correlation (Cohen, 1968; Fleiss & Cohen, 1973).

Table 1 shows, that fair to good weighted kappa coefficients, product-moment correlations, and intra-class correlations were obtained for only five CBCA criteria (3, 4, 6, 12, 13). However, the weighted kappa coefficient as well as the product-moment correlation and intra-class correlation do not exclusively reflect actual inter-rater agreement, but also depend on the distributions of the ratings. If the ratings vary only slightly between transcripts (which, for instance, should be the case when the transcripts actually do not differ with respect to the relevant CBCA criterion), the resultant kappa coefficients, product-moment correlations and intra-class correlations tend to underestimate the true inter-rater agreement (Kozlowski & Hattrup, 1992; McGraw & Wong, 1996). This problem was also immanent in the data of the present study. In order to illustrate this, the major parameters for the distributions of the ratings are depicted in Table 3. As can be seen in Table 3, the six CBCA criteria with the lowest kappa coefficients, product-moment correlations, and intra-class correlations (Criteria 1, 7, 10, 11, 16, 17; \(k_{AA}, r, r_{AA} \leq .05\); cf. Table 1\(^5\)) were the same criteria for which the standard deviations of the ratings were very small and for which the rating distributions of all three raters, respectively, were extremely right-skewed (rater chooses scale-degree 3 for most of the transcripts) or left-skewed (rater chooses scale-degree 0 for most of the transcripts). The low \(k_{AA}, r,\) and \(r_{AA}\) values of these six criteria are intelligibly inconsistent with the fact that these criteria achieved very high values of extended percentage agreement and even the highest values of simple percentage agreement (see Table 1). This suggests that, given low variability and extreme skewness of the ratings, neither the weighted kappa coefficient nor the product-moment correlation nor, finally, the intra-class correlation could be regarded as adequate indices of inter-rater reliability.

**Finn-coefficients**

For cases of very low variability between transcripts the Finn-coefficient has been proposed as an alternative to the intra-class correlation (Finn, 1970). The Finn-coefficient represents the proportion of observed to expected variability substracted from 1:

\[\text{Finn-coefficient} = 1 - \frac{\text{expected variability}}{\text{observed variability}}\]

\(^5\) With respect to CBCA criterion 11, neither the three pairwise product-moment correlations nor their average could be calculated, since the ratings of raters A and B did not show any variability at all. Accordingly, with Criterion 11, pairwise kappa coefficients could be calculated only for the rater combinations A–C and B–C (.00, respectively) but not for A–B, so that averaging all three pairwise kappa coefficients was not possible (see Table 2). With respect to Criterion 17, there was no variability in the ratings of rater B, so that calculating the pairwise product-moment correlation was possible only for the rater combination A–C (\(r = -.01\)). Accordingly, calculating the average of all three pairwise product-moment correlations was not possible (see Table 2).
Finn-coefficient $= 1.0 - \text{observed variance/expected variance}.$

The rationale of the Finn-coefficient is as follows: perfect agreement between raters exists if there is zero variance between their ratings of a given transcript. Zero inter-rater agreement, on the other hand, exists if the ratings of different raters simply reflect random placement of a given transcript into the available scale degrees; that is, the distribution of ratings across scale degrees is rectangular. Accordingly, the variance of rectangularly distributed ratings is conceptualized as the expected variance. The degree to which the observed variance is less than the expected variance then reflects the presence of non-error variance in the ratings. Thus, the ratio of observed variance to expected variance gives the proportion of random or error variance present in the observed ratings. Subtracting this ratio from 1.0 then gives the proportion of non-error variance in the ratings, which could be interpreted as a reliability coefficient.

Table 1 shows the Finn-coefficients for the respective CBCA criteria. For 14 criteria (1, 2, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18), excellent Finn-coefficients ($\geq .75$) were obtained; all other criteria (3, 4, 9, 14) achieved fair to good Finn-coefficients (from .40 to .74). Compared with the other criteria, however, the coefficient for Criterion 9 was relatively low.

**Table 3. Distribution parameters of the ratings, separated for CBCA criteria and raters**

<table>
<thead>
<tr>
<th>CBCA criterion</th>
<th>Rater A</th>
<th></th>
<th>Rater B</th>
<th></th>
<th>Rater C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>S (SD)</td>
<td>M (SD)</td>
<td>S (SD)</td>
<td>M (SD)</td>
<td>S (SD)</td>
</tr>
<tr>
<td>1. Logical structure</td>
<td>2.97 (0.17)</td>
<td>−5.65</td>
<td>2.93 (0.32)</td>
<td>−5.05</td>
<td>2.90 (0.33)</td>
<td>−3.55</td>
</tr>
<tr>
<td>2. Unstructured prod.</td>
<td>0.36 (0.58)</td>
<td>1.35</td>
<td>0.32 (0.57)</td>
<td>1.91</td>
<td>0.52 (0.79)</td>
<td>1.58</td>
</tr>
<tr>
<td>3. Quantity of details</td>
<td>1.33 (1.07)</td>
<td>0.18</td>
<td>1.86 (0.82)</td>
<td>0.26</td>
<td>2.21 (0.80)</td>
<td>−0.51</td>
</tr>
<tr>
<td>4. Contextual embedding</td>
<td>1.01 (0.81)</td>
<td>0.43</td>
<td>1.24 (0.58)</td>
<td>0.23</td>
<td>1.19 (0.98)</td>
<td>0.13</td>
</tr>
<tr>
<td>5. Descrip. of interact.</td>
<td>0.41 (0.65)</td>
<td>1.77</td>
<td>0.15 (0.38)</td>
<td>2.55</td>
<td>0.38 (0.70)</td>
<td>1.73</td>
</tr>
<tr>
<td>6. Reprod. of convers.</td>
<td>1.77 (1.04)</td>
<td>−0.33</td>
<td>1.50 (0.86)</td>
<td>0.14</td>
<td>1.88 (0.96)</td>
<td>−0.31</td>
</tr>
<tr>
<td>7. Unexp. complications</td>
<td>0.13 (0.41)</td>
<td>3.43</td>
<td>0.06 (0.24)</td>
<td>3.81</td>
<td>0.04 (0.20)</td>
<td>4.82</td>
</tr>
<tr>
<td>8. Unusual details</td>
<td>0.10 (0.30)</td>
<td>2.74</td>
<td>0.64 (0.48)</td>
<td>−0.58</td>
<td>0.45 (0.56)</td>
<td>0.73</td>
</tr>
<tr>
<td>9. Superfluous details</td>
<td>0.24 (0.55)</td>
<td>2.65</td>
<td>0.80 (1.10)</td>
<td>1.04</td>
<td>1.13 (0.92)</td>
<td>0.13</td>
</tr>
<tr>
<td>10. Details misunderstood</td>
<td>0.02 (0.14)</td>
<td>7.03</td>
<td>0.02 (0.14)</td>
<td>7.03</td>
<td>0.02 (0.20)</td>
<td>10.10</td>
</tr>
<tr>
<td>11. Rel. ext. associations</td>
<td>0.00 (0.00)</td>
<td>a</td>
<td>0.00 (0.00)</td>
<td>a</td>
<td>0.03 (0.17)</td>
<td>5.65</td>
</tr>
<tr>
<td>12. Subj. mental state</td>
<td>1.31 (0.70)</td>
<td>0.87</td>
<td>0.78 (0.65)</td>
<td>0.69</td>
<td>1.50 (0.74)</td>
<td>−0.37</td>
</tr>
<tr>
<td>13. Perp. mental state</td>
<td>0.49 (0.56)</td>
<td>0.56</td>
<td>0.26 (0.44)</td>
<td>1.08</td>
<td>0.81 (0.89)</td>
<td>0.38</td>
</tr>
<tr>
<td>14. Spont. corrections</td>
<td>1.27 (0.75)</td>
<td>0.67</td>
<td>0.80 (0.65)</td>
<td>0.43</td>
<td>1.65 (0.91)</td>
<td>−0.37</td>
</tr>
<tr>
<td>15. Lack of memory</td>
<td>0.06 (0.24)</td>
<td>3.81</td>
<td>0.26 (0.47)</td>
<td>1.38</td>
<td>0.23 (0.49)</td>
<td>2.09</td>
</tr>
<tr>
<td>16. Self-doubts</td>
<td>0.01 (0.10)</td>
<td>10.10</td>
<td>0.01 (0.10)</td>
<td>10.10</td>
<td>0.06 (0.24)</td>
<td>3.81</td>
</tr>
<tr>
<td>17. Self-deprecation</td>
<td>0.02 (0.20)</td>
<td>10.10</td>
<td>0.00 (0.00)</td>
<td>a</td>
<td>0.02 (0.20)</td>
<td>10.10</td>
</tr>
<tr>
<td>18. Pard. the perpetrator</td>
<td>0.19 (0.44)</td>
<td>2.32</td>
<td>0.14 (0.35)</td>
<td>2.14</td>
<td>0.41 (0.69)</td>
<td>1.59</td>
</tr>
</tbody>
</table>

*Note: M = mean, SD = standard deviation, S = skewness. (a) S could not be calculated since there was no variability in the ratings. The rater consistently chose scale degree 0.*

**Differentiation of truthful and fabricated accounts**

For analyses concerning validity of CBCA, ratings were averaged across the three raters. Thus, validity analyses were calculated on the basis of one score (average across raters) per transcript and CBCA criterion. In a first step, these average scores were summed across the 18 CBCA criteria in order to obtain one total CBCA score for each participant. The total
CBCA scores were then analysed for group differences. A one-way ANOVA with the between-subjects factor experimental group (perpetrators [fabricated accounts] vs. uninvolved subjects [fabricated accounts]) yielded a significant effect, $F(2, 99) = 5.515, p < .01$. Scheffé tests for pairwise group comparisons showed that the total CBCA scores in the witnesses group ($M = 14.49, SD = 3.07$) were significantly higher than the total CBCA scores obtained by the perpetrators ($M = 12.43, SD = 3.85$) and by the uninvolved subjects ($M = 11.98, SD = 2.98$), $p < .05$, respectively. Perpetrators and uninvolved subjects did not differ significantly.

In a second step, a multivariate ANOVA (MANOVA) with the between-subjects factor experimental group (perpetrators vs. witnesses vs. uninvolved subjects) as the independent variable and the scores for each of the 18 CBCA criteria as the dependent variables was calculated. The MANOVA yielded a significant simultaneous effect of the group factor on the 18 scores for the respective CBCA criteria, $PS = 0.872, F(36, 166) = 3.565, p < .01$. In order to further clarify this overall effect, three MANOVAs were calculated, in which only two groups (perpetrators vs. witnesses, perpetrators vs. uninvolved subjects, witnesses vs. uninvolved subjects) were considered, respectively. Bonferroni correction was applied, thus: $a' = .05/3 = .017$ was used for the statistical tests. These MANOVAs for pairwise group comparisons showed that the witnesses differed significantly from the perpetrators, $PS = 0.545, F(18, 49) = 3.263, p < .017$, and from the uninvolved subjects, $PS = 0.761, F(18, 49) = 8.690, p < .017$. Perpetrators and uninvolved subjects did not differ significantly.

In a third step, each of the 18 CBCA criteria was analysed for group differences by a one-way ANOVA. The $a$ priori significance level was not adjusted for multiple testing because of the expected loss of statistical power due to the rather high number of single comparisons. Significant overall effects were clarified by Scheffé tests for pairwise group comparisons. The results of the 18 ANOVAs and corresponding Scheffé tests as well as the mean scores and standard deviations for each CBCA criterion and experimental group are depicted in Table 4. Significant overall effects were obtained for seven CBCA criteria (3. quantity of details, 4. contextual embedding, 6. reproduction of conversation, 8. unusual details, 9. superfluous details, 14. spontaneous corrections, 15. admitting lack of memory; cf. Table 4). The Scheffé tests showed that Criteria 3 (quantity of details), 6 (reproduction of conversation), and 9 (superfluous details) were significantly more pronounced for the witnesses than for the perpetrators and uninvolved subjects (cf. Table 4). Criterion 14 (spontaneous corrections) and 15 (admitting lack of memory) were significantly more pronounced for the witnesses than for the uninvolved subjects, with both of these groups not differing significantly from the perpetrators (cf. Table 4). Criterion 8 (unusual details) was significantly less pronounced for the uninvolved subjects as compared with the two other experimental groups, no significant difference being found between the witnesses and perpetrators (cf. Table 4). Criterion 4 (contextual embedding) was more pronounced for the uninvolved subjects than for the witnesses, with both of these groups not differing significantly from the perpetrators (cf. Table 4).

**Discussion**

The present study examined the inter-rater reliability of CBCA (Steller & Köhnken, 1989) for statements given in a mock-crime scenario. Three raters received about 18 hours of CBCA training. The training utilized in the present study was conceived after a standardized training program developed by Krause (1997) and Petersen (1997; see also Höfer et al., 1999). After the training, each rater scored 102 transcripts of statements.
Table 4. Mean scores (M) and standard deviations (SD), separated for CBCA criteria and experimental groups; results of one-way ANOVAs and post hoc tests (Scheffé) for pairwise group comparisons; absolute value of Cohen's d for significant differences

<table>
<thead>
<tr>
<th>CBCA-criterion</th>
<th>Experimental group</th>
<th>ANOVA</th>
<th>Pairwise group comparisons (Scheffé)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>M (SD)</td>
<td>M (SD)</td>
</tr>
<tr>
<td>1. Logical structure</td>
<td>2.93 (0.16)</td>
<td>2.93 (0.20)</td>
<td>2.94 (0.15)</td>
</tr>
<tr>
<td>2. Unstructured prod.</td>
<td>0.50 (0.58)</td>
<td>0.44 (0.50)</td>
<td>0.26 (0.36)</td>
</tr>
<tr>
<td>3. Quantity of details</td>
<td>1.51 (0.73)</td>
<td>2.31 (0.70)</td>
<td>1.58 (0.62)</td>
</tr>
<tr>
<td>4. Contextual embedding</td>
<td>1.15 (0.70)</td>
<td>0.86 (0.55)</td>
<td>1.42 (0.60)</td>
</tr>
<tr>
<td>5. Descript. of interact.</td>
<td>0.21 (0.48)</td>
<td>0.32 (0.40)</td>
<td>0.41 (0.43)</td>
</tr>
<tr>
<td>6. Reprod. of convers.</td>
<td>1.48 (0.85)</td>
<td>2.27 (0.70)</td>
<td>1.40 (0.89)</td>
</tr>
<tr>
<td>7. Unexp. complications</td>
<td>0.08 (0.18)</td>
<td>0.05 (0.15)</td>
<td>0.10 (0.19)</td>
</tr>
<tr>
<td>8. Unusual details</td>
<td>0.45 (0.35)</td>
<td>0.57 (0.24)</td>
<td>0.17 (0.25)</td>
</tr>
<tr>
<td>9. Superfluous details</td>
<td>0.70 (0.61)</td>
<td>1.07 (0.60)</td>
<td>0.40 (0.38)</td>
</tr>
<tr>
<td>10. Details misunderstood</td>
<td>0.02 (0.08)</td>
<td>0.04 (0.14)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>11. Rel. ext. associations</td>
<td>0.01 (0.06)</td>
<td>0.02 (0.08)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>12. Subj. mental state</td>
<td>1.20 (0.55)</td>
<td>1.11 (0.59)</td>
<td>1.29 (0.64)</td>
</tr>
<tr>
<td>13. Perp. mental state</td>
<td>0.43 (0.53)</td>
<td>0.53 (0.55)</td>
<td>0.61 (0.59)</td>
</tr>
<tr>
<td>14. Spont. corrections</td>
<td>1.19 (0.64)</td>
<td>1.46 (0.45)</td>
<td>1.08 (0.51)</td>
</tr>
<tr>
<td>15. Lack of memory</td>
<td>0.19 (0.27)</td>
<td>0.27 (0.35)</td>
<td>0.09 (0.22)</td>
</tr>
<tr>
<td>16. Self-doubts</td>
<td>0.02 (0.08)</td>
<td>0.03 (0.10)</td>
<td>0.03 (0.10)</td>
</tr>
<tr>
<td>17. Self-deprecation</td>
<td>0.02 (0.11)</td>
<td>0.00 (0.00)</td>
<td>0.02 (0.11)</td>
</tr>
<tr>
<td>18. Pard. the perpetrator</td>
<td>0.36 (0.44)</td>
<td>0.20 (0.35)</td>
<td>0.18 (0.30)</td>
</tr>
</tbody>
</table>

Note: * p < .05.
referring to a simulated theft. Some of the statements were based on experience and some were confabulated. The raters used 4-point scales to judge the degree to which the first 18 CBCA criteria were fulfilled in each statement. Thus, for each transcript and each CBCA criterion, scores of three independent raters were obtained. Accordingly, for each of the 18 assessed CBCA criteria, the congruence of the judgments of the three raters across the 102 transcripts could be analysed. The statistical indices of inter-rater reliability calculated in the present study were: the simple and extended percentage agreement between the three raters, the average weighted kappa coefficient for the three possible two-way combinations of raters (A–B, A–C, B–C), the average product-moment correlation for the three paired rater combinations, the intra-class correlation, and the Finn-coefficient.

Simple percentage agreement was satisfactory (≥70%) for only six CBCA criteria (1, 7, 10, 11, 16, 17), with Criterion 15 only barely missing the 70% threshold. Extended percentage agreement, on the other hand, was satisfactory for 15 CBCA criteria (1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18). Here, Criteria 3 and 14 only barely missed the 70% threshold. Only for Criterion 9 was the extended percentage agreement definitely insufficient. However, measures of percentage agreement have the serious disadvantage that the proportion of chance agreements between raters (which increases with decreasing number of scale degrees) is not considered and corrected for (cf. Frick & Semmel, 1978). Thus, percentage agreement indices should, in principle, be interpreted with caution.

The mean weighted kappa coefficient, the mean product-moment correlation, and the intra-class correlation, respectively, were sufficiently high for only five CBCA criteria (3, 4, 6, 12, 13). However, the analysis of rater judgment distributions revealed that, with respect to the present data, the weighted kappa coefficient, the product-moment correlation, and the intra-class correlation were inadequate indices of inter-rater reliability. These three indices of reliability are, in principle, also very sensitive to the distribution of rater judgments across the scale degrees. They cannot sufficiently represent the true inter-rater agreement if the ratings given by the raters vary only slightly from transcript to transcript. Accordingly, the six criteria with the lowest kappa coefficients, product-moment correlations, and intra-class correlations (Criteria 1, 7, 10, 11, 16, 17) also stood out in terms of extreme skewness and extremely low variability of the ratings. Furthermore, the low \( k_w \), \( r \), and \( r_{\text{AA}} \)s obtained for these six criteria were strongly incongruent with the finding that, on the other hand, these criteria achieved very high extended percentage agreement and even the most pronounced simple percentage agreement.

Because of the problem of low variability between transcripts, the Finn-coefficient (Finn, 1970), which does not depend on variability between transcripts, was utilized as an alternative index of inter-rater reliability. For all CBCA criteria, Finn-coefficients of 0.40 or higher were obtained and thus at least indicated sufficient inter-rater agreement. Only the Finn-coefficient for Criterion 9 was clearly lower than 0.60. The Finn-coefficients were strongly consistent with the extended percentage agreements. In contrast, some of the weighted kappa coefficients, product-moment correlations, and intra-class correlations were considerably low and, thus, clearly contradicted the respective Finn-coefficients and percentage agreements. Taking into account that, on the one hand, the kappa coefficients, product-moment correlations, and intra-class correlations were not unambiguously interpretable (because of the distribution characteristics of the ratings) and that, on the other hand, the Finn-coefficients are independent of distribution parameters and stood in strong accordance with the extended percentage agreement.
agreements (that are also independent of the distribution), the Finn-coefficients were considered to be the most adequate indices of inter-rater reliability with respect to the present data. Accordingly, the overall conclusion could be drawn that, except for Criterion 9 (superfluous details), the inter-rater reliability of CBCA in the present mock-crime study was satisfactory.

In the two previous field studies focusing in detail on inter-rater reliability of CBCA (Anson et al., 1993; Horowitz et al., 1997), less favourable results were obtained than in the present study. In both studies, the raters had also been subjected to extensive training in the application of the CBCA criteria. But the ratings given were merely dichotomous; that is, the raters just had to decide whether the respective CBCA criteria were present in a statement or not. In both studies, Maxwell’s RE coefficient (Maxwell, 1977) was utilized as the crucial index of inter-rater reliability. Being conceived for dichotomous ratings only, this coefficient was not calculated in the present study. Whereas Anson et al. obtained insufficient inter-rater reliability (Maxwell’s RE < .30) for six CBCA criteria (2, 5, 11, 12, 15, 19), in the Horowitz et al. study the RE values for all CBCA criteria exceeded the threshold of .30. The mean percentage agreement between raters, as averaged across the CBCA Criteria 1 to 18, was 76.5% in the Anson et al. study and 82.1% (average of two rating sessions) in the Horowitz et al. study. Compared with this, the extended percentage agreement in the present study, as averaged across the CBCA Criteria 1 to 18, was 89.2% (cf. Table 1). This value is substantially higher than the respective values achieved in the studies by Anson et al. and Horowitz et al., especially if one considers that the rating scale utilized in the present study was relatively more sophisticated (the extended percentage agreement on the 4-point scale that tolerates disagreements by one scale degree is actually equivalent to simple percentage agreement on a 3-point scale). In addition, it is noteworthy that in the present study the calculation of percentage agreement was more conservative than in the studies by Anson et al. and Horowitz et al. In the present study, percentage agreement for each CBCA criterion was calculated in terms of the proportion of transcripts for which all three raters had chosen the same degree (or two adjacent degrees) on the 4-point scale. Contrary to this, Horowitz et al. first determined the pairwise percentage agreements for the three possible two-way combinations of the raters for each CBCA criterion. Subsequently, the overall percentage agreement between all three raters was calculated by averaging the three pairwise percentage agreements. Anson et al. had only two raters. If the calculation method from the Horowitz et al. study had been utilized for the data of the present study, much higher values of simple and extended percentage agreement than those depicted in Table 1 would have resulted. According to this calculation method, the overall percentage agreement between the three raters of the present study, as averaged across the CBCA Criteria 1–18, would have been 94.7%. Furthermore, compared with the present data, a greater proportion of chance agreements has to be supposed for the Anson et al. data as well as for the Horowitz et al. data, since the rating scales utilized in both of these studies were less graded than the rating scales in the present study. On the other hand, it has to be considered that the statements analysed in the Anson et al. and Horowitz et al. field studies related to different events, whereas in the present study all participants reported on the same mock-crime event. This might have partly contributed to the higher reliabilities obtained in the present study.

6 The lower coefficients obtained by Anson et al. (1993) – as compared with the results obtained by Horowitz et al. (1997) – may be due to the less extensive rater training and to the ratings being based on video-recordings rather than on transcripts.
Inter-rater reliability in terms of extended percentage agreement obtained in the present study was as high as the percentage agreement reported by other experimental studies using a binary scale (Vrij, Kneller, et al., 2000; Zaparniuk et al., 1995). Mean correlation coefficients and kappa values were substantially lower than those obtained in some other laboratory studies (Vrij et al., 2001a; Vrij, Edward et al., 2000). As already discussed, this could be ascribed to the extremely skewed distributions of ratings found for several CBCA criteria in the present study. The Finn-coefficients, on the other hand, indicated satisfactory inter-rater reliability also as compared with other experimental studies. A comparatively low reliability for Criterion 9 (superfluous details), as found in the present study, had already been reported by other researchers as well (e.g. Horowitz et al., 1997).

Previous field studies (Anson et al., 1993) and laboratory research utilizing the film paradigm (Akehurst et al., 2001; Vrij et al., 2001b) or the autobiographical paradigm (Sporer, 1997) found that the criteria reflecting general characteristics of the statement (CBCA Criteria 1, 2, 3) were judged less reliably than criteria in the specific contents and peculiarities of content categories (Criteria 4–13). In the present study, this was the case only for Criterion 3 (quantity of details), whereas Criteria 1 and 2 did not prove to be (substantially) less reliable than the criteria in the other CBCA categories. Motivation-related contents (Criteria 14–18) were rated highly reliable with the exception of Criterion 14, which was already found to be relatively unreliable in several other studies (Akehurst et al., 2001; Anson et al., 1993; Horowitz et al., 1997; Vrij et al., 2001a).

Sufficient inter-rater reliability is only useful if CBCA can help us to differentiate between truthful and fabricated accounts. Accordingly, the validity of CBCA was also analysed in the present study and could, on the whole, be corroborated. The total CBCA scores (calculated by summing the scores of the 18 CBCA criteria for each transcript, with the scores of the three raters being averaged, respectively) were significantly higher for the truthful statements (witnesses) than for the confabulated accounts (perpetrators and uninvolved subjects). Multivariate analyses also yielded a significant simultaneous effect of the group factor on the scores for the 18 CBCA criteria, with the truthful statements differing significantly from the fabricated ones. Finally, univariate analyses that were calculated separately for each CBCA criterion showed that the differentiation of truthful and fabricated accounts was mainly due to the Criteria 3 (quantity of details), 6 (reproduction of conversation), and 9 (superfluous details). These criteria were significantly more pronounced in the truthful accounts given by the witnesses than in the confabulated accounts given by the perpetrators and the uninvolved subjects. Furthermore, the Criteria 8 (unusual details), 14 (spontaneous corrections), and 15 (admitting lack of memory) were also significantly more present in the truthful accounts of the witnesses than in the confabulated accounts of the uninvolved subjects (but not of the perpetrators), with Criterion 8 also being more pronounced in the confabulated statements given by the perpetrators than in those given by the uninvolved subjects. Criterion 4 (contextual embedding), however, was significantly less pronounced in the truthful statements of the witnesses than in the fabricated accounts of the uninvolved subjects and, thus, was the only criterion that contradicted the basic axiom of SVA.

Interestingly, although Criteria 9 and 14 were judged less reliable than the other ones, they proved to be valid in differentiating between truthful and fabricated accounts (Criterion 9) and between witnesses and uninvolved subjects (Criterion 14), respectively. A further clarification or redefinition of these criteria with the aim of enhancing reliability might even improve their validity, too.
The differentiation of truthful and fabricated accounts obtained in the present study was satisfactory too, if compared with the results of a very similar mock-crime experiment in which the participants also reported on a simulated theft of money (Porter & Yuille, 1996). These authors found, that three out of 10 CBCA criteria assessed (1. logical structure, 3. quantity of details, 15. admitting lack of memory) differentiated significantly between truthful and fabricated statements. Unfortunately, Porter and Yuille reported no data on total CBCA scores.

Conclusions

In summary, the inter-rater reliability obtained in the present study was satisfactory not only if considered absolutely, but also if compared with results of other field and experimental studies. The present data suggest that, given thorough rater training, CBCA can be applied in the experimental mock-crime paradigm with a sufficient degree of inter-rater reliability. Furthermore, CBCA's validity in differentiating between truthful and fabricated accounts of a simulated theft could be corroborated suggesting that the mock-crime paradigm is a useful tool for the further investigation and development of CBCA/SVA. The present findings also corroborate the effectiveness of the CBCA training program as conceived by Krause (1997) and Petersen (1997; see also Höfer et al., 1999). In this regard, it has to be stressed that compared with the original proposal by Krause and Petersen, less time was devoted to the rater training in the present study. Probably, with a thorough implementation of the original training program, even higher inter-rater reliabilities would have been obtained for single CBCA criteria.

Acknowledgements

The authors are indebted to Katrin Dumbert and Kirsti Maron for their assistance during data collection, to Professor Günter Köhnken for supplying rater training manuals, and to Alfred Weber for his helpful comments during the preparation of this article. Additionally, we would like to thank Professor Siegfried L. Sporer and an anonymous reviewer for commenting on an earlier version of this manuscript.

References


Appendix

Detailed description of the rater training utilized in the present study

The first training session comprised the following steps:

- introduction and overview of the curriculum
- review and discussion of the scientific literature (see above)
- definition and discussion of the terms 'general credibility' (from German 'allgemeine Glaubwürdigkeit'; i.e. credibility conceived as a personality trait) versus 'specific credibility' (from German 'spezielle Glaubhaftigkeit'; i.e. validity of a specific statement, independent of personality traits), 'deception', and 'lie'
- exploration of the trainees’ knowledge about SVA/CBCA
- discussion of the main axioms (especially the so-called ‘Undeutsch-Hypothesis’, Undeutsch, 1967 [term after Steller, 1989, p. 145]) and principle components of SVA (structured interview, CBCA, validity checklist)
- introduction of the five superordinate categories of CBCA criteria (general characteristics, specific contents, peculiarities of content, motivation-related contents, details characteristic of the offence)
- general information on estimation biases and discussion of potential sources of error in CBCA scoring
The second training session comprised the following steps:
- rehearsal of the criterion definitions and scoring rules
- exercise on the basis of short transcripts (identification of single criteria without scaling the degree of presence, each trainee working on her own)
- group discussion of the short transcripts/feedback by the trainer (first author)
- homework: identifying several criteria in a transcript (without scaling the degree of presence).

The third training session comprised the following steps:
- rehearsal of the CBCA categories and criteria, description of the criteria by the trainees
- group discussion of the homework
- production training: trainees produce statements that contain single CBCA criteria. Subsequently, the other trainees who are blind with respect to the criteria produced try to identify them.
- homework: scoring of several transcripts for the presence/absence of all CBCA criteria (no scoring of the degree of presence).

The fourth training session comprised the following steps:
- rehearsal of the CBCA categories and criteria; discussion of ambiguities regarding their definition
- discussion of the homework results; feedback by the trainer
- specification of how the degree of presence of the criteria is scored correctly on the 4-point (0, 1, 2, 3) rating scales
- specification of how criterion-fulfilling statement passages have to be marked in the transcript in order to achieve maximum transparency/retraceability of the scoring
- homework: identifying several CBCA criteria in a transcript and scoring their degrees of presence.

The fifth training session comprised the following steps:
- rehearsal of the CBCA categories and criteria; discussion of ambiguities regarding their definition and scoring
- rehearsal of the scoring rules
- rehearsal of the principles regarding the marking of criterion-fulfilling transcript passages
- scoring of short transcripts (each trainee on her own), subsequently group discussion of the results
- discussion of the homework results
- Production training: each trainee generates, on her own, three examples for five of the CBCA Criteria 4–18, respectively. The three examples for a criterion have to differ with respect to the degree of presence of the respective criterion (scale points 1, 2, or 3). Subsequently, the two other trainees have to identify the criteria and rank them with respect to their degree of presence in the examples.

- Homework: identifying and scoring the CBCA Criteria 1–18 in a transcript.

The sixth training session comprised the following steps:
- Standardized test (paper-and-pencil) of CBCA-related knowledge
- Rehearsal of the CBCA categories and criteria; discussion of ambiguities regarding their scoring
- Discussion of the homework results
- Production training: each trainee generates, on her own, three examples for five of the CBCA Criteria 4–18, respectively. The three examples for a criterion have to differ with respect to the degree of presence of the respective criterion (scale points 1, 2, or 3). Subsequently, the two other trainees have to identify the criteria and score them on the 4-point scale.
- Homework: identifying and scoring the CBCA Criteria 1–18 in several transcripts.

The seventh training session comprised the following steps:
- Rehearsal of the CBCA categories and criteria; discussion of ambiguities regarding their scoring
- Feedback and discussion of the test results (CBCA knowledge) from Session 6
- Discussion of the homework results
- Final discussion of any unresolved ambiguities regarding CBCA.