韓國心理學會誌 : 社會 및 性格
Korean Journal of Social and Personality Psychology
2018, Vol. 32, No. 2, 67-83

성폭력 피해아동 진술신빙성 판단에 있어서 평가자간 신뢰도: 진술분석 전문가 집단을 대상으로*

이 미 선†

동양대학교

CBCA(Criteria-Based Content Analysis)는 성폭력 피해 아동 진술의 신빙성을 판단하기 위한 진술분석 기법이다. 총 19개의 준거로 구성되어 있으며 개별 준거는 진실한 아동에게서 더 많이 나타나는 것으로 고려된다. 본 연구는 경찰청 소속 진술분석 전문가들을 대상으로 CBCA 분석의 평가자간 신뢰도를 계산하였다. 총 아홉 명의 진술분석 전문가들은 동일한 다섯 건의 아동 성폭력 사건에 대하여 CBCA 준거의 존재 여부 및 종합적인 신빙성 판단을 실시하였다. 두 명의 평가자간 일치율 평균을 확인한 결과 총 19개 준거 중 11개 준거의 평가자간 신뢰도는 '적절한 수준'인 것으로 나타났다. 반면, 준거2(구조화되지 않은 표현), 준거5(상호작용), 준거6(대화의 인용), 준거7(사건 중 예기치 않은 일 발생), 준거8(독특한 세부내용), 준거9(부가적인 세부내용), 준거11(관련된 외적 연합), 준거14(자발적 수정)의 경우 평가가간 낮은 수준의 일치도를 보였다. 아동 진술의 신빙성에 대한 최종 판단에 있어서 평가자간 신뢰도는 '적절한 수준'으로, 비교적 평가자간 일관된 결론을 내리고 있는 것으로 나타났다. 마지막으로 본 연구 결과를 바탕으로 사법현장에서 진술분석 의견서 활용에 대해 논의하였다.

주요어: 진술분석, 진술타당도평가, 준거기반내용분석, CBCA, 진술분석전문가, 평가자간 신뢰도

* 본 연구는 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016S1A5B5A02022368).
† 교신저자: 이미선, 동양대학교 경찰행정과, 경북 영주시 풍기읍 동양대로 145 동양대학교 다산관 5205호
E-mail: msy23@dyu.ac.kr

최근 아동 대상 범죄가 지속적으로 증가하면서 아동이 법정에 피해자 또는 목격자로 참여하는 경우가 증가하고 있다. 특히, 아동 성폭력 사건의 경우 대부분 물적, 인적 증거가 존재하지 않으며 아동 진술이 사건의 유일한 증거가 되는 경우가 빈번하기 때문에 진술의 신빙성에 대한 판단은 사건을 해결하는데 있어서 결정적인 요인으로 작용한다. 그러나 안타깝게도 아동의 언어적, 인지적 능력의 미숙과(Fivush, Gray, & Fromhoff, 1987), 높은 피암시성(Whipple, 1912)으로 인해 아동 진술의 신빙성은 역사적으로 부정되어 왔다(Ceci, & Bruck, 1993).

진술타당도평가(Statement Validity Assessment: SVA)는 성폭력 피해아동 진술의 신빙성을 판단하는 기법이다(Steller, 1989). 1950년대 독일의 심리학자들은 '실제로 어떤 사건을 경험한 아동의 진술은 허구나 상상에 기초한 진술과는 내용과 질적인 측면에서 차이가 있다'라는 Undeutsch 가설을 기초로 진술의 신빙성에 대한 평가를 실시하였으며(Steller, 1989), 이후 Steller와 Koehnken (1989)는 진술신빙성 판단을 위해 사용되었던 다양한 내용 준거들을 체계화하여 통합된 절차로 발달시켰다. 진술타당도평가는 3단계로 이루어져 있다. 첫 번째는 반구조화된 면담을 통해 아동으로부터 온전한 진술을 확보하는 단계이다. 이때, 수사관의 영향을 배제한 아동의 자발적인

표 1. CBCA 준거 및 준거 설명

준거	준거 설명
1. 논리적 일관성	진술의 일관성과 통일성
2. 구조화되지 않은 표현	진술의 시간적 순서를 따르지 않고 자연스러운 방식으로 제시
3. 세부내용의 풍부함	언제, 어디서, 누가, 무엇을 등의 구체적이고 상세한 세부묘사
4. 맥락상 깊이	사건에서 맥락의 정보가 풍부하게 제시
5. 상호작용	가해자와 피해자간의 행동과 행동에 대한 반응이 제시
6. 대화의 인용	사건 당시 가해자와 나누었던 대화를 직접적으로 인용
7. 사건 중 예기치 않은 일 발생	사건 중 예상치 못한 중단, 어려움, 자발적인 종료
8. 독특한 세부내용	실제 그러한 상황이 존재했다는 것을 확인할 수 있는 특징적 진술
9. 부가적인 세부내용	사건 맥락을 이해하는데 도움을 주는 세부 묘사
10. 정확하지만 이해하지 못한 내용	자신이 경험한 것이 이해하지 못하였으나 설명 (특히 성적인 행동)
11. 관련된 외적 연합	사건에 대한 직접적 진술은 아니나, 사건이 있었음을 암시하는 진술
12. 주관적 심리상태 묘사	사건 당시 피해자 자신이 주관적 인지 또는 정서 상태 표현
13. 가해자의 정신상태 귀인	가해자의 정신 상태에 대한 추론
14. 자발적 수정	자신의 진술을 자연스럽게 수정
15. 기억의 부족 시인	기억이 나지 않은 부분에 대해 자연스럽게 시인
16. 자기 진술에 대한 의심제기	자신의 진술이 정확하지 않을 수 있음을 염려
17. 자기 비난	자기에게 불리하거나 혐의를 초래 할 수 있는 진술
18. 가해자 용서	가해자의 행동의 심각성을 최소화하거나 용서해주는 진술
19. 범죄 특징에 대한 세부내용	특정한 범죄에서 전형적으로 일어나는 방식으로 사건을 진술

※ Steller와 Koehnken(1989) 참고

진술을 받는 것은 진술분석 결과의 타당성을 확보하는데 있어서 매우 중요한 요소가 된다. 두 번째 단계는 SVA의 핵심으로 준거기반내용분석(Criteria-Based Content Analysis; CBCA)을 통해 진술의 신빙성을 객관적으로 판단하는 과정이다. CBCA는 진실한 진술에서 더 많이 존재하는 것으로 고려되는 19개의 내용 준거로 구성되어 있다(표 1 참조). 따라서 아동 진술에서 더 많은 수의 CBCA 준거가 더 명백하게 존재할수록 진술의 신빙성은 높다고 평가된다(Vrij, 2008). 마지막으로 타당도 체크리스트(Validity checklists)는 진술의 허위 가능성을 판단하는 단계이다. CBCA 분석과는 별도로 아동의 언어와 지식의 적합성, 정서의 적합성, 암시 취약성, 강압적 조사면담 및 허위로 고소할 동기 유무 등을 평가하여, CBCA 결과가 진술의 신빙성 이외 다른 요인에 의해 영향 받았을 가능성을 배제한다(Vrij, 2008).

CBCA가 진실한 진술과 허위 진술을 신뢰롭게 구별해 줄 수 있다는 것은 이미 다수의 현장연구(Boychuk, 1991; Craig, Scheibe, Raskin, Kircher, & Dodd, 1999; Esplin, Boychuk, & Raskin, 1988; Lamb, Sternberg, Esplin, Jershkowitz, Orbach, & Hovav, 1997) 및 실험연구를 통해 반복적으로 증명되고 있다(Amado, Arce, & Fariña, 2015; Oberlader, Naefgen, Koppehele-Gossel, Quinten, Banse, & Schmidt, 2016; Vrij, 2005 참고). 지금까지 실시된 CBCA 연구에 대한 메타분석 결과, CBCA은 진실한 진술과 허위 진술을 유의미하게 구별할 수 있으며, 정확률은 대략 70% 정도인 것으로 나타났다. 특히 실제 아동 성폭력 사건을 대상으로 한 현장연구는 가상적인 모의상황에서 이루어진 실험연구에 비해 더 높은 정확률을 보였다(Amado et al., 2015; Oberlader et al., 2016; Vrij, 2005). 우리나라에서 실시된 CBCA 타당도 연구 결과 역시 기존 다른 언어권 국가에서 실시된 연구 결과와 유사하였다. 즉, 실제 성폭력을 경험한 것으로 고려되는 아동의 진술에는 성폭력이 의심스러운 아동의 진술에서보다 더 많은 CBCA 준거들이 포함되는 것으로 나타났다(고은영, 채규만, 2008; 김현정, 2010; 이미선, 2004; 이수정, 2010).

비록 CBCA는 진술의 신빙성을 판단하는 데 있어 유용한 도구임은 분명하지만 평가자에 따라 분석 내용이 상이하다면 CBCA 분석의 타당성을 확보할 수 없다. 그럼에도 지금까지 CBCA와 관련된 대부분의 연구는 CBCA 기법의 타당도에 초점을 두었으며, CBCA 분석의 신뢰도에 대해서는 거의 알려진 바가 없다. 평가자간 신뢰도(inter-rater reliability)은 평가자간 분석 결과의 일관성을 의미하는 것으로(성태제, 2002), 다수의 평가자들이 동일한 사건에 대해 얼마나 일관된 결과를 도출할 수 있는지와 관련되어 있다. CBCA의 경우 명확한 채점기준이 부재하며, CBCA 분석 결과를 통한 신빙성 판단 기준이 존재하지 않기 때문에 진술의 신빙성에 대한 최종 결정은 평가자의 주관적인 판단에 맡겨지는 경향이 있다는 비판이 제기되고 있다(Mazzoni, & Ambrosi, 2003). 이에, CBCA 분석의 평가자간 신뢰도를 확인하는 것은 분석의 객관성을 확보하는데 있어서 매우 중요할 것으로 판단된다.

평가자간 신뢰도를 나타내기 위해 다양한 일치계수(coefficient of agrement)가 사용된다(박광배, 엄진섭, 1996). 일반적으로 분석 방법과 상관없이 일치계수가 .00 미만인 경우 '일치도 거의 없음'으로 판단하며, .00 ~ .20 이하는 '약간의 일치도 존재', .20 ~ .40 이하의 경우 '일정 수준 일치도 존재', .40 ~ .60 이하 '적절한 수준 일치도 존재', .60 ~ .80 이하 '좋은 수준 일치도 존재' 그리고 .80 이상의 경우 '거의 완벽한 수준 일치도 존재'를 의미한다(Fleiss, 1981; Vrij, 2005 재인용; Landis, & Koch, 1977). 따라서 연구에서는 일치도 계수가 .40 이상(적절한 수준)일 때 신뢰도가 존재하는 것으로 고려한다(Fleiss, 1981).

이러한 기준으로 볼 때, 지금까지 실시된 CBCA 연구에서 평가자간 신뢰도는 대부분 '좋은 수준' 으로 나타났다(Vrij, 2005). 최근 실시된

메타 연구 결과 역시 준거별 일치계수의 범위는 -.22에서 1.00 사이로 다양하였지만 전반적으로 '적절한' 수준의 평가자간 신뢰도를 보였다(Hauch, Sporer, Masip, & Blandon-Gitlin, 2017). 다만 앞서 설명한 메타분석(Vrij, 2005; Hauch et al., 2017)에 포함된 연구들은 CBCA 분석의 타당성을 확보하기 위한 목적으로 평가자간 신뢰도를 분석했기 때문에, 연구자들은 평가자간 일정 수준 이상의 일치도에 이르기까지 지속적이고, 반복적인 훈련을 실시하였다. 따라서 이러한 연구에서 평가자간 높은 신뢰도를 보고하는 것은 당연한 결과로 볼 수 있다.

지금까지 직접적으로 CBCA 분석의 평가자간 신뢰도를 확인한 연구는 기존 연구 결과와는 차이가 있는 것으로 나타났다(Akehurst, Manton, & Quandte, 2011; Anson, Golding, & Gully, 1993; Gödert, Gamer, Rill, & Vossel, 2005; Horowitz, Lamb, Esplin, Boychuck, Krispin, & Reiter-Lavery, 1997; Niveau, Lacasa, Berclaz, & Germond, 2015). Anson 등(1993)의 연구에서는 CBCA 워크숍에 참석했던 총 네 명의 연구자들을 대상으로 평가자간 신뢰도 분석을 실시하였다. 평가자들은 23개의 실제 성폭력 사건 비디오를 시청 후에 CBCA 분석을 실시하였으며, 평가자간 일치도를 확인하기 위해 Maxwell RE 계수를 계산하였다. 연구 결과 대부분 준거들은 '거의 완벽한 신뢰도'(준거18, 준거16, 준거13, 준거10), '좋은 수준의 신뢰도'(준거17, 준거6, 준거3, 준거1, 준거7), '적절한 수준의 신뢰도'(준거9, 준거4, 준거8, 준거4, 준거14)를 보였다. 다만 준거2(구조화되지 않은 표현), 준거5(상호작용), 준거11(관련된 외적 연합), 준거12(주관적 심리상태 묘사), 준거15(기억의 부족 시인), 준거19(범죄 특징에 대한 세부내용)의 경우 평가자간 일치도가 거의 없는 것으로 나타났다. Horowitz 등(1997)의 연구에서는 한 명의 CBCA 전문가와 두 명의 대학원생 연구자들이 아동 성폭력 사건에 대한 CBCA 분석을 실시하였으며, 이후 세 명의 평가자간 신뢰도를 계산하였다. 연구 결과 대부분은 '좋은 수준' 또는 '적절한 수준'이었으며, 준거9(부가적 세부내용), 준거14(자발적 수정), 준거15(기억부족 시인) 만이 '일정수준(.20 ~ .40)'의 일치율을 나타냈다.

앞서 설명한 연구들이(Anson et al., 1993; Horowitz et al., 1997) 실제 성폭력 사건에 초점을 두었다면, Gödert 등(2005)의 연구에서는 모의 범죄 실험을 통해 확보한 진술에 대한 CBCA 분석을 실시하였다. 평가자간 신뢰도 분석을 위해 세 명의 평가자들이 참여하였으며, 개별 준거에 대한 Weighted Kappa 계수를 산출하였다. Gödert 등(2005)의 실험연구에서 평가자간 신뢰도는 기존 현장 연구에서 산출된 평가자간 신뢰도에 비해 좀 더 낮은 일치율을 보였는데, 특히 준거1(논리적 일관성), 준거7(사건 중 예기치 않은 일 발생), 준거9(부가적인 세부내용), 준거10(정확하지만 이해하지 못한 내용), 준거16(자기 진술에 대한 의심제기), 준거17(자기 비난)은 일치도 계수가 .20 미만으로 나타났다.

지금까지 연구 결과를 종합해 보면, 개별 준거에 대한 일치율은 대부분 적절한 수준 이상으로 나타났지만 몇몇 준거들의 평가자간 신뢰도는 낮은 수준으로 사실상 평가자간 동일한 판단을 하고 있다고 보기 어려웠다. 특히 낮은 일치율을 보인 준거들은 실시된 연구마다 각각 다르게 나타났기 때문에 기존 연구만으로 CBCA 준거별 평가자간 신뢰도에 대한 종합적인 결론을 내리기는 어려울 것으로 판단된다. 또한 기존 연구에서는 연구를 목적으로 동일한 교육과 훈련을 받은 연구자들이 분석을 실시하였으며, 유일하게 Horowitz 등(1997)의 연구만이 한 명의 현장 실무자가 포함되어 있었다. 따라서 지금까지 연구 결과만으로 진술분석 실무자들이 얼마나 일관된 평가를 할 수 있는지를 확인할 수 없다. 현장 전문가들의 경우, 개인에 따라 교육수준과 배경지식 및 실무경험이 다르기 때문에 동일한 기간 동안 동일한 교육을 받은 연구자들이 실시한 분석과는 평가자간 신뢰도에 있어서 차이가 있을 수 있음을 예상해 볼 수 있으나, 지금까지

CBCA 실무자들의 진술분석 결과의 일치도에 대해서는 여전히 확인된 바 없다.

우리나라의 경우 아동 성폭력 피해자 진술의 신빙성을 평가하기 위해 2010년부터 진술전문가 제도를 도입하여, 2016년 현재 총 백사 명의 진술분석관들이 전국적으로 활동하고 있다(경찰청, 2017). 아동과 장애인 성폭력 사건에서 진술분석관들의 의견서는 사건 수사 및 재판에 중요한 영향을 미치고 있으며(박종선, 2013; 이수정, 2011), 이에 진술분석관들의 판단은 매우 높은 수준의 전문성이 요구되는 실정이다. 하지만 지금과 같은 진술분석관 양성 과정만으로 성폭력 피해 아동 진술의 신빙성 판단에 전문성을 확보할 수 있을지에 대해서는 의문이 제기된다. 일반적으로 진술분석관들은 3 ~ 5일의 이론 교육 및 30여 시간의 실습 후 적격심사를 통해 전문가로 위촉되어 활동하게 된다(경찰청, 2017). 비록 진술분석 전문가 위촉을 위해서는 심리학 분야의 전문가로부터 추천이 필요하기 때문에 진술분석관들의 전문성에 대한 판단을 실시하기는 하지만, 진술분석 경력은 전무한 상태에서 단기간의 교육과 실습을 이수한 후 전문가로 활동을 시작하게 되는 것이다. 그럼에도 최근까지 진술분석 전문가들의 전문성에 대한 실증적인 검증이 이뤄지지 않았으며, 진술분석관들이 실시한 CBCA 진술분석이 얼마나 일관된 결과를 도출하고 있는지에 대한 연구는 부재하다.

이에 본 연구의 목적은 다음과 같다. 첫째, CBCA 개별 준거 평가에 있어서 평가자간 신뢰도를 확인한다. 둘째, 아동 진술의 최종 신빙성 판단에 있어서 평가자간 신뢰도를 확인한다. 마지막으로 진술분석관들이 진술신빙성 판단 시 중요하게 고려하는 요인 및 근무환경에 대한 자유로운 의견을 수집하고자 한다.

연구방법

연구대상

총 아홉 명의 경찰청 소속 진술분석 전문가들이 본 연구에 참여하였다. 연구에 참여한 평가자들의 평균 연령은 40.67세(표준편차 6.42)이었다. 참가자들은 모두 여성으로 심리학 관련 분야 석사 이상의 학력을 가지고 있었다. 전공을 살펴보면 상담심리학, 발달심리학, 범죄심리학, 임상심리학 등으로 나타났다. 진술분석관으로 활동한 경력은 최소 10개월부터 최대 72개월이었으며, 평균 근무 경력은 50.1개월(표준편차 23.05)이었다. 평가자들은 주로 심리상담사, 임상심리사, 발달심리사, 범죄심리사, 피해상담사 등의 자격증을 소지하고 있었으며, 진술분석관으로 활동한 경력 이외에 기타 관련 업무 경력은 평가자에 따라 42개월에서 192개월로 나타났다. 진술분석 의견서 제출 회수는 평가자 2, 3, 4의 경우 각각 2회, 10회, 15회이었으며, 나머지 전문가들은 현재까지 대략 100회에서 270회까지 의견서를 제출한 것으로 나타났다(표 2 참고).

연구 도구

아동성폭력 사건 면담자료

진술분석을 위해 총 다섯 건의 아동 성폭력 사건의 녹취록이 제작되었다(표 3 참고). 아동 성폭력 사건 녹취록은 실제 진술분석 전문가들이 진술분석 시 활용하는 수사면담 녹취록과 동일한 형태로 수사관과 아동 간 문답으로 구성하였다. 우리나라의 경우, 성폭력 사건에 대한 경찰관의 수사면담은 NICHD 조사면담 프로토콜에 따라 사건관련 면담이 실시되기 이전에 기본 규칙 설명, 라포형성, 진술훈련 등 다양한 과정이 포함되어 있다(Lamb, Hershkowitz, Orbach, & Esplin, 2011). 다만 분석을 위해 제작된 녹취록은 사건 관련 면담 이외 다른 단계는 생략하였다. 개별 사건 면담 자료를 구성하기 위해 동일 연령의 피해자가 포함된 다수의 실제 아동 성폭력 사건의 수사면담 녹취록 내용을 참고하여 재구

표 2. 참가자의 특성

	나이	성별	최종 학력	전공 (석,박사)	진술분석 경력(개월)	의견서 제출회수	기타경력 (개월)	자격증
평가자 1	47	여	대학원 (박사)	상담심리, 교육상담	50	270	55	상담심리사2급 피해상담사2급 범죄심리사1급
평가자 2	37	여	대학원 (석사)	범죄심리	60	2	28	범죄심리사1급 피해상담사2급
평가자 3	42	여	대학원 (석사)	상담 및 임상심리	10	15	196	정신보건임상심리사2급
평가자 4	35	여	대학원 (석사)	발달심리	21	10	60	발달심리사2급
평가자 5	40	여	대학원 (석사)	심리학	72	200	96	재활심리치료사 청소년 상담사 임상심리전문가
평가자 6	51	여	대학원 (박사)	상담심리	58	110	122	범죄심리사1급 전문상담사1급 정신보건상담사2급
평가자 7	31	여	대학원 (석사)	범죄심리	72	100	42	범죄심리사1급 사회복지사 성폭력가정폭력상담사
평가자 8	46	여	대학원 (박사)	사회복지, 상담심리	72	130	192	사회복지사 언어재활사
평가자 9	37	여	대학원 (석사)	범죄심리	36	250	99	범죄심리사1급

성하였다. 성폭력 사건에 대한 아동 진술의 사실성을 높이기 위해서 진술 내용 및 진술방식 등은 실제 사건에 근거하여 사실적으로 표현하였다. 다만 개인이나, 특정 사건을 식별할 수 있는 어떠한 정보도 포함하지 않았다. 따라서 본 연구에 포함된 다섯 건의 사건은 실제 사건을 바탕으로 제작되었다는 점에서 사실성이 높지만 실제 사건으로 보기는 어렵다.

다섯 건의 사건 중 세 건은 강제추행 사건이었으며, 두 건은 강간 사건으로 피해 유형, 발생 장소, 가해자와 관계 등은 사건에 따라 다양하였다. 녹취록에 포함된 피해 아동의 연령은 12세 두 명, 9세 한 명, 7세가 두 명이었다. 면담 시 아동과 수사관 사이의 문답 수는 대략 55개이었다. 사건1의 경우 전체 문답수가 85개로 가장 많았으며, 사건2의 문답 수는 41개로 가장 적었다. 아동 진술의 양적인 평가를 위해 응답수와 질문 당 응답수를 계산하였다. 아동의 응답은 우리나라 표준 맞춤법 기준으로 띄어쓰기를 수정한 후, 띄어쓰기를 단위로 하여 응답수를 계산하였다. 전체 아동의 응답 수는 사건1과 2의 경우 각각 152개와 144개이었으며, 사건3은 832개, 사건4는 414개, 사건5는 234개이었다. 질문 당 응답 수는 사건3의 경우 13.21개로 가장

표 3. CBCA 분석에 사용된 사건

	아동 연령	아동 성별	피해 유형	문답수	전체 응답수	질문당 응답수	사건내용
사건 1	12	여	강제 추행	85	152	1.79	아동은 집에 가는 길에 뒤쫓아 온 낯선 남성이 승강기에서 자신을 성추행 했다고 주장
사건 2	12	여	강간	41	144	3.51	아동은 자신의 집에서 삼촌으로 부르는 사람에게 강간을 당했다고 주장
사건 3	9	여	강간	63	832	13.21	아동은 함께 살고 있는 엄마의 남자친구에게 자신의 집에서 강간을 당했다고 주장
사건 4	7	남	강제 추행	44	414	9.41	아동은 동생과 집에 가는 길에 낯선 남성이 자신을 그의 집으로 데리고 가서 성추행 했다고 주장
사건 5	7	여	강제 추행	43	234	5.44	아동은 2층에 사는 오빠가 자신을 10층 계단으로 데리고 가서 성추행 했다고 주장

많았고, 사건4의 경우 9.41개, 사건5는 5.44개, 사건2는 3.51개, 사건1은 1.79개로 가장 적었다(표 3 참고).

진술분석

연구에 참여한 전문가들은 총 다섯 건의 아동 성폭력 사건에 대해 CBCA 분석을 실시하였다. 분석은 개별 녹취록을 검토한 이후 (1) CBCA 준거의 존재 여부에 따라 '부재', '존재', '강하게 존재'로 평가 하였다. 다만 우리나라에서 진술분석 의견서를 작성 시 진술분석관들의 개별 준거에 대한 평가는 3점 척도가 아닌 2점 척도(존재/부재) 상에서 이루어지는 점을 고려하여, 개별 CBCA에 대한 평가는 이후 3점 척도(부재/존재/강하게 존재)에서 2점 척도(유/무)로 수정되었다.

CBCA 준거 존재 여부와는 별개로 아동 진술의 신빙성에 대한 종합적인 판단을 실시하였다. 진술신빙성 판단은 진술분석 결과를 바탕으로 개별 사건에 대해 진술의 (2) 신빙성 존재 유무(신빙성 있음/ 신빙성 없음) 및 (3) 백분율(0% 신빙성 없음 ~ 100% 신빙성 있음)로 나타냈다.

설문지구성

설문지는 세 부분으로 구성되어 있다. 첫 번째는 연구 참가자들의 성별, 연령, 경력 등 기본 정보에 대해 질문하였다(표 2 참고). 두 번째 부분은 진술신빙성 판단 시 가장 중요하게 고려하는 요소 및 어려움에 대해 자유 서술하도록 하였다. 세 번째 부분에서는 근무환경과 관련된 내용으로 (1) 근무하면서 어려운 점, (2) 전문가 업무와 관련하여 필요한 교육, 그리고 (3) 진술분석관 제도에서 개선되어야 할 사항에 대해 자유 서술하였다.

연구 절차

연구 참가자 모집을 위해 전국 원스톱센터의 연락처 및 팩스번호를 홈페이지를 통해 확인하였다. 이후 참가자 모집 공고문을 전국 원스톱센터에 팩스로 전송한 후, 전화로 연락하여 연구의 취지를 소개하고 진술분석 전문가들의 참여를 독려하였다. 추가적으로 진술분석 전문가들의 인터넷 커뮤니티 사이트에 연구 참가자 모집 공고를 게시하였으며, 참가를 희망하는 사람은 이메일 또는 전화로 참가 신청을 할 수 있도록 하였다. 연구 참여를 신청한 진술분석 전문가에게는 전화로 연구의 목적과 진행과정을 설명하였으며 최종적으로 서명한 참가 동의서를

이메일을 통해 회수하였다. 연구 참가를 희망하는 전문가들은 이메일 통해 다섯 건의 성폭력 사건 녹취록과 설문지를 전송 받았다. 진술분석 전문가들은 CBCA 분석 결과 및 설문지를 3주 이내로 연구자에게 이메일을 통해 회신해야 하며, 연구에 참여하는 동안 연구 내용 및 분석 결과를 타인에게 알리는 것이 금지되었음을 안내 받았다. 최종적으로 아홉 명의 전문가들이 연구에 참여하였으며, 아홉 명 모두 3주 이내에 CBCA 분석 결과 및 설문지를 제출하였다. 연구에 참여한 전문가에게는 20만원의 참가비가 제공되었다.

통계분석

개별 CBCA 준거 및 최종 신빙성 판단에 대한 아홉 명의 평가자간 신뢰도를 확인하기 위해 평가자간 일치율(percentage agreement), Fleiss 카파계수, Maxwell Random Error 계수, 그리고 Intra-class correlation coefficient(ICC)를 계산하였다. 일치율의 경우 평가자간 어떤 유목이나 범주에 대한 분류의 일치도를 추정하는 방법으로 평가자간 동일한 답변을 한 비율을 확인하게 된다(성태제, 2002). 아홉 명의 평가자에 대한 일치율의 경우 전원이 동일하게 답변한 경우만을 일치하는 것으로 판단하기 때문에, 일치율 판단은 과소 추정될 가능성이 있다. 따라서 본 연구에서는 아홉 명의 평가자간 일치율에 대한 분석 이외 가능한 모든 두 명의 평가자 조합을 만들어서 두 명의 평가자의 일치율을 구한 후, 이에 대한 평균을 추가적으로 제시하였다. 총 아홉 명의 평가자들의 가능한 두 명 조합은 36개(예: 평가자1-평가자2, 평가자1-평가자3, 평가자1-평가자4.... 평가자8-평가자7, 평가자8-평가자9)이다. 따라서 본 연구에서는 총 36개의 두 명의 평가자간 일치율들의 평균을 제시하였다. 일치율의 경우 .70 이상인 경우 '평가자간 일관성이 높다'라고 판단된다(Wellershaus, & Wolf, 1989; Gödert et al., 2005 재인용).

다만 일치율의 경우 우연에 의해 평가자가 동일한 답변을 했을 가능성이 고려되지 않기 때문에 일치도는 과대 추정되는 경향이 있다(Frick, & Semmel, 1978). Fleiss 카파는 우연에 의한 가능성을 고려하여 두 명 이상이 평가자의 범주형 데이터 평가에 대한 일치도를 확인하는 분석이다(Conger, 1980). 본 연구는 아홉 명의 평가자가 이분형 범주에 대해 평가한 것으로, 아홉 명간 일치율 분석을 위해 Fleiss 카파계수를 산출하였다.

추가적으로 두 명의 평가자간 일치도 평균을 계산하기 위해서는 Maxwell RE 계수를 확인하였다. 일반적으로 두 명의 평가자간 카파계수를 계산하는 경우 Cohen의 카파계수가 사용되지만, Cohen의 카파의 경우 발생 기저율이 .50에서 확연히 벗어날 때에는 일치도가 낮게 추정될 수 있다는 한계를 갖는다(Spitznagel, & Helzer, 1985). CBCA의 경우 어떤 준거는 좀 더 빈번하게 나타나는 반면 다른 준거들은 거의 나타나지 않기 때문에 개별 준거 발생 기저율을 .50로 가정하기 어렵다(Anson et al., 1993, Horowitz et al., 1997). 반면 Maxwell RE의 경우 기저율이 .50로 가정되지 않은 상황에서 평가자들의 수행을 바탕으로 기저율을 설정하게 되므로(Maxwell, 1997), 본 연구에서는 RE 계수를 계산하는 것이 적합하다고 판단하였다. Maxwell RE 계수는 두 명의 평가자간 이분형 자료 분석에 사용되며, 카파계수와 거의 유사하게 해석된다(Maxwell, 1977). 마지막으로 진술의 진실 가능성에 대한 백분율은 ICC 분석을 실시하였다(Bartko, 1966). 본 연구에서 실시된 모든 일치도 분석은 R 소프트웨어(버전 3.4.2)을 위한 통계패키지 'irr' (Gamer, 2015)를 사용하여 분석하였다.

결 과

준거별 평가자간 신뢰도

총 아홉 명의 평가자간 개별 준거 일치율의

표 4. 준거별 평가자간 신뢰도

	9명 평가자간 일치율		2명 평가자간 일치율의 평균	
	일치율†	Fleiss 카파‡	일치율	Maxwell RE‡
1. 논리적 일관성	.40	.126	.74	.489
2. 구조화되지 않은 표현	**.00**	**-.113**	**.70**	**.178**
3. 세부내용의 풍부함	.40	.039	.78	.556
4. 맥락상 깊이	.20	-.069	.79	.578
5. 상호작용	**.40**	**.189**	**.79**	**.333**
6. 대화의 인용	**.20**	**.244**	**.63**	**.244**
7. 사건 중 예기치 않은 일 발생	**.00**	**-.001**	**.65**	**.228**
8. 독특한 세부내용	**.00**	**.180**	**.63**	**.178**
9. 부가적인 세부내용	**.00**	**-.049**	**.47**	**-.044**
10. 정확하지만 이해하지 못한 내용	.20	.389	.79	.578
11. 관련된 외적 연합	**.00**	**-.037**	**.56**	**.111**
12. 주관적 심리상태 묘사	.20	.567	.79	.578
13. 가해자의 정신상태 귀인	.80	.196	.90	.800
14. 자발적 수정	**.20**	**.120**	**.58**	**.156**
15. 기억의 부족 시인	.20	.442	.72	.433
16. 자기 진술에 대한 의심제기	.60	.375	.80	.600
17. 자기 비난	1.00	a	1.00	1.000
18. 가해자 용서	1.00	a	1.00	1.000
19. 범죄 특징에 대한 세부내용	.80	-.023	.96	.911

† 일치율이 .70 이상인 경우 높은 수준의 일관성으로 평가됨

‡ 일치계수가 .00 미만(일치도 거의 없음), .000-.200(약간의 일치도 존재), .201-.400(일정수준의 일치도 존재), .401-.600(적당한 수의 일치도), .601-.800(상당한 수준의 일치도), .801-1.000(거의 완벽한 수준의 일치도)(Landis 등, 1977)

a. 분석되지 않음

평균은 .37이었다. 개별 준거에 대한 일치율을 살펴보면, 준거17(자기 비난), 준거18(가해자 용서)의 일치율은 1.00으로 다섯 건 사건에 대해 전원 동일한 평가를 내렸다. 평가자들의 분석 내용을 살펴본 결과 100% 일치율을 보인 두 준거에 대해서는 평가자 전원이 모든 사건에 있어 해당 준거들이 '부재'하는 것으로 판단하였다. 반면, 준거2(구조화되지 않은 표현), 준거7(사건 중 예기치 않은 일 발생), 준거8(독특한 세부내용), 준거9(부가적인 세부내용), 준거11(관련된 외적 연합)의 경우 일치율은 .00으로 나타났다. 즉, 해당 준거에 대해서는 다섯 건의 사건 중 전원이 일치된 판단을 한 경우는 존재하지 않았다.

CBCA 개별 준거에 대한 아홉 명의 평가자간 Fleiss 카파계수의 범위는 -.04에서 1.00 사이로 나타났다. 아홉 명의 평가자간 일치계수는 총 19개의 준거 중 13개 준거에서 .20 이하로 평가자간 신뢰도가 거의 없거나, 또는 약간의 일치

도만이 존재하는 것으로 나타났다. 준거6(대화의 인용), 준거10(정확하지만 이해하지 못한 내용), 준거16(자기 진술에 대한 의심제기)의 경우 일정 수준의 신뢰도가 존재하였다. 반면 준거12(주관적 심리상태 묘사)와 준거15(기억의 부족 시인) 경우 '적절한 수준'의 신뢰도를 보였다. 종합해 보면, 총 19개의 준거 중 15개의 준거의 일치도 계수는 .400 미만으로 평가자간 일관된 평가를 하고 있다고 판단하기 어려운 것으로 나타났다.

반면, 두 명의 평가자간 개별 준거에 대한 평균 일치율의 평균은 .73으로 높은 수준의 일치율 보였다. 준거13(가해자의 정신상태 귀인), 준거17(자기 비난), 준거18(가해자 용서), 준거19(범죄 특징에 대한 세부내용)의 경우 일치율은 .900 이상으로 평가자간 매우 높은 수준의 일관성을 보였다. 반면 준거9(부가적인 세부내용)의 경우 .47로 가장 낮은 일치율을 보였으며, 준거11(관련된 외적 연합), 준거14(자발적 수정)의 경우 각각 .56과 .58로 절반에서 조금 높은 수준이었다.

추가적인 Maxwell RE 계수 분석에서, 두 명의 평가자간 평균 RE 계수는 -.044에서 1.00으로 나타났다. 준거1(논리적 일관성), 준거3(세부내용의 풍부함), 준거4(맥락상 깊이), 준거10(정확하지만 이해하지 못한 세부내용), 준거14(자발적 수정), 준거16(자기 진술에 대한 의심 제기)은 RE 계수가 .400에서 .600사이로 적당한 수준의 일치도가 존재하였으며, 준거13(가해자의 정신상태 귀인), 준거17(자기 비난), 준거18(가해자 용서), 준거19(범죄 특징에 대한 세부내용)의 경우 높은 수준의 일치도를 보이고 있다. 반면, 준거2(구조화되지 않은 표현), 준거8(독특한 세부내용), 준거9(부가적인 세부내용), 준거11(관련된 외적 연합), 준거14(자발적 수정)의 경우 RE 계수가 .200 미만으로 평가자간 약간의 일치도만이 존재하였다. 준거5(상호작용), 준거6(대화의 인용), 준거7(사건 중 예기치 않은 일 발생)의 경우 RE 계수가 .200에서 .400 사이로 일정수준의 일치도가 확인되었다. 종합하면, 두 명의 평가자간 일치계수를 고려할 때 19개의 준거 중 11개 준거(준거1, 3, 4, 10, 12, 13, 15, 16, 17, 18, 19)는 일치도 계수가 .400 이상으로 적절한 수준 이상의 신뢰도가 존재하였으나, 나머지 8개의 준거는 평가자간 동일한 판단을 한다고 보기는 어려운 것으로 나타났다.

진술신빙성 판단 시 평가자간 신뢰도

대부분의 평가자들은 사건1, 사건3, 사건4의 아동 진술은 신빙성이 있는 것으로, 사건2와 사건5의 경우 신빙성이 없는 것으로 판단하였다. 신빙성이 존재하는 것으로 판단한 사건1, 사건3, 사건4의 경우 순서대로 평균 7.56개(표준편차 1.94), 11.89개(표준편차 1.76), 9.78개(표준편차 1.48)의 준거가 존재하였으며, 신빙성이 없다고 판단된 사건2와 사건5의 경우 준거의 존재 개수 평균은 6.44개(표준편차 2.128)와 5.00개(표준편차 1.73)로 나타났다.

진술신빙성 유무 판단 시 두 명의 평가자간 평균 일치율은 .87로 높은 수준이었다. 사건1과 사건2의 경우 각 .78이었으며, 사건3과 사건5의 경우 .89, 사건4은 전원 일치율을 보였다. 진술신빙성 판단의 일치계수 또한 아홉 명 평가자간 일치율과 두 명 평가자간 일치율 평균으로 구분하였다. 아홉 명 및 두 명의 평가자간 일치계수는 모두 '적절한 수준'인 것으로 나타났다(Fleiss 카파계수 = .491; Maxwell RE 계수 = .511). 개별 사건에 대한 아동 진술의 신빙성을 백분율로 판단하였을 때, 대부분의 전문가들이 신빙성이 없다고 판단한 사건2와 사건5의 경우 신빙성 평균은 각각 43.33%, 39.44%이었다. 반면, 대부분의 전문가들이 신빙성 있는 것으로 판단한 사건1, 사건3, 사건4의 경우 신빙성 평균은 순서대로 62.22%, 81.89%, 76.67%로 나타났다. 급내상관계수를 분석한 결과 평가자간 신뢰도는 유의미하였다(ICC = .936, $p < .001$)(표 5 참고).

비록 본 연구의 참여한 평가자 모두 경찰청 소속 진술분석 전문가로 활동하고 있으나 전문가들 사이에 경력 및 의견서 작성 경험에서 차

표 5. 진술신빙성 판단 시 평가자간 일치도

	CBCA 준거 개수			신빙성 판단 (유/무)				신빙성 판단 (%)		
	최소값	최대값	평균 (표준편차)	신빙성	일치율	Fleiss 카파	Maxwell RE	최소값	최대값	평균 (표준편차)
사건1	4	10	7.56 (1.94)	유 (7/9)	.78			30	80	62.22 (16.60)
사건2	4	10	6.44 (2.13)	무 (7/9)	.78			20	80	43.33 (18.71)
사건3	9	15	11.89 (1.76)	유 (8/9)	.89	.491	.511	50	99	81.89 (15.22)
사건4	7	12	9.78 (1.48)	유 (9/9)	1.00			60	90	76.67 (10.00)
사건5	2	8	5.00 (1.73)	무 (8/9)	.89			10	60	39.44 (16.29)
전체				일치율 평균 = .87				ICC = .936, p < .001		

표 6. 평가자간 상관분석

	1	2	3	4	5	6	7	8	9
평가자1	1								
평가자2	.469**	1							
평가자3	.563**	.579**	1						
평가자4	.540**	.628**	.549**	1					
평가자5	.407**	.544**	.521**	.431**	1				
평가자6	.520**	.549**	.505**	.527**	.484**	1			
평가자7	.463**	.474**	.457**	.671**	.452**	.546**	1		
평가자8	.442**	.538**	.518**	.614**	.610**	.381**	.530**	1	
평가자9	.421**	.613**	.606**	.556**	.503**	.399**	.479**	.454**	1

** 은 .001 수준에서 유의미한 상관이 있음을 의미함

이가 있었다. 따라서 특정 전문가의 판단이 다른 전문가들의 판단과 차이가 있는지 확인하였다. 아홉 명의 평가자들의 CBCA 개별 준거 평가에 대한 상관분석 결과 모든 평가자들의 판단은 다른 평가자들의 판단과 유의미한 정적 상관관계가 있는 것으로 나타났다(표 6 참고).

진술신빙성 판단 기준

추가적으로 진술신빙성 판단 시 가장 중요하게 고려하는 요소에 대해 전문가들의 자유로운 의견을 확인하였다. 전체 아홉 명 중 여덟 명의 전문가들은 '구체적이고 상세한 진술묘사'가 진술의 신빙성 판단 시 가장 중요하게 고려하는 요인으로 응답하였다. 또한 'CBCA 평가 결과', '피해 사건 발고 과정 및 피해자 정보', '조사면담 전, 중, 후 피해자 진술태도', '객관적인 증거 존재 여부', '사건 조사 진행과정', '진술내용의 일관성', '발달 연령에 따른 피해자 진술양상', '보호자의 반응과 왜곡 가능성', '진술 내용의 현실성' 등을 고려한다고 응답하였다.

신빙성 판단 시 가장 어려운 점은 '특수한 피

해자(예: 지적 장애인, 매우 어린 아동)에 대한 전문 지식 부족', '개방형 질문 부족으로 진술내용이 제한적인 경우', '사건 발생 후 시간이 경과함에 따른 기억의 부족', 'CBCA 준거들이 발생하였으나, 평가의 질이 낮은 경우', '제한된 자료만으로 신빙성 판단을 해야 하는 상황', '추가면담, 보호자 면담 등과 같은 추가적인 정보 탐색의 제한', '말을 번복하거나 진술이 일관적이지 않은 경우' 등이었다.

진술분석관 활동 시 어려움 및 개선점

진술분석관으로 활동하면서 어려운 점에 대해서는 '경찰 수사관과의 민감한 관계', '수사관들의 특정 피해자(예: 지적 장애인)에 대한 이해가 낮아 협조를 구하기 어려움', '업무의 중요성에 비해 적은 수당', '법원 출석으로 인한 심리적 부담', '긴 대기시간, 갑작스러운 호출 등 불규칙한 일정', '부적절한 질문 방식으로 인한 진술분석 어려움'을 언급하였다. 진술분석관으로 근무하면서 개선되어야 할 업무 환경 및 처우와 관련해서는 '보수 인상', '상주 배치', '평가를 위한 자료(예: 녹화 영상 등)에 대한 자유로운 접근', '예약 취소 및 2차 면담 등에 따른 비용처리'등을 언급하였다.

전문가 업무와 관련한 교육에 대해서는 절반 정도의 전문가들이 '의견서 내용에 대한 지속적인 슈퍼비전과 피드백'을 희망하였으며, 동료 사례 연습 및 슈퍼비전이 필요하다고 응답하였다. 또한 특정 사례별 전문 교육이 필요하며, 전문가에 대해서 정기적 교육 및 관리가 제공되기를 희망하였다.

논 의

본 연구는 경찰청 소속 진술분석 전문가들을 대상으로 CBCA 분석에 대한 평가자간 신뢰도를 확인하였다. 연구 결과, 개별 준거에 대한 두 명의 평가자간 평균 일치도는 대략 '일정 수준'에서 '적절한 수준'을 보이고 있다. 일반적으로 카파계수가 .40 이상(적절한 수준)일 때 '신뢰도가 있다'라고 판단하는 것을 고려할 때(Fleiss, 1981), 본 연구에서는 평가자간 '신뢰도가 있다'라고 볼 수 있는 준거는 총 19개의 준거 중 11개로 나타났다. 반면 준거2(구조화되지 않은 표현), 준거5(상호작용), 준거6(대화의 인용), 준거7(사건 중 예기치 않은 일 발생), 준거8(독특한 세부내용), 준거9(부가적인 세부내용), 준거11(관련된 외적 연합), 준거14(자발적 수정)의 경우 평가자간 일치도가 거의 없거나 낮은 수준으로, 동일한 사건에 대해 전문가들이 동일한 CBCA 분석 결과를 도출한다고 보기는 어려웠다. 이와 같은 결과는 기존 연구 결과와도 일정 부분 유사한 것으로 보인다. 예를 들어, 본 연구에서는 준거2, 준거8, 준거9는 비교적 낮은 평가자간 신뢰도를 보인 반면, 준거16, 준거17, 준거18의 경우 신뢰도가 높은 것을 나타났는데, 이는 기존의 영어를 사용하는 문화권에서 실시된 연구 결과와 유사하다(Akehurst et al., 2011; Anson et al., 1993; Gödert et al., 2005; Horowitz et al., 1997; Niveau et al., 2015).

연구 결과 신뢰도가 낮은 준거들의 경우, 다른 준거에 비해 판단하는데 있어서 다소 모호한 부분이 존재하는 것으로 보인다. 특히, 준거2(구조화되지 않은 표현)는 진술이 시간적 순서에 따르지 않고 산발적으로 나타나는 것을 확인하는 준거이다. 일반적으로 강간 피해자의 경우, 피해사실에 대해 시간 순서대로 구조화된 방식으로 진술하지 않는 경우가 빈번이 발생한다(Winkkel, Vrij, Koppelaar, & Van der Steen, 1991). 따라서 진술분석에서는 진술이 시간의 순서를 따르기 보다는 의식의 흐름에 따라 자연스러운 방식으로 나타날 때 진실의 가능성이 높다고 판단한다. 다만 '산발적'이라는 것의 의미가 모호하여 진술의 비일관성으로 잘못 인식되는 경우가 존재할 수 있으며, 이러한 경우 준거에 대한 코딩은 일관성이 떨어질 수 있을 것으로 보인

다. 준거8(독특한 세부내용)과 준거9(부가적인 세부내용)의 경우에도 진술 내에 일반적이지 않은 독특한 진술(준거8) 또는 본 사건을 이해하는데 있어서 생략되어도 무관한 부가적인 진술(준거9)이 존재했는지를 평가한다. 여기에서 '독특한' 또는 '부가적인'과 같은 개념 역시 다소 모호하며 정확한 판단 기준이 존재하지 않다는 점에서 평가자간 낮은 일치도를 보인 것으로 생각된다.

CBCA 분석에 근거한 진술신빙성에 대한 최종 판단은 평가자간 적절한 수준의 일치도를 보였다. 특히 사건4의 경우 아홉 명 전원이 동일하게 판단하였으며, 사건3과 사건4의 경우 아홉 명 중 한 명만이(89%) 다른 전문가들과 다른 판단을 하는 것으로 나타났다. 사건1과 사건2의 경우 아홉 명 중 두 명의 전문가가 다른 전문가들과 다른 의견을 제시하여 약 78%의 일치율을 보였다.

진술분석 의견서는 법정에서 유무죄 판단에 영향을 미칠 수 있다는 점에서 진술분석관들의 CBCA 분석은 정확해야 하며 또한 전문성을 갖추어야 한다. 본 연구는 일부 준거 활용에 있어서 전문가들 간 일관성이 부족할 수 있다는 가능성이 제기되었다. 다만 최종 판단의 모호성 및 주관성에 대한 기존 비판에도 불구하고(Mazzoni, & Ambrosi, 2003), 본 연구는 CBCA 분석을 기초로 진술의 신빙성 판단 시 평가자간 어느 정도 일관된 결과를 도출할 수 있다는 것을 확인하였다는 점에서 매우 의미 있는 것으로 보인다. 이와 같은 연구 결과를 바탕으로 사법현장에서 진술분석 의견서의 활용 및 진술분석관들의 전문성 제고를 위한 학문적, 정책적 노력은 반드시 이루어져야 할 것이다.

본 연구에는 몇 가지 제한점이 존재한다. 첫째, 본 연구는 아홉 명의 진술분석 전문가들이 참여하였다. 2016년 통계에 따르면 현재 우리나라에는 총 백사 명의 진술분석 전문가들이 활동하고 있다(경찰청, 2017). 이를 고려할 때, 연구에 참여한 전문가는 현재 활동하는 전체 전문가들의 8 ~ 9%에 불과하다. 따라서 본 연구의 결과만으로 전체 진술분석관들의 전문성을 판단하기는 다소 부족한 것으로 보인다. 또한 연구에 참여한 전문가들은 참여하지 않은 전문가들에 비해 자신의 판단에 더 자신감을 가지고 있거나, 또는 더 많은 경력을 가지고 있는 등의 특성을 포함하고 있을 가능성을 배제할 수 없다. 박노섭 등(2013)의 연구에서 이십팔 명의 진술분석 전문가를 대상으로 설문조사를 실시하였는데, 조사 대상자들의 근무 기간은 평균 28.12개월이었으며, 의견서 제출 회수는 10회 미만인 경우가 약 10%, 20 ~ 50회가 30%, 50회 이상이 51%로 나타났다. 비록 기존 연구가 2013년 실시되었으며 이후 4 ~ 5년이 경과되었음을 고려하더라도 기존 연구와 비교하여 본 연구에 참여한 전문가들의 경력과 의견서 제출 회수는 상대적으로 더 높은 것으로 보인다. 이러한 차이는 본 연구의 결과를 일반화 하는데 있어서 제한점이 될 수 있으며, 연구 결과를 해석하는데 있어서 반드시 고려되어야 할 것이다.

두 번째로 본 연구에서는 단지 다섯 건의 사건에 대한 평가자간 일치도를 확인하였다. 신뢰도 분석을 위해 적은 수의 사례만을 포함하는 경우 CBCA 분석 결과는 사건의 난이도 등에 의해 영향을 받을 수 있다. 따라서 본 연구 결과가 실제 평가자간 신뢰도를 과소 혹은 과대평가했을 가능성이 존재한다. 본 연구는 현장에서 근무하는 전문가들을 대상으로 실시하였기 때문에 비용과 시간에 따른 어려움으로 인해 많은 사례를 포함하는데 있어서 현실적인 어려움이 있었다. 이는 본 연구의 제한점이 될 수 있다.

추가적으로 후속 연구에서는 진술분석 전문가들의 판단의 정확성에 대한 평가가 반드시 필요하다. 본 연구에서는 평가자간 얼마나 일관된 결과를 도출할 수 있는지만을 고려하였으며, 전문가들의 CBCA 분석 결과의 정확성 또는 타당성에 대한 분석은 실시하지 않았다. 그러나 높은 수준의 일치도를 보이는 세 개의 준거, 즉, 가해자의 16정신상태 귀인(13번), 자기 비난(17번), 범죄 특징에 대한 세부내용(19번)의 경우 거

의 모든 평가자들은 모든 사건에서 해당 준거가 '부재' 하는 것으로 평가하였으나, 본 연구자가 녹취록을 검토하였을 때, 다섯 건 중 적어도 한 건 이상에서 준거13, 준거17, 준거19가 존재하는 것으로 판단되었다. 이와 동일하게 준거7과 준거8 역시 적어도 한 사건 이상에서 해당 준거들이 존재하였으나, 대부분의 연구자들은 부재한다고 판단하였다. 따라서 평가자간 신뢰도와는 별개로 진술분석 전문가들의 CBCA 평가의 정확성은 재검토 되어야 할 것이다.

진술분석 전문가들의 CBCA 분석에 있어서 평가자간 일치도는 실무적으로도 매우 중요한 의미를 갖는다. 만일 사건을 어떤 전문가에게 배당하는지에 따라 그 결과가 달라진다면 진술분석 결과를 신뢰할 수 없게 된다. 본 연구 결과 최종 신빙성 판단에 있어서 일치율은 높았지만 CBCA 분석에 있어 일부 준거들의 평가자간 일관성은 낮은 것이 확인되었다. 이러한 평가자간 차이는 진술분석 전문가 집단이 동일한 사건에 대한 분석 시에도 분석 결과가 다르게 나타날 수 있음을 시사하는 것으로 전문성을 위협하는 요인이 될 수 있다. 대부분의 진술분석 전문가들은 심리학 또는 그와 관련 분야에서 석사 학위 이상을 받았으며, 임상전문가, 상담전문가 또는 청소년 상담 등 유사 업무에 종사한 경험이 있거나, 이와 관련된 훈련을 받은 자들이다(박노섭 등 2013). 다만 법정임상 또는 진술분석은 기존의 심리학의 영역과는 차이가 있기 때문에, 기존 관련 경력만으로 진술분석 전문가로 활동하는데 있어서는 어려움이 있을 수 있다. 진술분석관들의 자유의견에서도 관련 교육과 슈퍼비전의 부족 및 부재는 문제로 제기되고 있다. 따라서 진술분석 전문가들의 CBCA 활용에 있어서 전문성 제고를 위한 교육 및 슈퍼비전은 진술분석 전문가들의 전문성을 향상하는데 매우 중요한 역할을 할 것으로 기대한다.

추가적으로 진술분석 전문가들의 업무환경의 개선 역시 중요한 부분이다. 최근 진술분석관들이 법정에 출석하여 전문가 증언의 형태로 의견을 제시하는 사례가 증가하고 있으며, 진술분석 의견서는 기소율와 유죄판결율에 영향을 미치는 것으로 나타났다(이수정, 2010). 이를 고려할 때, 진술분석 전문가들의 의견은 매우 높은 수준의 전문성이 요구되는 실정이나, 전문가들이 이에 합당한 대우를 받는지는 여전히 의문이다. 기존 진술분석 전문가들을 대상으로 실시한 설문조사 연구 결과에 따르면(박노섭 등, 2013), 전문가들은 의견서 작성에 따른 수당에 불만을 가지고 있으며 수사관과의 관계에서 부적절한 대우를 받는다고 느끼는 경우가 다수 존재하는 것으로 나타났다. 이와 같은 의견은 본 연구에서도 동일하게 지적되었으며, 추가적으로 2차 면담 시 추가적인 보상을 받지 못하거나 불규칙한 상황에 대한 불만을 표현하기도 하였다. 아동과 장애인의 수사면담 과정에서 2차 피해를 방지하고 수사와 재판의 효율성을 위해 전문가 의견조회 제도가 시작된 지 대략 7년의 시간이 흘렀다. 지금까지 사법현장에서 실무자들은 전문가 의견조회 제도에 대해 긍정적인 평가를 하고 있으나(박종선, 2013), 전문가 의견의 타당도 검증 및 적절한 제도적 방안은 여전히 미비한 것으로 보인다. 아동성폭력 범죄의 심각성 및 사회적인 피해를 고려할 때, 진술분석 전문가의 전문성 제고 및 발전방안에 대한 지속적인 관심과 변화가 필요할 것이다.

참고문헌

경찰청 (2017). 진술분석가 의견 작성 매뉴얼. 경찰청.

고은영, 채규만 (2011). 성폭력 피해 아동의 진술에 대한 준거기반 내용분석의 활용 가능성 연구. 한국범죄심리연구, 7(3), 5-22.

김현정 (2010). CBCA와 RM을 이용한 성폭력 피해 아동이 진술신빙성 평가. 한국심리학회지: 여성, 15(3), 355-379.

박광배, 엄진섭 (1996). 평가자간 일치도를 파악

하기 위한 통계절차들. 한국심리학회 학술대회 자료집, 1996(1), 623-642.

박노섭, 조은경, 이미선 (2013). 성폭력 근절 관련 대책의 개선방안. 경찰청.

박종선 (2013). 전문가 의견조회의 성과와 발전 방안: 판검사 설문조사를 중심으로. 형사법의 신동향 통권 제 41호 86-117.

성태제 (2002). 타당도와 신뢰도. 학지사.

이미선 (2004). 성폭력 피해 아동진술에 대한 준거기반 내용분석의 타당화를 위한 연구. 한림대학교 석사 학위 논문.

이수정 (2010). 성폭력 피해 진술에 대한 신빙성 분석도구들의 타당도 연구. 한국심리학회지: 사회 및 성격, 24(2), 105-116.

이수정 (2011). 아동 및 장애인에 대한 성폭력 사건에서 경찰단계에서의 전문가 참여가 기소율과 유죄판결률에 미치는 영향. 한국범죄학, 5(1), 111-138.

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, 7(1), 3-12.

Anson, D. A., Golding, S. L., & Gully, K. J. (1993). Child sexual abuse allegations: Reliability of criteria-based content analysis. *Law and Human Behavior*, *17*(3), 331-341.

Akehurst, L., Manton, S., & Quandte, S. (2011). Careful calculation or a leap of faith? A field study of the translation of CBCA ratings to final credibility judgements. *Applied Cognitive Psychology*, *25*(2), 236-243.

Bartko, J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, *19*(1), 3-11.

Boychuk, T. (1991). *Criteria-Based Content Analysis of children's statements about sexual abuse: a field based validation study*. Doctoral dissertation. Arizona State University.

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: a historical review and synthesis. *Psychological bulletin, 113(3)*, 403-439.

Craing, R., Scheibe, R., Raskin, D. C., Kircher, J. C., & Dodd, D. H., (1999). Interviewer questions and content analysis of children's statements of sexual abuse. *Applied Developmental Science, 3,* 77-85.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322-328.

Esplin, P. W., Boychuk T., & Raskin, D. C. (1988). A *field validity study of Criteria-based content analysis of children's statements in sexual abuse cases*. Paper presented at the NATO Advanced Study Institute on credibility assessment in Maaratea, Italy, June, 1988.

Fivush, R., Gray, J. T., & Fromhoff, F. A. (1987). Two-year-old talk about the past. *Cognitive development*, *2*(4), 393-409.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*(1), 105-112.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, *48*(1), 157-184.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2015). Package 'irr' http://www.r-project.org

Gödert, H. W., Gamer, M., Rill, H. G., & Vossel, G. (2005). Statement validity assessment: Inter rater reliability of criteria based content analysis in the mock crime paradigm. *Legal and criminological psychology*, *10*(2), 225-245.

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of criteria-based content analysis. *Psychological Assessment, 29*(6), 819-834.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter Lavery, L. (1997). Reliability of criteria based content analysis of child witness statements. *Legal and Criminological Psychology*, *2*(1), 11-21.

Lamb, M. E., Hershkowitz, I., Orbach, Y., & Esplin, P. W. (2011). *Tell me what happened: Structured investigative interviews of child victims and witnesses* (Vol. 56). John Wiley & Sons.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Jershkowitz, I., Orbach, Y., & Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse and Neglect,* 21, 255-264.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Maxwell, A. E. (1997). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry,* 130, 79-83.

Mazzoni, G., & Ambrosio, K. (2003). L'analisi del resoconto testimoniale in bambini: impiego del metodo di analisi del contenuto CBCA in bambini di 7 anni. Disponibile online: http://www. psicologiagiuridica. com/numero, 20006.

Niveau, G., Lacasa, M. J., Berclaz, M., & Germond, M. (2015). Inter rater reliability of criteria based content analysis of children's statements of abuse. *Journal of forensic sciences*, *60*(5), 1247-1252.

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and human behavior, 40(4),* 440-457.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, *42*(7), 725-728.

Steller, M. (1989). Commentary: Rehabilitation of the Child Witness. In J. Doris (Ed.) *The suggestibility of children's recollections.* Washington, DC: American Psychological Association.

Steller, M., & Koehnken, G. (1989). *Criteria-based content analysis.* In D. C. Raskin (Ed.), Psychological methods in criminal investigation and evidence. New York, NJ: Springer-Verlag.

Vrij, A. (2005). Criteria-Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, *11*(1), 3-41.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities.* (2nd ed.) West Sussex, England: John Wiley & Sons.

Whipple, G. M. (1912). Psychology of testimony and report. *Psychological Bulletin,* 9(7), 264-269.

Winkel, F. W., Vrij, A., Koppelaar, L., & Van der Steen, J. (1991). Reducing secondary victimzation risks and skilled police intervention: Enhancing the quality of police rape victim encounters through training programmes. *Journal of Police and Criminal Psychology, 7,* 2-411.

1 차원고접수 : 2018. 01. 15.
수정원고접수 : 2018. 02. 28.
최종게재결정 : 2018. 05. 21.

韓國心理學會誌 : 社會 및 性格
Korean Journal of Social and Personality Psychology
2018, Vol. 32, No. 2, 67-83

Inter-rater reliability in assessing the credibility of allegedly sexually abused child victims' statements: Focusing on the CBCA expert group

Mi Sun Yi

Dongyang University

Criteria-Based Contents Analysis(CBCA) is a statement-analysis technique that determines the veracity of sexually abused child victims' statements. It consists of the 19 criteria that are most likely to exist in statements of children who are telling the truth. This study examined the inter-rater reliability of nine CBCA experts working for the Korean Police Agency. The experts evaluated the existing 19 CBCA criteria with five child sexual abuse cases. They were also asked to make a final decision as to whether the children's statements were reliable or not. The results showed that 11 out of the 19 criteria indicated adequate inter-rater reliabilities, with coefficients of criterion 2(Unstructured production), 8(Unusual details), 9(Superfluous details), 11(Related associations), and 14(Spontaneous corrections) being low. The agreement coefficient of the overall judgments of the reliability of the children's statements was .511, showing an adequate level of reliability. The implications of these findings for the use of CBCA reports in making credibility decisions in the forensic contexts are discussed.

Key words : Statement Validity Assessment, Criteria-Based Content Analysis, expert testimony, inter-rater reliability