

실무 자백 사건들을 활용한 검찰 심리생리검사의 타당성 연구

김 시 온* 김 미 영 김 석 찬 이 상 현
대검찰청, 대검찰청
연세대학교 인지과학협동과정

2016년부터 2017년까지 2년간 검찰에서 실시한 심리생리검사 사건들 중 수사 및 재판 과정에서 피고인이 자백한 91건을 확인하여 피고인 자백 내용과 심리생리검사 결과와의 일치 여부를 확인하였다. 검찰에서 시행 중인 의사 결정 과정에 따른 정확성을 계산해 보았을 때, 모든 단계에서 90%가 넘는 정확률을 보였고, 이는 컴퓨터 자동채점 및 분석 프로그램이 기록한 정확성이나 미국 폴리그래프 협회에서 유사한 기법을 통해 검사하였을 때 보고했던 정확성보다 높은 수준이었다. 신호탐지론을 통해 민감도(d')와 반응 기준(c)을 계산한 결과, 담당 검사관의 최초의견부터 대검찰청 소속 분석관 검토의견 및 최종의견에 이르기까지 3을 넘는 민감도를 확인할 수 있었고, 이는 비교대상으로 삼은 컴퓨터 자동채점 프로그램이나, 이전 연구 결과에서 보고한 값보다 더 큰 결과였다. 반응 기준의 경우, 대검소속 검사관들은 상대적으로 보수적인 판단 기준을 가지고 의사 결정을 수행하였으나, 컴퓨터 자동채점 프로그램의 경우 보다 더 관대한 판단 기준으로 의사 결정을 하는 것을 확인할 수 있었다. 또한 검사관들의 경우, 판단불능률은 더 높았으나 컴퓨터 자동채점이 보여준 높은 오경보율 등의 오류가 더 작았던 것을 확인할 수 있었다. 본 연구는 비록 자백 연구의 특성 상 실제보다 정확성이 과장되어 나타날 가능성이 존재할 지라도, 실제 사례들을 통해 검찰에서 수행하는 심리생리검사의 성능 및 특성을 파악할 수 있는 실증적인 자료를 제공한다는 데 의의를 가진다.

주요어 : 심리생리검사, 신호탐지론, 타당성, 민감도, 반응 기준

* 교신저자: 김시온, 서울시 서초구 반포대로 157, 대검찰청 국가디지털포렌식센터(NDFC) 105호

E-mail: htmwys@spo.co.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. Copyright ©2024, The Korean Association of Psychology and Law

심리생리검사(Psychophysiological Detection of Deception: PDD)는 사건 관련된 질문에 대한 피검사자의 생리적 반응의 변화를 측정 및 분석함으로써 피검사자의 진술의 신빙성 및 범죄 연루 여부를 평가하는 검사 방법이다 (Honts, 1994). 사건 관계자들의 진술은 현대 형사소송절차에서 실제적 진실을 알려주는 중요한 가치를 지닐 뿐만 아니라(양동철, 2013; 김석찬 외, 2015), 물리적 증거가 부족한 사건의 경우 사실 인정에 큰 영향을 미치기 때문에 그 가치가 더욱 더 증가한다(방철, 2023). 심리생리검사는 이러한 진술의 신빙성을 판단할 수 있는 도구 중의 하나로써, 검찰에서만 매년 2천 건 넘게 실시될 정도로(방철, 2023), 검찰, 경찰, 해경, 군 등 다양한 국가 기관에서 널리 활용되고 있다. 다만 이렇게 널리 사용되는 심리생리검사의 활용성과는 무관하게, 심리생리검사의 정확률 및 타당성에 대해서는 다양한 의견들이 있어 왔다. 예를 들어, 가장 널리 사용되는 기법인 비교질문기법(Comparison Question Technique: CQT)을 사용한 심리생리검사의 정확률을 추정한 연구들은 그 결과가 우연에 비슷한 수준부터 거의 100%에 가까운 수준까지 다양하다(Grubin, 2010). 미국 국립연구위원회(National Research Council)는 이 중 신뢰할 수 있는 연구결과들을 바탕으로 검토하였을 때 정확률이 81%~91% 사이에 있을 것이라 결론 내렸다(NRC, 2003). 한편 미국 폴리그래프 협회(American Polygraph Association)는 심리생리검사의 정확률에 대한 메타분석을 실시하였는데, 형사사건에서 주로 사용되는, 질문순서 등이 포함된 검사 기법 중 하나인 You-Phase 기법으로 질문을 구성하고, 호흡, 혈압, 피부전도 반응 등 다양한 생리적 반응을 수치화하는 차트 채점 기법 중 하나로써 널리

사용되는 경험적 채점 기법(Empirical Scoring System: ESS, Nelson, 2011)을 통해 채점했을 때 추정되는 정확률의 평균이 .904($sd = .032$)이며, 판단불능률은 .192($sd = .033$)로 보고하였다. 이와 같은 추정된 정확률에 대해서 일부는 10% 가까운 오류율을 지적하며 심리생리검사가 수사의 도구로 활용된다거나 재판의 증거로써 사용하기 부적당하다고 판단하고 있다.

하지만 진술의 신빙성 판단과 관련된 기존 연구들은 일반인들이 실험 상황에서 다른 이의 거짓말을 탐지해 낼 수 있는 확률이 우연보다 조금 나은 수준에 지나지 않으며, 강한 진실 편향을 보이고 있음을 반복해서 보고하고 있다(Bonds & DePaulo, 2006; DePaulo & Rosenthal, 1979; Vrij, 2000; Vrij, 2008; Zukerman et al., 1982). 피검사자의 진술을 보고 듣는 것만으로 거짓말을 올바르게 판단할 수 있는 능력이 우연보다 조금 나은 수준인 것은 일반인 뿐만 아니라 오랜 기간 수사경험이 있는 전문가들도 크게 다르지 않다는 연구 결과들 또한 반복해서 검증되고 있다(Bull, 1989; Kassin & Fong, 1999; Vrij, 1994). 이런 측면을 고려하면, 심리생리검사는 진실 혹은 거짓을 판단하는데 있어 다소 부족할 수 있는 인간의 의사결정 능력을 비약적으로 향상시켜줄 수 있는 도구로써 수사 및 법적 의사결정에서 유용하게 활용될 수 있다.

그렇다면 심리생리검사는 얼마나 정확하게 진실과 거짓을 구별해 낼 수 있을까? 기존 연구들은 주로 두 가지 방법을 활용하여 심리생리검사의 정확률을 추정해왔다(Pollina et al., 2004). 첫 번째는 모의범죄를 활용한 실험실 연구들이다(Kircher & Raskin, 1988; Kircher et al., 1988). 실험실 연구들은 무엇보다 대상자의

실체적 진실(유죄 혹은 무죄 여부)을 연구자가 명확하게 알 수 있다는 장점으로 널리 사용되어 왔다. 다만, 모의 범죄의 특성 상 피검사자에게 연구 윤리에 위반되는 행위를 시킬 수 없기 때문에, 연구에 사용할 수 있는 범죄의 종류가 절도와 같은 경범죄로 한정될 뿐 아니라, 동기 측면에서도 실사 거짓말이 발각된다고 할지라도 피검사자에게 돌아올 수 있는 벌칙이 실제 범죄자와 비교할 수 없기 때문에 생태학적 타당성에 있어서도 부족하다는 점이 자주 지적되어 왔다. 두 번째는 경찰, 검찰 및 다양한 기관에서 실제 실시된 사건들을 토대로 정확성을 유추하는 현장연구이다(김석찬 등, 2015; Elaad, 1990; Ginton et al., 1982, Mangan et al., 2008). 이 방법의 경우, 살인, 강간, 문서위조, 사기 등 다양한 사건들에 대한 검사 자료들이 연구 대상이 되고, 실제 피의자 및 피해자들의 반응이 분석에 활용되었다는 데서 높은 생태학적 타당성을 가지고 있다고 할 수 있다. 하지만 수사 과정에서의 의사결정이 실체적 진실을 얼마나 반영하고 있는지에 대해서 명확한 결론을 얻을 수 없다는 문제점이 있다. 따라서 현장 연구들은 수사 및 검사 과정 중의 자백(Mangan et al., 2008)이나 검찰 처분 일치도 및 판결 처분 일치도(김석찬과 동료들, 2015) 등의 자료들을 실체적 진실로 가정하여 심리생리검사의 정확성을 추정하고 있다. 이 중 가장 정보가가 높은 것은 자백으로써 일부러 자신의 죄를 인정하면서 벌을 받고 싶어 하는 경우는 극히 드물기 때문에, 인간의 불완전할 수 있는 판단이 개입될 수 있는 검찰처분 혹은 판결일치도(Garrett, 2011)보다 실체적 진실에 더 가깝다고 할 수 있을 것이다.

다만 최근 연구들은 거짓말 탐지 연구에서

정확률만을 사용하는 것이 부적당할 수 있다는 점을 지적하고 있다(Zloteanu & Vuorre, 2024). 정확률이라는 개념이 가지고 있는 가장 커다란 문제점은 수행에 있어 검사도구의 구별 능력(Discriminability)과 편향(Bias)에 의한 결과를 구별할 수 없다는 점이다(Wagner, 1993). 즉, 검사를 통한 옳은 반응이 검사 도구가 거짓 집단과 진실 집단을 세밀하게 구별해 낸 결과물인지, 아니면 애매한 자극을 진실 혹은 거짓으로 판단하는 개인적 경향에 따른 우연의 결과물인지를 구별할 수 없다. 만약 편향이 의사결정에 있어 커다란 영향을 미치는 상황이라면 분석 대상이 되는 집단의 분포(거짓말을 하는 사람과 진실을 말하는 사람의 비율)에 따라 동일한 도구로 검사했음에도 다른 정확률을 얻을 수도 있다. 이를 구별하기 위한 방법으로 제시되는 것은 신호탐지론 분석을 통해 올바르게 구별한 결과물과 편향에 의해 우연히 맞은 결과물들을 구별하는 것이다. 기존 정확률에서는 단지 옳음과 그름, 두 가지 가능성만 생각한다면 신호탐지론을 이용한 심리생리검사 결과 분석에서는 실체적 진실이 진실과 거짓일 가능성, 그리고 심리생리검사 결과가 진실 혹은 거짓일 2x2 상황을 가정한다. 그 결과, 실체적 진실이 거짓인 사건에 대해서 실제 거짓 결과를 얻을 적중(Hit), 실체적 진실이 거짓인 사건에 대해서 진실 결과를 얻을 누락(Miss, 2종 오류), 실체적 진실이 진실인 사건에 대해서 거짓 결과를 얻을 오경보(False Alarm, 1종 오류), 그리고 실체적 진실이 진실인 사건에 대해서 진실 결과를 얻을 정확 기각(Correct Rejection)으로 사건들을 분류하게 된다. 이후 이 값들을 토대로 얼마나 잘 서로 다른 자극들을 구별해 내는지를 의미하는 민감도(d')와 애매한 자극을 구별하는 기준을 의

미하는 반응 기준(d')을 계산함으로써 특정 도구의 구별 능력과 편향을 확인할 수 있다. 따라서 심리생리검사 기법 역시 신호탐지론 분석을 통해 민감도와 반응 기준을 계산함으로써 검사가 얼마나 적중과 정확기각을 잘 해낼 수 있는지, 또한 검사를 통해 발생하는 오류의 특징이 어떤지에 대한 추가정보를 획득할 수 있다.

매해 많은 수의 검사를 실시하고 있음에도 불구하고 검찰에서 실시되는 심리생리검사의 성능을 직접적으로 유추할 만한 자료는 많지 않다. 기존 검찰 심리분석실에서 실시한 연구들로 모의범죄를 통한 실험연구(정재영 등, 2010) 및 기소 및 판결 일치도 연구(김석찬 등, 2015) 등이 있었으나, 이는 모의 범죄를 통해 얻어진 자료들을 사용하거나, 실제적 진실 대신 검사 및 판사의 판결에 대한 일치도를 기준으로 분석한 결과라는 데서 그 한계가 있었다. 또 다른 방법으로 다른 유사 연구들의 결과를 통해 정확률을 유추하는 방법을 생각해 볼 수 있는데, 예를 들면 미국 폴리그래프 협회에서 발표한, 검찰에서 사용하는 기법과 유사한 기법('You-Phase 기법')으로 진행한 연구들에 대한 메타분석 결과(Gordon, 2011) 등을 통해 추정해 보는 방법이다. 하지만 해당 연구는 다른 문화권에서 얻어진 결과이고, 검사 진행 방법에 있어서도 사소한 차이들이 존재할 가능성이 있기 때문에 메타분석 결과들이 실제 검찰 심리생리검사의 성능을 잘 반영할 수 있을지 판단하기 어렵다. 따라서 본 연구에서는 실제적 진실을 가정할 수 있는 방법 중 가장 정보가가 높은 사례들인 자백 사건들을 수집하여 연구를 진행하였다. 구체적으로 2016년 및 2017년 검찰에서 수행한 6708건 11413명의 데이터 중 판결문 상 피고인의 자

백이 확인된 91건에 대하여 이를 실제적 진실로 가정하고 심리생리검사 결과와 비교함으로써 검찰 심리생리검사 시스템이 얼마나 진실과 거짓을 올바르게 분류할 수 있는지를 확인해 보고자 하였다. 또한 이전 연구들이 보여주었던 정확률 및 판단불능률 외에 신호탐지론 분석을 수행하여 민감도와 반응 기준 또한 확인해 보고자 하였다. 비교 대상으로서, 검찰 심리생리검사 결과 외에 미국에서 널리 사용되는 자동채점 프로그램인 OSS-3(Objective Scoring System: OSS-3)과 앞서 서술한 Gordon (2011)의 메타분석 결과 또한 함께 확인해 보았다. 마지막으로 검찰에서 정확성 및 신뢰성 향상을 목적으로 담당검사관 이외 다른 검사관들의 독립적인 판단을 참고하여 최종 결과를 도출하는 의사결정 과정(품질관리 시스템)이 검사 결과에 어떠한 영향을 미치는지도 함께 알아보려고 하였다.

방 법

자료

검찰에서 2016과 2017년 동안 대검찰청을 포함하여 총 24개청에서 실시한 6708건 11413명의 심리생리검사 결과를 대상으로 판결문을 확인하였고, 심리생리검사 시 자백하지 않았으나 판결문 상 재판과정에서 자백이 확인된 사건들(예: '피고인이 이 사건 범행을 인정하고 반성하고 있는 점', '피고인은 자백하며 반성하고 있다' 등)를 수집하였다. 그 결과 총 139건의 자료를 확인할 수 있었다. 이 139건 중 판결문 상 자백 사안과 검사 사안이 불일치한 경우(8건), 판결문 상 피고인과 검사 대

상이 불일치한 경우(2건), 심한 잡음(Artifacts)으로 인해 컴퓨터를 이용한 자동 채점이 어려운 경우(12건), 혈압 측정 상태가 불량한 경우(16건), 피부전도반응 측정 상태가 불량한 경우(5건), 시각장애인이 피검사자였던 경우(2건), 관련 질문으로 부적당한 질문을 사용한 경우(3건) 등 총 48건을 제외하였다. 따라서 최종 91건이 분석에 활용되었다.

차트 채점

분석에 활용한 자료들은 모두 Lafayette사의 LX-4000 및 LX-5000을 이용해서 측정되었으며, 단일 사안에 대한 형사사건에서 널리 사용되는 You-phase 질문구성의 원리에 따른 질문 배열을 가지고 있었다(Nelson, 2012; Nelson et al., 2016). 관련 질문의 형태가 ‘당신은 무엇을 어찌 했습니까?’의 형태를 가지고 있어서 ‘You-Phase’라 불리는 해당 기법은 일반적으로 9개의 질문으로 구성되어 있는데, 이 중 채점과 관련된 질문은 총 5개로, 표 1과 같이 2개의 관련 질문과 3개의 비교 질문을 가지고 있는 형태로 구성되어 있다. 관련 질문은 사건과 직접적으로 관련된 질문으로 같은 사안에 대해 다르되 어휘만 약간 수정한 2개의 질문이었고, 비교 질문은 관련 질문에 의한 생리

적인 반응과 비교하기 위한 질문으로 본 사건과는 관련 없으나 비슷한 유형의 범죄 상황에 대하여 피검사자가 거짓말을 할 법한 질문들로 구성되었다. 검사관은 하나의 관련 질문과 이를 둘러싼 두 개의 비교 질문들에 대한 호흡, 혈압, 피부전도반응을 시각적으로 탐색하고 비교하여, 관련 질문에 대한 반응이 더 클 경우에 -1~3점의 점수를, 비교질문에 대한 반응이 더 클 경우에 +1~+3점까지의 점수를 부여하며, 둘 간에 차이가 없을 경우 0점을 부여하였다. 검사는 최소 3회 반복 실시되었고 그 때 얻어진 점수들의 총합이 진실과 거짓을 판별하는 의사결정에 사용되었다. 진실과 거짓, 판단불능을 결정하는 절단점(cut-off score)의 경우, 검찰청에서 자체적으로 사용하고 있는 기준을 사용하였는데, -12점 이하면 ‘거짓’, -11점~3점 사이면 ‘판단불능’, -3점 이하면 ‘진실’로 판정하였다. 각 사건들에 대한 채점 및 의사결정은 2016년 및 2017년 당시 대검찰청을 포함한 25개청에서 근무하던 심리생리검사관 자격을 갖춘 검사관들에 의해 실시되었다.

컴퓨터 자동 채점의 경우, Lafayette사에서 제공하고 있는 LxSoftware 11.8.5에 포함된 OSS (Objective Scoring System)-Version 3(이하 OSS-3) 자동채점 및 결과분석 프로그램을 연구에 사

표 1. You-phase 기법에서의 비교 질문 및 관련 질문 제시 형태

형태	질문 내용 예시
비교 질문	이 사건과 상관없이, 비난받을 만한 거짓말을 한 적 있습니까?
관련 질문	본 건 당시, 당신은 xxx의 얼굴 부위를 때렸습니까?
비교 질문	이 사건과 무관하게, 화가 나서 다른 사람을 때린 적이 있습니까?
관련 질문	그날 그 장소에서 당신은 xxx의 얼굴 부위를 때렸습니까?
비교 질문	이 사건과 관련없이, 감정을 주체하지 못해 타인을 때렸던 적이 있습니까?

용하였다(Krapohl & McManus, 1999; Krapohl, 2002). OSS는 ‘커셔 특징(Kircher Features)’이라고 불리는 심리생리검사의 세 가지 채점 원리(호흡선의 길이, 피부 전기 반응의 위상 진폭 증가, 심혈관 반응의 위상 진폭 증가)에 기반을 두고 있는 기법으로(Dutton, 2000), 각 생리적 반응의 차이를 비율로 변환하여 비교하는 방식이다. 계산된 비율은 -3에서 +3까지의 정수값을 가지는 균일한 7분위 분포로 변환되는데, 이후 총점은 합산되고 가우시안 신호 탐지모델(Wickens, 1991)에 적용된다. 생리적 반응들의 비율의 차를 정밀하게 계산하기 위해서는 컴퓨터 알고리즘을 통한 각 생리적 반응들의 객관적 측정이 필요하므로, OSS는 대표적인 컴퓨터 알고리즘 채점 프로그램으로 널리 사용되게 되었다(Nelson et al., 2008). OSS를 기반으로 한 OSS-3는 가장 널리 사용되는 자동채점 및 결과분석 프로그램 중 하나로, 스크리닝 검사를 위한 DLST 기법을 사용한 타당성 연구에서 일반적으로 널리 사용되는 수동 채점 방식인 ESS 검사 기법과 비교하였을 때 오히려 더 나은 결과를 보여준 바 있다(Nelson et al., 2012). OSS-3를 사용하기 위해 검토 과정에서 눈에 띄는 잡음들은 분석에서 제외시키는 과정을 거쳤고, 측정된 자료 전반적으로 심한 잡음이 있거나 잡음의 수가 너무 많아 분석에 필요한 만큼의 충분한 데이터가 확보될 수 없다고 판단되는 자료들은 분석에서 제외하였다.

분석 방법

본 연구의 자백의 주체는 피고인이었기 때문에, 심리생리검사에서 검사한 사안에 대해서 피고인은 거짓된 대답을 하였고, 고소인은

진실로 대답했다는 사실을 확인할 수 있었다. 따라서 표 3과 같이, 피고인에 대한 심리생리 검사 결과가 거짓으로 분류되었을 때는 자백과 일치하는 결과로, 진실로 분류되었을 때는 자백과 불일치하는 결과로 분석하였다. 고소인의 경우, 피고인과는 반대로 심리생리검사 결과가 거짓으로 분류되었을 때는 자백과 불일치하는 결과로, 심리생리검사 결과가 진실로 분류되었을 때는 일치하는 결과로 분류하였다. 또한, 일반적으로 심리생리검사는 거짓말을 탐지하기 위한 검사이기 때문에, 실제적 진실이 거짓인 경우 이를 제대로 탐지했으면 적중, 그렇지 않으면 누락, 실제적 진실이 진실인 경우, 이를 제대로 기각한 것을 정확기각, 이를 기각하지 못하고 거짓인 것으로 잘못 반응한 경우를 오경보로 분류하였다.

검찰에서는 심리생리검사를 실시한 후, 검사관의 의견만으로 결과를 통보하는 것이 아니라 대검찰청에서 근무하는 동료 검사관들의 검토의견 및 심리분석실장의 종합의견을 참고한 후 최종적으로 담당 검사관이 판단하여 그 결과를 통보하는 품질관리시스템을 운영하고 있다(그림 1 참조). 이런 과정을 통해 저연차 대검소속 검사관들의 차트 해석 능력을 향상시키고, 고연차 심리분석실장의 의견을 담당 검사관에게 독립적으로 제시함으로써, 혹시나 존재할 수 있는 담당 검사관의 오류를 수정할 기회를 제공하고 있다. 본 연구에서는 각각의 의사결정 단계에 따른 수행에 차이가 있는지를 확인하기 위해 대검소속 검사관들을 저연차 검사관과 고연차 검사관으로 나누어 각각의 정확률 및 판단불능률을 살펴보았다. 저연차 검사관들은 경력 3년 이하의 다수의 검사관으로 이루어져 있었고 한 사건에 여러 명의 검토 의견이 포함되어 있었기 때문에 본 연구

표 2. 자백사건에서 각 신분 및 결과에 따라 얻을 수 있는 경우의 수

신분	실체적 진실	총 수(N)	가능한 검사 결과	결론	탐지결과 분류
피고인	거짓	61	거짓	일치	적중
			진실	불일치	누락
고소인	진실	30	거짓	불일치	오경보
			진실	일치	정확기각

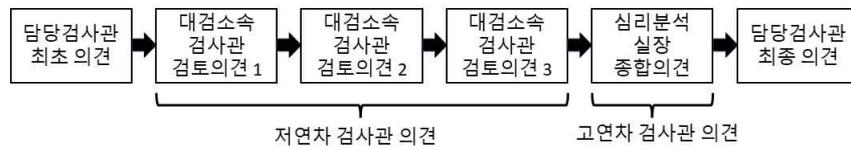


그림 1. 검찰 심리생리검사결과 의견 도출 과정

에서 사용된 자료 수($n=91$)보다 더 많은 총 220개의 자료를 분석에 사용하였다. 고연차 검사관은 경력 10년 이상인 심리분석 실장이었다. 본 연구에서는 검사를 담당한 검사관의 최초 의견, 저연차 검사관 의견, 고연차 검사관 의견 및 이를 참고한 최종 의견의 정확률 및 판단불능률을 살펴봄으로써 담당 검사관의 최초 의견이 여러 검사관의 독립적인 의견 제시에 따라 어떻게 변화하는지도 살펴보고자 하였다.

한편, OSS-3를 통한 분석을 위해 다음과 같은 소프트웨어의 옵션들을 사용하였다. 먼저 본 연구에서 사용한 'You-phase' 기법의 분석을 위하여 Event Specific/Single Issue(Zone) 옵션을 사용하였고, 채점을 위해 OSS-3 Two-stage 옵션(Senter, 2003)을 사용하였다. 검사관들의 의견과 마찬가지로, OSS-3의 결과 역시 자백 결과로부터 도출한 자료와 비교하여 정확률 및 판단불능률을 계산하였다.

또한 거짓말 탐지 연구에서도 자주 사용되고 있는 신호 탐지론(Green & Swets, 1966;

Lynn, Barrett, 2014; Zloteanu & Vuorre, 2024)을 활용해서 각 의견 도출 단계별 민감도 및 반응 기준을 계산하였다. 또한 정확률 및 판단불능률과 마찬가지로 OSS-3의 민감도 및 반응 기준을 계산하였다. 민감도와 반응 기준은 다음과 같은 방법으로 계산되었다(Green & Swets, 1996). 여기서 $HR(Hit Rate)$ 은 적중 확률을, $FAR(False Alarm Rate)$ 은 오경보 확률을 의미한다.

$$d'(\text{민감도}) = z(HR) - z(FAR)$$

$$c(\text{반응기준}) = \frac{z(HR) + z(FAR)}{2}$$

민감도 및 반응 기준의 경우, 각 단계 별 판단불능으로 결정된 자료들은 분석 대상에서 제외된 후 계산하였다. 또한, 본 연구에서 계산된 민감도가 기존 연구들에서 확인한 민감도와는 어떻게 다른지 비교해 보기 위해 Gordon(2011)의 통계치를 바탕으로 민감도를 계산한 후 이를 비교 자료로 활용하였다.

결 과

기술적 통계

본 연구에 사용된 자료들의 특성을 파악하기 위하여 기술적 통계를 실시하였다. 표 3은 본 연구에 사용된 자료들을 성별, 연령별, 신분별, 검사형태별, 사건유형별로 분류한 자료이다. 분류 결과, 남성이 67%로 여성보다 많

표 3. 성별, 연령별, 신분별, 검사형태별, 사건유형별 피검사자 수와 비율

		피검사자 수(명)	비율(%)
성별	남성	67	73.6
	여성	24	26.4
연령	20대 미만	3	3.3
	20대	22	24.2
	30대	14	15.4
	40대	19	20.9
	50대	27	29.7
	60대 이상	6	6.6
신분	피의자/피고인	61	67
	피해자/고소인	30	33
검사 형태	단독	46	73.6
	양당사자	45	26.4
사건 유형	성범죄	52	57.1
	강도와 절도	3	3.3
	교통법 및 도교법	2	2.2
	사기와 공갈	17	18.7
	상해와 폭행	11	12.1
	기타	6	6.6
	전체	91	100

았고 연령은 50대(29.7%)가 제일 많았고 20대(24.2%), 40대(19%), 30대(14%), 60대 이상(6%), 20대 미만(3%)이었다. 신분으로는 피의자/피고인이 61명(67%)으로 피해자/고소인에 비해 2배 정도 많았다. 심리생리검사를 양당사자가 모두 받았는지, 아니면 둘 중 한 명만 받았는지를 조사한 결과, 45명(49.5%)의 자료는 양당사자 사건으로 피의자/피고인 및 피해자/고소인이 동일한 검사관에게 심리생리검사를 받은 사건이었고, 46명(50.5%)의 자료는 이들 중 한 명만 단독으로 심리생리검사를 받은 사건이었다. 사건 유형으로는 김석찬과 동료들(2015)에서 사용한 기준을 토대로 분류하였는데, 성범죄 사건이 52건(57.1%)으로 제일 많았고, 사기와 공갈(18.7%), 상해와 폭행(11%), 강도와 절도(3.3%), 교통법 및 도교법 위반 사건(2.2%) 순이었으며 이에 해당하지 않는 기타사건이 총 6건(6.6%) 있었다.

의사결정 단계별 정확률 및 판단불능률

총 91건의 심리생리검사 결과에 대해서 판결문 상의 자백을 기반으로 정확률 및 판단불능률을 계산해 본 결과는 표 4와 같다. 담당 검사관의 최초 의견의 정확률은 93.5%, 판단불능률은 15.4%로 확인되었다. 이에 대한 저연차 검사관들의 의사결정을 살펴보았을 때 정확률은 93.1%로 큰 차이가 나지 않으나 판단불능률이 40.1%로 크게 상승하였다. 고연차 검사관의 의사결정의 정확률은 95.7%로 담당 검사관의 최초 의견이나 저연차 검사관들의 검토의견들보다 더 나은 정확률을 기록하였고 판단불능률은 23.1%로 담당 검사관의 최초의 의견의 판단불능률보다 다소 상승하였다. 대검찰청의 의견을 검토한 후의 담당검사관 최종

의견의 경우, 최초 의견에 비해 정확률과 판단불능률 모두 상승하는 결과를 확인할 수 있었다.

OSS-3 정확률 및 판단불능률

총 91건에 대하여 OSS-3를 활용하여 분석한 결과, 정확률은 72.7%로 검사관들의 결과에 비해 낮았다(표 4). 판단불능률의 경우, 3.2%로 검사관들의 판단불능률에 비해 낮은 판단불능률을 보였다.

의사결정 단계 및 OSS-3 결과에 대한 신호 탐지론을 활용한 분석

각 의사결정 단계 및 OSS-3 결과 자료들의 적중, 누락, 오경보, 정확기각에 대한 결과들 표 5와 같이 확인하였다. 분석 결과, 거짓 피고인들의 경우, OSS-3가 가장 높은 적중률을 보이면서 가장 낮은 판단불능률을 기록하는 등, 검사관들에 비해 나은 수행을 보인 것과 달리 진실 고소인들에 대해서는 높은 오경보율을 보임을 확인할 수 있었다. 검사관들은

표 4. 품질관리시스템 의사결정 단계 별 정확률 및 판단불능률과 컴퓨터 자동채점 분석프로그램(OSS-3)을 통한 정확률 및 판단불능률

	최초의견	저연차 검사관 검토의견	고연차 검사관 검토의견	최종의견	OSS-3 2-stage
총 수(N)	91	220	91	91	91
일치	72	121	67	70	64
불일치	5	9	3	3	24
판단불능	14	90	21	18	3
정확률	93.5%	93.1%	95.7%	95.9%	72.7%
판단불능률	15.4%	40.1%	23.1%	20%	3.2%

표 5. 자백사건을 기반으로 한 실제적 진실과 의사결정 단계 및 컴퓨터 자동채점 분석 프로그램으로 분석한 결과를 비교했을 때의 적중, 누락, 오경보, 정확기각 및 판단불능 건수

	최초의견	저연차 검사관 검토의견	고연차 검사관 검토의견	최종의견	OSS-3 2-stage	
거짓	적중	49(80.3%)	70(46.7%)	44(72.1%)	46(75.4%)	52(85.2%)
	누락	8(13.1%)	7(4.7%)	2(3.3%)	2(3.3%)	6(9.8%)
	판단불능	4(6.6%)	73(48.6%)	15(24.6%)	13(21.3%)	3(5%)
진실	오경보	1(3.3%)	6(8.6%)	1(3.3%)	1(3.3%)	18(60%)
	정확기각	23(76.7%)	49(70%)	23(76.7%)	24(80%)	12(40%)
	판단불능	6(20%)	15(21.4%)	6(20.0%)	5(16.7%)	0

20%에 가까운 판단불능률에도 불구하고 70% 이상의 높은 정확기각을 기록하였고, 오경보율은 10% 이하로 낮았으나 OSS-3의 경우 판단불능이 없었음에도 불구하고 60%의 오경보율로 인해 정확기각률이 매우 떨어지며 많은 진실 고소인들을 거짓으로 분류하는 오류를 저지르고 말았다. 이러한 오류가 검사관들에 비해 OSS-3의 성능을 하락시킨 가장 큰 원인이 되었다.

일반적으로 민감도가 0이면 진실과 거짓을 전혀 구별할 수 없는 상황을 의미하고, 민감도가 커질수록 이를 구별해 낼 수 있는 능력이 커진다고 할 수 있다(Green & Swets, 1966; Macmillan & Creelman, 2005). 본 연구에서 검사관들이 기록한 민감도는 3이 넘었는데, 이는 기존 미국 폴리그래프 협회의 메타분석 자료들로부터 계산된 결과($d=2.20$)들에 비해 높았다(그림 2 참조). 각각의 의사결정 단계별 민감도를 비교하였을 때, 최초 담당 검사관 의사결정의 민감도($d = 3.167$)가 대검 검토를 거친 후 최종 의견 시에 증가($d = 3.468$)하는 것

을 확인할 수 있었다. OSS-3의 경우 민감도는 1.011로써 이는 검찰 검사관들의 의사결정 단계에서의 민감도나 Gordon(2011)에서 제시한 결과에 미치지 못하였다.

의사결정에 있어서의 편향을 나타내는 반응 기준을 계산한 결과가 표 6에 제시되어 있다. 일반적으로 반응 기준이 0보다 크면 애매한 자극에 대해서 오경보보다 누락을 선호하는 보수적인(Conservative) 경향을 보이고, 반응기준이 0보다 작으면 애매한 자극에 대해서 누락보다는 오경보를 선호하는 관대한(Liberal) 경향을 보인다고 해석한다. 분석 결과, 검찰소속 검사관들은 보수적으로 의사결정을 하는 것을 확인할 수 있었는데, 거짓말 탐지와 관련된 의사결정 상황에서 무죄 피고인을 유죄로 평가하기 보다는 유죄 피고인을 무죄로 평가하는 경향이 높았다고 해석할 수 있다. 반대로 OSS-3 프로그램의 경우 높은 성향의 관대한 기준을 가지고 의사결정을 한다는 것을 확인할 수 있었다. 즉 무죄인 피고인을 유죄로 판단하는 성향이 유죄인 피고인을 무죄로 판단

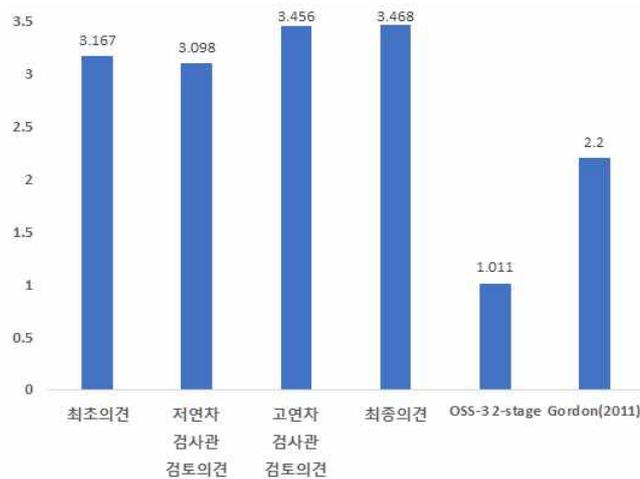


그림 2. 의사결정 단계 및 컴퓨터 자동채점 프로그램(OSS-3), 그리고 미국 APA 협회 메타분석 자료(Gordon, 2011)를 기준으로 계산한 민감도 비교

표 6. 의사결정 단계 및 컴퓨터 자동채점 분석프로그램의 정확도 및 민감도, 반응 기준

	최초의견	저연차 검사관 검토의견	고연차 검사관 검토의견	최종의견	OSS-3 2-stage
정확도(%)	93.5%	93%	95.8%	95.8%	73.7%
민감도(d')	3.167	3.098	3.456	3.468	1.011
반응 기준(c)	0.144	0.23	0.011	0.017	-0.759
기준	보수적	보수적	보수적	보수적	관대함

하는 성향보다 높았다.

검사 형태(단독/양당사자)에 따른 탐지 결과 분류

검사관들이 OSS-3에 비해 높은 정확기각를 보인 것과 관련하여 검사관들만이 오직 양당사자 정보를 채점에 활용할 수 있었기 때문

일 가능성을 확인하기 위해 단독사건과 양당사자 사건을 검사형태에 나누어 신호탐지론 분석을 실시하였다. 만약 검사관들이 피의자와 고소인 둘 중 한 명은 거짓이고 한 명은 진실인 것을 아는 상태에서 의사결정을 내렸다면 양당사자 검사에서 단독 검사보다 더 우수한 수행을 기대할 수 있을 것이다. 하지만 OSS-3의 경우, 사람과는 다르게 이러한 정보를

표 7. 검사 형태에 따른 적중, 누락, 오경보, 정확기각 및 판단불능 분류

검사형태	실체적 진실	분류	최초의견	저연차	고연차	최종의견	OSS-3 2-stage
				검사관 검토의견	검사관 검토의견		
양당사자 (n=45)	거짓	적중	19(79.2%)	29(50%)	17(70.8%)	17(70.8%)	21(87.5%)
		누락	1(4.2%)	0	0	0	2(8.3%)
	진실	판단불능	4(16.7%)	29(50%)	7(29.2%)	7(29.2%)	1(4.2%)
		오경보	1(4.8%)	2(3.9%)	1(4.8%)	1(4.8%)	13(61.9%)
		정확기각	14(66.7%)	32(62.7%)	15(71.4%)	15(71.4%)	8(38.15)
단독 (n=46)	거짓	판단불능	6(28.6%)	17(33.3%)	5(23.8%)	5(23.8%)	0
		적중	30(81.1%)	41(44.6%)	27(73%)	29(78.4%)	31(83.8%)
		누락	3(8.1%)	7(7.6%)	2(5.4%)	2(5.4%)	4(10.8%)
	진실	판단불능	4(10.8%)	44(47.8%)	8(21.6%)	6(16.2%)	2(5.4%)
		오경보	0(0%)	0	0	0	5(55.6%)
		정확기각	9(100%)	17(89.5%)	8(88.9%)	9(100%)	4(44.4%)
	판단불능	0(0%)	2(10.5%)	1(11.1%)	0	0	

활용할 수 없다는 점에서 차이가 생겼을 가능성이 있다. 표 7은 양당사자와 단독 검사 시 검사관 및 OSS-3의 수행을 보여준다. 예상과는 다르게, 검사관들의 경우, 양당사자나 단독검사 모두에서 5% 이하의 낮은 오경보율을 보임을 확인할 수 있었다. OSS-3의 경우, 단독 및 양당사자 사건에서 모두 높은 수준의 오경보율(양당사자: 61.9%, 단독: 55.6%)을 보임을 확인할 수 있었다. 이러한 결과는 검사관들의 낮은 오경보율의 원인이 의사결정 시 검사형태와 관련된 정보를 활용했기 때문은 아닐 가능성을 시사한다.

논 의

본 연구에서는 실제 검사 자료인 2016~2017년 동안 검찰에서 실시된 심리생리검사 결과들 중 피고인의 자백으로 인해 실제적 진실이 확인된 생태학적 타당성이 높은 사건들을 대상으로 심리생리검사 결과와의 일치성을 확인함으로써 현재 검찰에서 실시되고 있는 심리생리검사의 성능을 확인해 보기 위한 목적으로 연구를 진행하였다. 연구 결과, 검찰 심리생리검사의 각 의사결정 단계에서 확인된 정확성은 OSS-3와 같은 자동채점 프로그램이 제공하는 정확성이나 기준에 알려진 유사한 기법을 사용하는 메타분석 결과(Gordon, 2011)에 비해 높은 수준으로 확인되었다. 또한 신호탐지론 분석을 통한 민감도 역시 기준에 알려진 것보다 높은 결과를 얻을 수 있었다. 이러한 결론은 본 연구에서 사용된 자료들을 대상으로 했을 때 심리생리검사관들이 사용한 분류 메커니즘이 기준에 알려진 정보들에 비해 높은 신호판별 성능을 가진 분류기였음을

보여준다고 할 수 있다.

또한 이 연구에서 확인된 반응 기준이 보수적이라는 데서도 그 의의를 찾을 수 있다. 대다수의 과학적인 검사들은 필연적으로 오류를 수반할 수밖에 없는데, 애매한 사건에 대해서 유죄 편향적인 결론을 내릴지, 무죄 편향적인 결론을 내릴 지가 매우 중요한 문제가 된다. 만약 한 기관의 심리생리검사가 관대한 판단 기준을 가지고 있다면 억울한 피의자가 생길 가능성이 높아질 것이고, 반대로 보수적인 판단 기준을 가지고 있다면 일부 피의자를 놓치더라도 억울한 피의자를 양산할 가능성이 감소할 것이다. 검찰은 최종 수사기관으로써 유력한 용의자가 특정되고, 해당 피의자가 실제 범인인지를 치열하게 판단해야 하는 단계이다. 무죄 추정의 원칙에서 고려해 볼 때, 무리한 판단으로 억울한 피고인을 만들어 잘못된 법적 처벌의 가능성을 높이는 것 보다는 일부 거짓말을 하는 피고인을 놓치는 편이 인권 보호의 측면에서도 더 나은 선택일 수 있다. 이런 상황에서 OSS-3가 보여주는 지나치게 높은 관대한 기준은 수많은 죄 없는 피고인들을 만들어낼 수 있다는 데서 검찰에서 사용하기 부적절하다고 할 수 있다. 오히려 현재 검사관들이 보여주고 있는 보수적인 반응 기준이 죄 없는 피고인을 양산할 확률을 줄인다는 측면에서 더욱 바람직하다고 할 수 있을 것이다.

이 연구를 통해 심리생리검사 이후의 의사결정에 있어서 판단불능의 역할 또한 중요함을 알 수 있다. 판단불능은 심리생리검사에서 차트 채점 결과 진실 혹은 거짓 판정의 기준을 충족하지 못한 경우에 나타나며 본 연구를 통해 의사결정 단계별로 차이는 있었으나 2~30%의 자료들이 판단불능 결과를 얻는 것을 확인할 수 있었다. 이는 Gordon(2011)에서

제시하고 있는 판단불능률의 범위 안에 있음으로 일반적으로 심리생리검사에 대해 기대할 수 있는 수준의 판단불능률이라 할 수 있다(Deception INC: .128(.046), Truth INC: .255(.044)). 컴퓨터 자동채점 프로그램인 OSS-3의 경우, 이러한 판단불능률이 5% 내외로 검사관들에 비해 낮은 판단불능을 기록함을 확인할 수 있었고, 이에 따라 적중의 경우 검사관들에 비해 오히려 더 나은 수행을 보이는 것을 확인할 수 있었다. 하지만 고소인이 대상자인 경우, 검사관들은 20%가량의 사건들에 대해 판단불능 결정을 내리며 보수적인 입장을 취한 것과는 달리 OSS-3의 경우는 쉽게 거짓으로 판단해 버림으로써 높은 수준의 오경보율을 보여 전체적으로 정확성과 민감도가 하락하는 결과를 확인할 수 있었다. 즉, 진실인지 거짓인지 판별이 어려운 사건들에 대해서 무리하게 결론을 내리기 보다는 판단불능의 결과를 내리며 의사결정을 유보하는 것이 해당 사건을 제외한 다른 사건들에 대한 구별능력 및 정확성을 향상시키는 훌륭한 방법이 될 수 있다.

검찰에서 실시 중인 품질관리 시스템과 관련하여 검사관의 최초 의견이 저연차 검사관 및 고연차 검사관의 검토 이후 민감도가 향상되는 것을 확인할 수 있었다. 다만 저연차 검사관과 고연차 검사관의 수행에는 차이가 있었는데, 저연차 검사관들의 민감도에 비해 고연차 검사관의 민감도가 높으면서도 판단불능률은 낮았다. 이러한 차이는 복잡한 생리적 반응들을 패턴화하고 이를 수치화한 후 의사결정을 내리는 과정에서 오랜 경험을 통해 획득된 고연차 검사관의 전문성이 저연차 검사관들이 가지고 있는 그것에 비해 우수한 데서 나타날 가능성도 존재할 것이다. 차트 분석과

마찬가지로 복잡한 지각적인 요소를 정확하고 빠르게 구별해야 하는 지문감정의 경우도 검사관의 전문성이 초보자들과 매우 다른 수행을 이끌어 낼 수 있음을 보여주고 있다(Thompson & Tangen, 2014; Thompson et al., 2013). 다만 본 연구에 참여한 검사관의 수가 소수이고 경력 이외의 다양한 요소들이 각 검사관들의 전문성 형성에 영향을 미칠 수 있는 만큼 추후 연구를 통해 어떤 요소가 검사관들의 수행 차이에 영향을 미치는지를 확인해 보아야 할 것이다.

본 연구는 실제적 진실을 알기 어려운 실무 사건들을 피검사자의 자백을 통해 실제적 진실 확인한 후 이를 토대로 분석했다는 데서 실험실 연구들에 비해 높은 생태학적 타당도를 지닌다고 할 수 있지만, 몇 가지 한계점도 분명히 존재한다. Iacono(2008)는 Mangan과 동료들(2008)의 연구를 비판하며 심리적으로 취약한 사람이 심리생리검사에서 거짓을 받은 후 자백을 하기 때문에 이러한 사례들만 취사선택되어 정확성이 지나치게 부풀려질 수 있음을 지적하고 있다. 물론 본 연구도 이러한 비판에서 자유롭기는 어려우나, 그럼에도 한국 검찰과 미국의 심리생리검사 결과의 활용이 동일하지 않다는 점은 참고해야 할 필요성이 있다. 미국의 경우, 심리생리검사를 실시한 후 피검사자들이 명백하게 거짓을 이야기하는 것으로 보인다면 바로 검사 후 면담을 실시하고 여기에서 자백을 얻어내는 경우들이 많다. 이런 식으로 자백 사건들을 수집한다면 심리생리검사가 자백의 직접적인 원인이 되는 것이고, 따라서 Iacono(2008)가 지적하는 것과 같이 검사와 자백의 독립성을 가정하기 어려울 것이다. 하지만 한국 검찰의 경우 심리생리검사 후 검사 후 면담을 실시하지 않은 채 검사

를 종료하고, 품질관리 시스템 등 다층적인 검토를 통해 최종 의견을 정한 후 의뢰 기관에 그 결과만을 통보하는 것이 일반적이다. 본 연구 또한 검사를 실시한 지 오랜 시간이 지난 후 피의자의 자백 유무를 법원 판결문을 통해 확인하였을 뿐이다. 따라서 심리생리검사와 피검사자 자백의 시점 사이에 시간적 차이가 존재하고 심리생리검사가 자백의 직접적인 원인이 되지 않는 경우도 많다. 그럼에도 불구하고 추후 심리생리검사가 수사나 재판 과정에서 피검사자에게 심리적 압박을 부과할 수 있는 가능성은 존재할 수 있기 때문에 그 정확성이 부풀려질 가능성을 무시할 수는 없을 것이다.

또한 일반적으로 형사사건의 자백이란 피고인이 본인이 잘못했다고 인정한 사건들을 의미하기 때문에, 반대로 고소인이 거짓말을 했다 자백한 사건들을 연구를 위한 사건들이 분석에 사용될 수 없다. 이는 피의자의 경우, 적중과 누락으로만 분류되고 고소인의 경우는 오경보와 정확기각으로만 분류되기 때문에 피의자에 대한 오경보 및 정확기각, 고소인에 대한 적중과 누락을 확인할 수 없어 큰 문제가 될 수 있다. 하지만 일반적으로 고소인이 거짓말을 한 사건의 경우, 고소를 스스로 취하한다거나 검사실에서 불기소를 하는 식으로 처리가 될 것이나 해당 사유를 상세히 기재한 자료를 얻기 힘들기 때문에 현재 법 체계에서는 수집이 매우 어렵다. 앞으로 높은 생태학적 타당성을 지니며 고소인이 거짓말을 한 경우에 대한 자료를 수집할 수 있는 방법들에 대해 고민해 보아야 할 필요성이 있다.

또 다른 가능성으로 거짓말에 특히 취약한 피검사자들이 심리생리검사에서 거짓으로 쉽게 판단되고, 이후 심리생리검사 결과에 의해

자백했을 가능성 또한 생각해 볼 수 있다. 다만 본 연구에서 사용한 판결문에는 자백 유무에 대해서만 기술했을 뿐, 어느 단계에서 어떻게 자백이 이루어졌는지 구체적인 상황까지 알기는 어려우므로 추후 연구에서 구체적 자백 경위까지 확인 가능한 자료로 연구가 이루어진다면 더 높은 생태학적 타당성을 가진 결과를 얻을 수 있을 것이다. 또한 피고인들이 법적으로 유죄가 확실시되는 상황에서 형량 등을 감할 목적으로 거짓 자백했을 가능성 또한 배제하기 어렵다. 즉, 거짓자백의 경우 자백사건이 가지고 있는 생태학적 타당도를 떨어뜨릴 수 있다는 데서 커다란 문제가 될 수 있다. 하지만 해당 사건이 거짓자백의 결과물 인지를 확인할 수 있는 방법은 사실상 존재하지 않기 때문에 이러한 부분들은 실무 데이터들을 대상으로 한 연구의 한계로써 감안되어야 할 것이다.

그럼에도 불구하고, 본 연구는 생태학적 타당성이 떨어지는 실험 연구가 아니라, 실제 2년간 심리생리검사가 실제로 실시된 사건들을 통해 검찰에서 실시하는 심리생리검사의 정확성을 추측할 수 있는 자료를 제공했다는 데 의미를 가진다. 검찰을 비롯하여 경찰, 군 등 다양한 조직에서 심리생리검사를 사용하고 있지만 방법론적 차이나 채점 방식의 차이 등에 의해 다양한 사건들에 대해 유사한 수준으로 진실 및 거짓을 분류하고 있는지 알기 어려운 것이 사실이다. 실제 심리생리검사관들의 검사 결과와 OSS-3를 통한 결과를 비교해 보았을 때 확인할 수 있듯, 여러 요소들에 의해 동일한 심리생리검사 차트를 분석할 지라도 서로 다른 민감도와 판단 기준을 보일 수 있다. 비록 정확성이 과대평가될 가능성이 있는 자백 사례들을 대상으로 실제적 진실을 확

인하였다는 데서 검찰 심리생리검사의 타당성을 정확히 확인하기는 어려울 지라도, 이와 특정 상황에서 높은 수준의 분류를 할 수 있음을 확인할 수 있다.

참고문헌

- 김석찬, 장은희, 이상현, 방철, 김시은, 김현택 (2015). 폴리그래프 검사 요인에 따른 검찰 처분 및 판결 일치도 연구: 검찰청 폴리그래프 실증 연구. *한국심리학회지: 법*, 6(1), 13-31.
- 방철 (2023). 심리생리검사 결과의 증거능력에 관한 소고. *중앙법학*, 25(1), 147-170.
- 양동철 (2013). 진술서, 진술녹취서의 증거능력. *경희법학*, 48(1), 433-468.
- 정재영, 김재홍, 김미영, 강민국, 지형기, 김기호, 이장한 (2010). 상반된 주장에 대한 폴리그래프 검사의 판별 정확성. *한국심리학회지: 사회및성격*, 24(1), 1-10
- Bond, C.F., & DePaulo, B.M. (2006). Accuracy of deception judgements. *Personality and Social Psychology Review*, 10, 214-234.
https://doi.org/10.1207/s15327957pspr1003_2
- Bull, R. (1989). Can training enhance the detection of deception? *Credibility assessment*, 83-99.
https://doi.org/10.1007/978-94-015-7856-1_5
- DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of personality and social psychology*, 37(10), 1713.
<https://doi.org/10.1037//0022-3514.37.10.1713>
- Dutton, D. (2000). Guide for performing the objective scoring system. *Polygraph*, 29(2), 177-184.
- Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigation. *Journal of Applied Psychology*, 75, 521-529.
<https://doi.org/10.1037//0021-9010.75.5.521>
- Garrett BL. (2011). *Convicting the Innocent: Where Criminal Prosecutions Go Wrong*. Cambridge, MA: Harvard Univ. Press.
<https://doi.org/10.4159/harvard.9780674060982>
- Ginton, A., Daie, N., Elaad, E., & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology*, 67, 131-137.
<https://doi.org/10.1037//0021-9010.67.2.131>
- Gordon, N. (2011). Meta-Analytic Survey of Criterion Accuracy of Validated Polygraph Techniques. *Polygraph*, 40(4), 194-305.
- Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. Wiley.
- Grubin, D. (2010). The polygraph and forensic psychiatry. *Journal of the American Academy of Psychiatry and the Law Online*, 38(4), 446-451.
- Honts, C. R. (1994). Psychophysiological detection of deception. *Current Directions in Psychological Science*, 3(3), 77-82.
<https://doi.org/10.1111/1467-8721.ep10770427>
- Iacono, W. G. (2008). Accuracy of polygraph techniques: Problems using confessions to determine ground truth. *Physiology & Behavior*, 95(1-2), 24-26.
<https://doi.org/10.1016/j.physbeh.2008.06.001>
- Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior*, 23, 499-516.
<https://doi.org/10.1023/a:1022330011811>

- Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology, 73*, 291-302.
<https://doi.org/10.1037//0021-9010.73.2.291>
- Kircher, J. C., Raskin, D. C., Honts, C. R., & Horowitz, S. W. (1988). Generalizability of mock crime laboratory studies of the control question polygraph technique. *Psychophysiology, 25*, 462-463.
- Krapohl, D. (2002). Short Report: An update for the Objective Scoring System. *Polygraph, 31*, 298-302.
- Krapohl, D. & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph, 28*, 209-222.
- Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*(1), 66.
<https://doi.org/10.1037//0021-9010.70.1.66>
- Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological science, 25*(9), 1663-1673.
<https://doi.org/10.1177/0956797614541991>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates.
- Mangan, D. J., Armitage, T. E., & Adams, G. C. (2008). A field study on the validity of the Quadri-Track Zone Comparison Technique. *Physiology & behavior, 95*(1-2), 17-23.
<https://doi.org/10.1016/j.physbeh.2008.03.001>
- National Research Council, Division of Behavioral, Committee on National Statistics, Board on Behavioral, Sensory Sciences, & Committee to Review the Scientific Evidence on the Polygraph. (2003). *The polygraph and lie detection*. National Academies Press.
<https://doi.org/10.17226/10420>
- Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., ... & Oelrich, M. (2011). Using the empirical scoring system. *Polygraph, 40*(2), 67-78.
- Nelson, R. (2012). Monte Carlo Study of Criterion Validity of Backster You-Phase Examinations. *Polygraph, 41*(1), 44-53.
- Nelson, R., Handler, M., Blalock, B., & Cushman, B. (2016). Blind scoring of confirmed federal You-Phase examinations by experienced and inexperienced examiners: Criterion validity with the Empirical Scoring System and the seven-position model. *Polygraph, 45*(1).
- Nelson, R., Handler, M., & Senter, S. (2012). Monte Carlo study of criterion validity of the directed lie screening test using the Empirical Scoring System and the Objective Scoring System version 3. *Polygraph, 41*(3), 144-155.
- Nelson, R., Krapohl, D., & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph, 37*(3), 185-215.
- Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J., & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of applied psychology, 89*(6), 1099.

- <https://doi.org/10.1037/0021-9010.89.6.1099>
Senter, S. M. (2003). Modified general question test decision rule exploration. *Polygraph, 32*(4), 251-263.
- Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: Experts can do a lot with a little. *PloS one, 9*(12), e114759.
<https://doi.org/10.1371/journal.pone.0114759>
- Thompson, M. B., Tangen, J. M., & McCarthy, D. J. (2013). Expertise in fingerprint identification. *Journal of forensic sciences, 58*(6), 1519-1530.
<https://doi.org/10.1111/1556-4029.12203>
- Vrij, A. (1994). The impact of information and setting on detection of deception by police detectives. *Journal of Nonverbal Behavior, 18*, 117-136. <https://doi.org/10.1007/bf02170074>
- Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley. <https://doi.org/10.1002/acp.743>
- Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3-28. <https://doi.org/10.1007/bf00987006>
- Wickens, T. D. (1991). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology, 36*, 213-234.
[https://doi.org/10.1016/0022-2496\(92\)90037-8](https://doi.org/10.1016/0022-2496(92)90037-8)
- Zloteanu, M., & Vuorre, M. (2024). A tutorial for deception detection analysis or: How I learned to stop aggregating veracity judgments and embraced signal detection theory mixed models. *Journal of Nonverbal Behavior, 1-25*.
<https://doi.org/10.1007/s10919-024-00456-x>
- Zuckerman, M., Spiegel, N. H., DePaulo, B. M., & Rosenthal, R. (1982). Nonverbal strategies for decoding deception. *Journal of Nonverbal Behavior, 6*, 171-187.
<https://doi.org/10.1007/bf00987066>
- 1 차원고접수 : 2024. 09. 25.
심사통과접수 : 2024. 11. 24.
최종원고접수 : 2024. 11. 25.

The Validity of Psychophysiological Detection of Deception(PDD) Tests in Prosecutorial Data Using Confession Cases

Sion Kim

Supreme Prosecutors' Office,
Graduate Program in Cognitive
Science, Yonsei University

Mi Young Kim

Seok Chan Kim

Sang Hyun Lee

Supreme Prosecutors' Office

Between 2016 and 2017, a study examined 91 cases in which defendants confessed during the investigation and trial processes, verifying the consistency between these confessions and the results of psychophysiological detection of deception(PDD) tests conducted by the Supreme Prosecutors' Office of the Republic of Korea(SPO). The accuracy during all decision-making stages conducted by the SPO exceeded 90%, surpassing the accuracy recorded by automated scoring and analysis programs and that reported by the American Polygraph Association using similar techniques. Signal detection theory was applied to calculate sensitivity (d') and decision criterion (c), revealing that the sensitivity at each decision-making stage—from the examiner's initial opinion to the review opinion by SPO examiners and the final opinion—exceeded 3. This sensitivity was higher than that recorded by automated scoring programs or in previous research findings. In terms of decision criteria (c), SPO examiners used relatively conservative criteria in decision-making, whereas the automated scoring program made decisions based on more liberal criteria. Additionally, in the case of the examiners, although the rate of inconclusive judgments was higher, it was observed that errors such as the high false alarm rates seen in the automated scoring program were smaller. This study, despite the possibility that the accuracy may be overstated due to the nature of confession-based research, provides empirical data based on actual cases to better understand the performance and characteristics of PDD tests conducted by SPO.

Key words : Psychophysiological Detection of Deception, Signal Detection Theory, Validity, Sensitivity, Decision Criteria