감성과학, vol. 11, No. 4 pp.553-564. December 2008

# Backster ZCT를 사용한 폴리그라프 검사절차의 일반화가능도: 관련 질문의 개수, 반복측정 횟수, 채점자의 수에 따른 신뢰도의 변화*

# Generalizability of Polygraph Test Procedures using Backster ZCT: Changes in reliability as a function of the number of relevant questions, the number of repeated tests, and the number of raters

엄진섭**† · 한유화** · 지형기*** · 박광배**
Jin-Sup Eom**† · Yu-Hwa Han** · Hyung-Ki Ji*** · Kwang-Bai Park**

충북대학교 심리학과**
Department of Psychology, Chungbuk National University

대검찰청***
Forensic section, S.P.P.O. in Korea

Abstract : Generalizablity theory was employed to examine how the reliability of polygraph test is affected by the number of relevant questions, the number of repeated tests (the number of of charts), and the number of raters (scorers). The data consisted of the results of the polygraph tests administered to 31 crime suspects. The sample was drawn from the real polygraph tests based on Backster ZCT and archived by the Prosecutor's Office of the Republic of Korea. The numerical scores assigned by thirteen raters to the test charts were analyzed to determine the generalizablity of the scores. The largest variance component was accounted for by the examinee factor (43.97%) and the residual variance component was 16.84% of the total variance. The variance component due to the interaction between the examinee and the chart factors was 12.17% and the variance component due to the three way interaction of the examinee, the repeated test, and the relevant question factors was 10.31%. The generalizability coefficient for the current measurement procedure as practiced by the Korean Prosecutor's Office was 0.74 which suggests that the current procedure is acceptable. However, measurement procedures with the combination of more than two relevant questions, more than three repeated tests, and more than two raters were generally found to yield generalizability coefficients larger than 0.80. Therefore, such procedures need to be considered seriously in order to significantly improve the reliability of polygraph test.

Keywords : polygraph, Backster ZCT, reliability, generalizability theory

---

* 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (M10740030003-07N4003-00310)

† 교신저자 : 엄진섭 (충북대학교 심리학과)
E-mail : jseom2003@hanmail.net
TEL : 043-261-2188
FAX : 043-271-1713

**요약** : 본 연구에서는 일반화가능도 이론을 이용하여 폴리그라프 검사에 사용된 관련 질문의 개수와 반복측정 횟수 (차트의 수), 채점자 수가 폴리그라프 검사의 신뢰도에 미치는 영향을 평가하였다. 검찰청에서 형사피의자를 대상으로 Backster ZCT를 사용한 폴리그라프 검사자료 중 31명의 폴리그라프 검사자료를 표본추출하였으며, 31명의 검사자료를 13명의 채점자가 수치적 채점방법을 이용하여 채점한 점수에 대하여 일반화가능도 이론을 적용하여 분석하였다. 분석결과, 피검사자의 변량성분이 43.97%로 가장 컸으며, 다음으로 잔여오차변량성분이 16.84%, 피검사자와 반복측정 횟수의 상호작용오차변량성분이 12.17%, 피검사자와 반복측정 횟수, 관련 질문 개수의 삼원상호작용오차변량성분이 10.31%였으며, 나머지 변량성분은 모두 7% 미만이었다. 관련 질문의 개수와 반복측정 횟수, 채점자의 수에 따른 일반화가능도 계수를 산출한 결과, 바람직한 일반화가능도 계수인 0.80 이상을 보이는 조건은 관련 질문 2개 이상과 반복측정 3회 이상, 채점자 2명 이상의 조합인 것으로 나타났다.

**주제어** : 폴리그라프, Backster ZCT, 신뢰도, 일반화가능도 이론

## 1. 서 론

1961년 국방부 과학수사연구소에서 절도사건 용의자에 대해 한국 최초의 폴리그라프 검사가 실시된 이후, 검찰과 경찰, 군 수사기관 등에서 조사대상자가 거짓을 말하는지 혹은 진실을 말하는지를 판단하기 위한 폴리그라프 검사가 널리 사용되고 있다. 그러나 국방부과학수사연구소와 경찰청, 대검찰청 등에서 폴리그라프 검사관을 양성하기 위한 교육을 수행하고 있음에도 불구하고[3], 폴리그라프 검사 자체에 대한 과학적 연구는 이제 시작단계에 들어서 있는 실정이다.

폴리그라프 검사는 스트레스 반응에 근거한 것으로, 어떤 사람에게 물리적이거나 심리적인 위협이 가해지면 복합적인 형태의 생리적 반응이 자동적으로 유발되는데[33], 손바닥의 피부전도수준과 혈압을 증가시키고, 호흡활동을 감소시키는 형태로 나타난다[32]. 일반적인 폴리그라프 검사 도구는 특정한 상황 하에서 피검사자의 호흡과 혈압 및 맥박, 피부전도수준을 측정하도록 고안되었다.

심리검사는 사람들의 마음속에 존재하는 것으로 가정하는 구성개념(construct)을 측정하기 위하여, 그 구성개념에 의해 유발되었을 것으로 여겨지는 행동을 표집하고 측정하여 구성개념의 상태를 추론한다[16]. 이러한 추론은 관찰된 행동이 측정하려고 했던 구성개념을 잘 반영하고 있는 정도를 의미하는 타당도(validity)와 신뢰도(reliability)의 문제를 발생시킨다. 심리검사와 마찬가지로 폴리그라프 검사도 피검사자가 거짓을 말하고 있는지 혹은 진실을 말하고 있는지를 평가하기 위하여, 거짓말에 의해 유발되었을 것이라고 여겨지는 행동표본인 생리적 반응을 측정하여 피검사자가 거짓을 말하는지의 여부를 추론하게 된다. 따라서 폴리그라프 검사도 심리검사의 경우와 마찬가지로 타당도와 신뢰도의 문제에 직면하게 된다.

폴리그라프 검사의 타당도는 주로 거짓을 말하는 피검사자와 진실을 말하는 피검사자를 정확하게 구분해내는 정도로 평가된다. 폴리그라프 검사의 타당도와 관련된 연구들은 메타분석을 수행한 연구들이 다수 있을 정도로 지속적으로 수행되고 있으며[8, 20, 26, 34], 최근에 한국에서도 폴리그라프 검사의 타당도와 관련된 실험실 연구와 현장 연구가 수행되었다[4, 5, 6].

폴리그라프 검사의 신뢰도는 측정상황에 따른 검사점수의 안정성을 나타내는 지수로,

일반적으로 세 가지 방법으로 추정된다[26]. 첫 번째 방법은 동일한 피검사자를 동일한 방법으로 두 번 측정하여 검사결과의 일치도를 보는 것으로, 검사-재검사 신뢰도(test-retest reliability)라고 한다. 그러나 현실적 상황에서 동일한 피검사자를 두 번 반복 측정하는 것이 어렵기 때문에, 검사-재검사 신뢰도를 보고한 연구는 찾아보기 어렵다고 한다[13].

두 번째 방법은 피검사자에 대한 폴리그라프 검사 차트를 여러 채점자들이 평가한 결과가 동일한 정도를 보는 것으로, 채점자간 신뢰도(inter-rater reliability)라고 한다. 채점자간 신뢰도는 다시 구체적인 절차에 따라 몇 가지로 구분될 수 있다. 폴리그라프 검사를 실시했던 검사자(이하 검사자라고 언급함)의 평가결과가 정확한 정도와 피검사자가 누구인지를 알지 못하는 채점자(이하 채점자라고 언급함)의 평가결과가 정확한 정도를 본 연구도 있으며[22, 23], 검사자의 평가결과와 채점자의 평가결과가 일치하는 정도를 본 연구도 있고[21, 25], 채점자들간의 평가결과가 일치하는 정도를 본 연구도 있다[27].

폴리그라프 검사의 신뢰도를 추정하는 세 번째 방법은 여러 개의 질문에 대한 생리적 반응이 일관성이 있는지를 보는 것으로 내적 일관성 신뢰도(internal consistency reliability)라고 한다. 덧셈 계산 능력이 있는 아동은 여러 개의 덧셈문제 각각에 대해 정답을 할 가능성이 높은 것처럼, 거짓을 말하는 피검사자는 조사 중인 사안과 관련된 모든 질문에 거짓을 말할 것이므로 이 질문들과 관련된 모든 생리적 반응들은 서로 유사한 반응을 보여야 한다. 생리적 반응들이 서로 다른 정도는 측정 오차로 간주되며, 신뢰도를 낮추는 결과를 가져온다. 그러나 폴리그라프 검사에서 내적 일관성 신뢰도를 평가하는 것이 중요함에도 불구하고 2002년까지는 발표된 연구가 없었다고 보고되고 있고[18], 최근에 발표된 연구에서 폴리그라프 검사의 내적일관성 신뢰도를 찾아 볼 수 있다[12].

위에서 언급한 세 가지 방법 외에 일반화가능도 이론(generalizability theory)으로 폴리그라프 검사의 신뢰도 추정이 가능하다[26]. 일반화가능도 이론은 검사상황에서 개입하는 다수의 오차요인들을 동시에 분석하는 방법으로, 폴리그라프 검사의 채점자들 간에 발생하는 오차와 여러 질문에 대한 생리적 반응들 간에 발생하는 오차를 동시에 분석할 수 있으며, 분석결과를 바탕으로 신뢰도를 높일 수 있는 조건을 제시할 수 있다.

검사이론(test theory)[16]에 따르면, 하나의 검사에 포함된 문항의 개수가 신뢰도에 직접적인 영향을 미치는데, 문항의 개수가 많을수록 신뢰도는 높아진다. 세계적으로 널리 사용되는 폴리그라프 검사기법인 비교질문기법(comparison question technique)에는 조사 중인 사안에 속하는 관련 질문(relevant question)이 여러 개가 포함되어 있으며, 폴리그라프 검사를 실시할 때 한 번에 여러 번 반복 실시하여 다수의 검사차트를 얻는 것이 일반적이다(검사를 한 번 실시할 때마다 한 개의 차트가 얻어지므로, 이하 '반복측정 횟수'와 '차트의 수'는 동일한 개념으로 사용한다). 결과적으로 폴리그라프 검사에 포함된 관련 질문의 개수와 검사를 반복한 횟수가 신뢰도에 직접적인 영향을 미치게 된다. 그러나 폴리그라프 검사에 포함된 관련 질문의 개수와 검사를 반복한 회수가 신뢰도에 미치는 영향을 평가한 연구는 찾아볼 수 없으며, 채점자의 수가 신뢰도에 미치는 영향을 평가한 연구도 또한 찾아보기 어렵다.

본 연구에서는 비교질문기법에 포함된 관련 질문의 개수와 검사를 반복수행한 회수, 채점자의 수가 신뢰도에 미치는 영향을 평가하고자 하였으며, 이를 위하여 한국 검찰청에서 형사사건에 대해 가장 많이 사용하고 있는 폴리그라프 검사기법인 Backster zone

comparison technique (Backster ZCT)을 사용한 폴리그라프 검사자료를 일반화가능도 이론을 이용하여 분석하였다.

## 1.1 Backster ZCT와 채점체계

거짓말을 하지 않는 정직한 사람도 범죄와 관련된 질문을 받으면 생리적 각성 반응이 나타날 것이 자명하므로, 거짓말에 의한 폴리그라프의 반응과 다른 정서적 동요에 의한 반응을 구별한기 위한 몇 가지 방법들이 개발되었는데[2], 본 논문에서는 한국 검찰에서 가장 많이 사용하는 기법인 Backster ZCT를 중심으로 살펴볼 것이다.

관련-무관련 질문 검사(relevant-irrelevant question test)는 10~15개의 범죄 관련 질문(예; 당신이 어제 밤에 OO 편의점에서 물건을 훔쳤습니까?)과 중성적인 범죄 무관련 질문(예; 당신은 때때로 TV를 봅니까?)으로 구성되어 있다. 이 검사기법은 죄가 있는 피검사자는 관련 질문에는 거짓을 말하고 무관련 질문에는 진실을 말할 것이므로, 무관련 질문에 대한 반응에 비하여 관련 질문에서 더 큰 생리적 반응을 보일 것이라고 가정하고, 죄가 없는 피검사자는 관련 질문과 무관련 질문 모두에 진실을 말할 것이므로, 두 가지 유형의 질문에 대한 생리적 반응이 비슷할 것이라고 가정한다. 그러나 관련 질문에 더 큰 반응을 유발하는 수많은 요인들이 존재하고, 안면타당도가 낮기 때문에 현재는 거의 사용되지 않는다[28].

Reid 비교질문검사(Reid comparison question test: CQT)는 관련-무관련 검사의 단점을 보완한 검사기법으로 범죄 관련 질문과 비교 질문, 다수의 무관련 질문을 포함하고 있다[30]. 비교 질문은 관련 질문의 주제와 동일한 것이지만 피검사자가 일생동안에 한 번 이상 범했을 것으로 추정되는 행동에 관한 것으로 광범위하고 일반적인 질문으로 구성되며(예; 어떤 물건을 훔쳐본 적이 있습니까?), 검사자는 피검사자가 이 질문에 '아니오'라고 답하도록 유도한다. 정직한 사람은 관련 질문보다는 비교 질문에 더 큰 관심을 가질 것이므로 비교 질문에 대한 반응이 더 클 것으로 가정하고, 거짓을 말하는 사람은 비교 질문보다 관련 질문에 더 큰 관심을 가질 것이므로 관련 질문에 대한 반응이 더 클 것으로 가정한다.

Backster[11]는 Reid의 비교질문검사를 수정하여 영역비교검사 (zone comparison test: Backster ZCT)를 만들었다. Backster는 질문을 세 가지 영역으로 구분하였는데, 붉은 영역(red zone)은 관련 질문을 의미하고, 녹색 영역(green zone)은 거짓가능성 질문(혹은 비교 질문)을 의미하고, 검은 영역(black zone) 은 주제외 질문을 의미한다. Backster는 비교 질문을 거짓가능성 질문(probable-lie question)이라고 이름을 바꾸어 불렀는데, 그 이유는 피검사자가 비교 질문에 거짓말을 할 것이라고 가정하기 때문이었다. 거짓가능성 질문은 조사하고 있는 주제와 비슷한 사실들을 다루지만, 좀 더 일반적이고, 다소 애매하며, 피험자의 일생에서 긴 시간 동안에 발생한 일들에 관한 것이다. 거짓가능성 질문을 잘 만든다면, 대부분의 사람들은 "예"라고 말하기 어려울 것이다. Backster의 ZCT는 2개 내지 3개의 관련 질문을 포함하고 있는데, 근본적으로 동일한 질문을 약간 다르게 표현한 것이다. 관련 질문과 비교 질문간의 내용영역에서의 중복을 막기 위하여 "2007년 이전에" 와 같은 단어들 포함시켰다 (절도 사건이 2008년에 발생했다면, 비교 질문의 예로 '2007년 이전에 남의 물건을 훔쳐본 적이 있습니까?').

Backster는 폴리그라프 차트를 평가하기 위한 수치적 채점(numerical scoring) 방법을 도입하였다. 수치적 채점은 관련 질문과 관련 질문 전후에 있는 비교 질문에 대한 생리적 반응을 비교하여 체계적인 방법으로 점수를 부여한다. 7점 척도로 평가하는데, 관련

질문에 대한 생리적 반응이 극적으로 강하면 -3점을 부여하고, 관련 질문과 비교 질문에 차이가 없으면 0점을 부여하고, 비교 질문에 대한 반응이 극적으로 강하면 +3점을 부여한다. 전체 검사에서 각 관련 질문에 대한 네 가지 생리적 측정치(흉부호흡과 복부호흡, 피부전도수준, 혈압/맥박)들 각각에 대해 점수를 부여하며, 총점을 이용하여 최종 판정을 내린다.

관련 질문의 개수와 판정에 이용한 차트의 개수(반복측정 횟수)에 따라 진실판정과 거짓판정을 위한 기준점수가 달라진다[8]. 한국검찰에서는 관련 질문이 2개이고 판정에 이용한 차트의 개수가 3개일 때, 총점이 0보다 크면 진실판정을 내리고, 총점이 -12점 이하이면 거짓 판정을 내린다. 총점이 -11 ~ 0점인 경우는 판단불능 판정을 내린다.

### 1.2 일반화가능도 이론

고전검사이론은 관찰점수를 진점수(true score)와 오차점수(error score)로만 구분하므로 고전검사이론을 통하여 산출된 신뢰도는 채점자, 검사 환경, 검사 시기, 문항개수 등의 다양한 오차요인들을 복합적으로 고려하지 못한다. 반면에 일반화가능도 이론은 변량분석(ANOVA)의 틀을 이용하여 측정상황에서 발생하는 복합적인 오차요인(국면, facet)을 동시에 분석하여 오차점수에 기여하는 다양한 원천을 구분할 수 있으며[1, 14, 17], 이를 통하여 관찰점수의 신뢰도에 미치는 오차요인들의 상대적 영향력을 파악하고, 바람직한 신뢰도를 확보하기 위한 효율적인 측정을 설계하도록 도와준다.

일반화 가능도 이론은 G 연구(generalizability study)와 D 연구(decision study)로 구분된다. G 연구에서는 검사점수에 영향을 미치는 요인들을 파악하고, 각 요인들이 검사점수에 미치는 영향을 변량성분(variance component)으로 추정한다. D 연구에서는 G 연구에서 추정된 변량성분의 상대적 크기를 이용하여 일반화가능도 계수(generalizability coefficient)를 추정하고, 효율적인 측정 구조를 결정하게 된다. 고전검사이론의 신뢰도 계수에 해당하는 일반화가능도 계수는 전집점수변량($\sigma_p^2$)과 오차점수변량($\sigma_\delta^2$)의 합에 대한 전집점수 변량($\sigma_p^2$)의 비이며, 일반화가능도 계수는 다음와 같이 계산된다.

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$$

## 2. 방 법

### 2.1 폴리그라프 검사 자료

2006년에 대검찰청과 지방검찰청에서 형사피의자를 대상으로 실시한 폴리그라프 검사 자료 중 31명의 폴리그라프 검사 자료를 표본추출하여 연구 자료로 사용하였다. 31명의 평균연령은 41.52세(범위 25세 ~ 58세) 였으며, 성별은 남자가 25명이었고 여자가 6명이었다. 죄명은 도로교통법 위반관련 7명, 사기 5명, 폭력행위 5명, 성폭력 및 강제추행 4명, 사/공문서 위조 3명, 기타 7명이었다. 31명에 대한 폴리그라프 검사에 사용된 기법은 모두 Backster ZCT 였으며, 관련 질문은 2개만 있었고, 3번의 검사를 반복 실시하여 3개의 검사차트가 있었다. 31명의 형사피의자에게 폴리그라프 검사를 실시했던 검사관이 채점한 점수는 표 1과 같다.

폴리그라프 검사에 사용된 도구는 흉부와 복부 호흡의 변화, 피부전도수준, 혈압 및 맥박 등을 측정하는 장비로 Ultrascribe, CPS, Factfinder 및 LX-2000W였다.

표 1. 31명의 형사피의자에 대한 폴리그라프 검사 점수분포

| 점수 | -28 | -26 | -22 | -21 | -17 | -16 | -15 | -14 | -13 |
|---|---|---|---|---|---|---|---|---|---|
| 빈도 | 1 | 1 | 1 | 1 | 2 | 5 | 1 | 1 | 2 |
| 점수 | -12 | -3 | -2 | 0 | 3 | 4 | 6 | 7 | |
| 빈도 | 5 | 1 | 4 | 2 | 1 | 1 | 1 | 1 | |

## 2.2 채점자

대검찰청과 지방검찰청에서 폴리그라프 검사를 수행하고 있는 13명의 폴리그라프 검사관이 31명에 대한 폴리그라프 검사자료를 모두 채점하였다. 채점자에게는 31명의 검사자료를 3개의 차트로 구분하고, 각 차트에서 4가지 생리적 반응(흉부호흡, 복부호흡, 피부전도수준, 혈압 및 맥박)들을 따로 구분하여, 총 31×3×4=372 쪽으로 구성된 검사자료를 제공하였으며, 각 피검사자내에서 채점의 독립성을 유지하도록 하기 위하여 372쪽으로 구성된 검사자료를 무작위로 섞어서 제공하였다. 채점은 대검찰청에서 실무에 사용하고 있는 7점 체계를 이용하여 각각의 관련 질문에 대한 생리적 반응을 수치적으로 채점하도록 하였다. 수치적 채점기준은 현재 대검찰청에서 실무에 사용하고 있는 채점기준을 이용하도록 하였다. 채점자의 폴리그라프 실무경력은 평균 5.9년(범위 3년~10년)이었다.

## 2.3 점수산출

31명의 검사자료는 3개의 차트로 구성되어 있으며, 각 차트에는 2개의 관련 질문이 있고, 각 관련 질문에 대한 4가지 생리적 반응을 13명의 채점자가 7점 체계(-3점~+3점)로 채점하도록 하였으므로, 총 31×3×2×4×13 개의 점수가 산출되었다. 그러나 4가지 생리적 반응에 대한 각 점수는 신뢰도 추정의 대상이 아니므로 4가지 생리적 반응 점수를 합산하여, 최종적으로 31(피검사자) × 13(채점자) × 3(반복측정) × 2(관련 질문) 점수행렬을 분석에 이용하였다.

## 2.4 분석방법

일반화가능도 이론의 G 연구에서 사용한 국면은 피험자(P), 채점자의 수(R), 반복측정 횟수(C), 관련 질문의 개수(I)였으며, 이 국면들은 무수히 많은 값들 중 일부를 표본추출하여 사용한 것이므로 모든 국면들을 무선요인으로 간주하였다. 반복측정변량분석 모형을 이용하여 각 국면과 상호작용효과의 평균자승합을 산출한 후, 이를 이용하여 변량성분을 계산하였다. 전체 측정에 대한 각 변량성분의 상대적 기여도는 각 변량성분을 모든 변량성분의 합으로 나눈 값으로 파악하였다.

G 연구 결과를 이용하여 D 연구를 수행하였다. D 연구에서는 오차 국면의 수를 조절해 가면서 측정설계들을 평가하고, 효율적인 측정구조를 탐색하였다. 채점자의 수는 1~3명까지 조절하였으며, 반복측정 횟수는 2~4개 까지 조절하였고, 문항의 수는 2개와 3개로 조절하면서 D 연구의 신뢰도 계수인 G 계수를 산출하였다. 0.8 이상의 G 계수가 바람직하며, 0.70 ~ 0.79의 신뢰도 계수는 수용가능한 수준인 것으로 해석할 수 있다[31].

일반화가능도 분석은 GENOVA[15]를 이용하였으며, 완전 교차된 31(피검사자) × 13(채점자) × 3(반복측정) × 2(관련 질문) 무선효과 반복측정 변량분석 모형을 사용하였다.

# 3. 결 과

## 3.1 일반화가능도 분석 결과

### 3.1.1 G 연구 결과

피검자자와 채점자의 수, 반복측정 횟수, 관련 질문의 개수를 국면으로 가지는 G 연구 결과로 산출된 변량성분 추정치와 변량성분의 상대비율이 표 2에 제시되어 있다. 피

검사자 변량은 고전검사이론에서 진점수 변량에 해당하며, 나머지는 모두 오차변량에 해당한다. 피검사자의 변량성분백분율은 43.97%로 높은 값을 보였으며, 채점자 수의 변량성분백분율은 3.56%, 반복측정 횟수의 변량성분백분율은 0.08%, 관련 질문 개수의 변량성분백분율은 0.00%로 각 오차의 주변량성분은 작은 것으로 나타났다.

채점자와 관련된 상호작용변량성분을 보면, 피검사자와 채점자 수의 상호작용변량성분은 3.97%로 비교적 낮았으며, 채점자 수와 반복측정 횟수의 상호작용변량성분은 0.01%, 채점자 수와 관련 질문 개수의 상호작용변량성분은 0.06%으로 낮았다. 피검사자와 채점자 수, 반복측정 횟수의 삼원상호작용변량성분은 6.57%로 다소 높았으며, 피검사자와 채점자 수, 관련 질문 개수의 삼원상호작용변량성분은 0.78%로 낮았고, 채점자 수와 반복측정 횟수, 관련 질문 개수의 삼원상호작용변량성분은 0.00%로 낮았다. 결과적으로 채점자 수와 관련된 가장 큰 변량성분은 피검사자와 채점자 수, 반복측정 횟수의 삼원상호작용이었으나 큰 값은 아니었다.

반복측정과 관련된 상호작용변량성분을 보면, 피검사자와 반복측정 횟수의 상호작용변량성분이 12.17%로 비교적 높은 값을 보이고 있으며, 채점자 수와 반복측정 횟수의 상호작용변량성분은 0.01%, 반복측정 횟수와 관련 질문 개수의 상호작용변량성분은 1.69%였다. 피검사자와 채점자 수와 반복측정 횟수의 삼원상호작용변량성분은 6.57%, 피검사자와 반복측정 횟수, 관련 질문 개수의 삼원상호작용변량성분은 10.31%로 각각 비교적 높은 값을 보였으며, 채점자 수와 반복측정 횟수, 관련 질문 개수의 삼원상호작용변량성분은 0.00%로 낮았다. 결과적으로, 반복측정 횟수와 관련된 변량성분 중 피검사자와 반복측정 횟수의 상호작용과 피검사자와 반복측정 횟수, 관련 질문 개수의 삼원상호작용은 주요한 오차성분인 것으로 나타났다.

관련 질문 개수와 관련된 상호작용변량성분을 보면, 피검사자와 관련 질문 개수의 상호작용변량성분은 0.00%, 채점자 수와 관련 질문 개수의 상호작용변량성분은 0.06%, 피

표 2. p×r×c×i 설계에 대한 G 연구결과

| 효과 | 자유도 | 자승합 | 평균자승 | 변량성분추정치(%) |
|---|---|---|---|---|
| 피검사자 | 30 | 6507.36 | 216.91 | 2.446 (43.97) |
| 채점자 | 12 | 477.94 | 39.83 | 0.198 ( 3.56) |
| 반복측정 | 2 | 135.97 | 67.99 | 0.005 ( 0.08) |
| 관련질문 | 1 | 6.06 | 6.06 | 0.000 ( 0.00) |
| 피검사자×채점자 | 360 | 1124.14 | 3.12 | 0.221 ( 3.97) |
| 피검사자×반복측정 | 60 | 1603.62 | 26.73 | 0.677 (12.17) |
| 피검사자×관련질문 | 30 | 193.37 | 6.45 | 0.000 ( 0.00) |
| 채점자×반복측정 | 24 | 29.77 | 1.24 | 0.001 ( 0.01) |
| 채점자×관련질문 | 12 | 10.81 | 0.90 | 0.003 ( 0.06) |
| 반복측정×관련질문 | 2 | 91.68 | 45.84 | 0.094 ( 1.69) |
| 피검사자×채점자×반복측정 | 720 | 1201.31 | 1.67 | 0.366 ( 6.57) |
| 피검사자×채점자×관련질문 | 360 | 383.93 | 1.07 | 0.043 ( 0.78) |
| 피검사자×반복측정×관련질문 | 60 | 503.56 | 8.39 | 0.574 (10.31) |
| 채점자×반복측정×관련질문 | 24 | 11.44 | 0.48 | 0.000 ( 0.00) |
| 피검사자×채점자×반복측정×관련질문 | 720 | 674.66 | 0.94 | 0.937 (16.84) |

검사자와 채점자 수, 관련 질문 개수의 삼원상호작용변량성분은 0.78%, 피검사자와 반복측정 회수, 관련 질문 개수의 삼원상호작용변량성분은 10.31%, 채점자 수와 반복측정 횟수, 관련 질문 개수의 삼원상호작용변량성분은 0.00%로 나타났다. 결과적으로, 관련 질문 개수와 관련된 변량성분은 피검사자와 반복측정 횟수, 관련 질문 개수의 삼원상호작용변량성분에서만 높은 값을 보였다.

잔여오차변량을 나타내는 모든 국면들 간의 상호작용변량성분은 16.84%였다.

### 3.1.2 D 연구 결과

채점자의 수를 1명에서 3명까지 변화시키고, 반복측정의 횟수를 2개에서 4개까지 변화시키고, 관련 질문의 개수를 2개에서 3개로 변화시키면서 일반화가능도 계수를 산출한 결과가 표 3에 제시되어 있다. 채점자의 수를 1명에서 2명으로 증가시키면, 일반화가능도 계수는 .060~.070 증가하였으며, 채점자의 수를 2명에서 3명으로 증가시키면, 일반화가능도 계수가 .021~.026 증가하는 것으로 나타났다. 반복측정의 횟수를 2번에서 3번으로 증가시키면, 일반화가능도 계수는 .050~.063 증가하였으며, 3번에서 4번으로 증가시키면 .280~.036 증가하였다. 관련 질문의 수를 2개에서 3개로 증가시키면 .012~.027 증가하는 것으로 나타났다.

현재 한국 검찰청에서 사용하고 있는 것처럼, 채점자 1명, 반복측정 3회, 관련 질문 2개를 사용할 경우의 일반화가능도 계수는 .742로 수용 가능한 수준인 것으로 해석된다. 그림 1에서 볼 수 있는 것처럼, 바람직한 수준의 일반화가능도 계수인 .8 이상을 보이는 조건은 채점자가 2명 이상이면서 반복측정 횟수가 3번 이상인 경우였다.

## 3.2 검사자와 채점자간 일치율

31명의 형사피의자에게 폴리그라프 검사를 실시했던 검사관이 채점한 점수와 13명의 채점자가 채점한 점수 간 일치율을 산출하였다. 일치율은 두 가지 방법을 이용하여 산출하였는데, 한 가지는 검사관이 채점한 점수와 각 채점자가 채점한 점수간 상관계수를 산출하였으며, 다른 한 가지는 검사관이 채점한 점수와 각 채점자가 채점한 점수를 근거로 내린 최종 판정(-12점 이하이면 거짓판정, 0점 보다 크면 진실 판정, -11~0점은 판정불능)간 일치율을 산출하였다.

검사관이 채점한 점수와 채점자들이 채점한 점수 간 상관계수의 평균은 .792(범위 .668~.867, 모두 $p<.001$)였으며, 13명의 채점자간 상관계수의 평균은 .876(범위 .758~.956, 모두 $p<.001$)이었다.

검사관이 채점한 점수에 근거한 판정과 채점자들이 채점한 점수에 근거한 판정의 일치율은 평균 .700(범위 .516~.871)이었으며, 우연에 의한 일치율을 고려한 일치도 지수인 Cohen's Kappa는 평균 .503(범위 .272~.755)으로 보통 수준(moderate)이었다[9]. 판정결과가 뒤바뀐 경우는 1명의 채점자에서 1개의 사례가 있었을 뿐이다. 13명의 채점자간 판정 일

**표 3.** p×R×C×I 설계에 대한 여러 측정조건에서의 일반화가능도 계수

| 채점자 | 반복측정 횟수 2번 | | 반복측정 횟수 3번 | | 반복측정 횟수 4번 | |
|---|---|---|---|---|---|---|
| | 관련질문 2개 | 관련질문 3개 | 관련질문 2개 | 관련질문 3개 | 관련질문 2개 | 관련질문 3개 |
| 1명 | .679 | .706 | .742 | .763 | .778 | .796 |
| 2명 | .749 | .770 | .806 | .823 | .840 | .852 |
| 3명 | .775 | .795 | .830 | .845 | .861 | .873 |

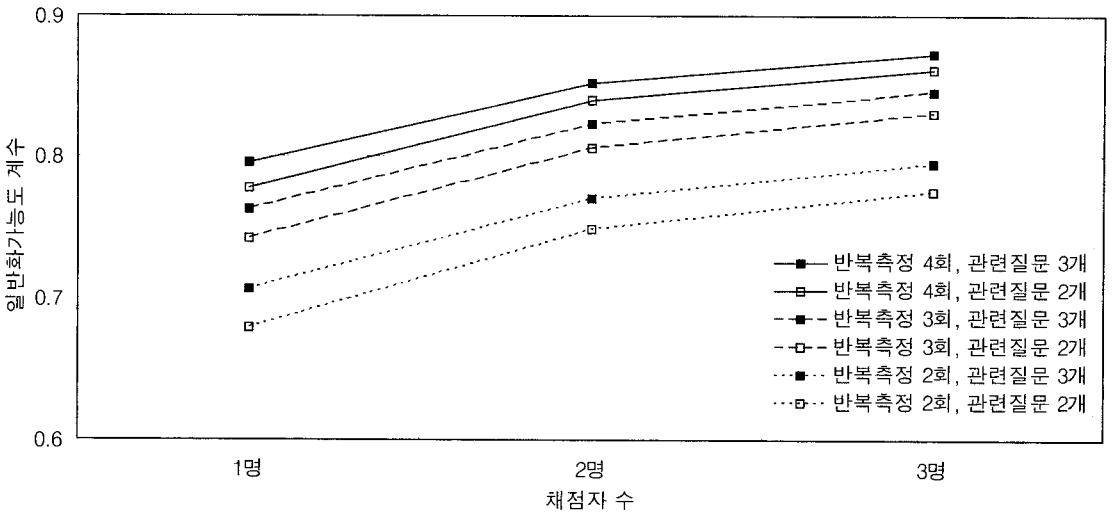


그림 1. 채점자의 수와 반복측정 횟수, 관련 질문의 수에 따른 일반화가능도 계수

치율의 평균은 .691(범위 .451~.903)이었으며, Cohen's Kappa는 평균 .505(범위 .206~.827)로 보통 수준(moderate)이었다.

### 3.3 내적일관성 신뢰도

13명의 채점자 각각에 대하여, 6 문항(반복측정 3회×관련 질문 2개)에 대한 생리적 반응을 채점한 점수의 Cronbach α를 산출하였다. 13명의 채점자 평균은 .868(범위 .748~.926)이었다.

## 4. 논 의

폴리그라프 검사의 신뢰도와 관련된 기존의 연구들은 검사를 실시한 검사자가 폴리그라프 검사 차트를 채점한 점수와 검사를 실시하지 않은 채점자가 채점한 점수가 일치하는 정도를 평가한 것들이 대부분이었다. 그러나 일반적인 심리검사에서와 같이, 채점자와 더불어 검사에 포함된 문항의 수와 검사를 반복한 횟수가 폴리그라프 검사의 신뢰도에 중요한 영향을 미친다.

본 연구에서는 기존의 폴리그라프 검사의 신뢰도 연구와는 달리, 일반화가능도 이론을 이용하여 채점자의 수와 검사를 반복한 횟수(차트의 수), 검사에 포함된 관련 질문의 개수가 신뢰도에 미치는 영향을 분석하고, 나아가 적절한 수준의 신뢰도를 보장하기 위한 측정조건을 제시하고자 하였다.

일반화가능도 이론의 G 연구 결과, 폴리그라프 검사점수의 전체변량성분 중 피검사자 변량성분이 43.97%로 가장 큰 비율을 차지하는 것으로 나타났으며, 잔여오차변량을 나타내는 모든 국면들 간의 상호작용변량성분이 16.84%로 두 번째로 높았고, 피검사자와 반복측정 횟수의 상호작용변량성분이 12.17%, 피검사와 반복측정 횟수, 관련 질문 개수의 상호작용변량성분이 10.31%로 높은 것으로 나타났다. 고전검사이론의 신뢰도 계수에 해당하는 일반화가능도 계수를 추정하기 위한 D 연구결과, 현재 검찰청에서 일반적으로 사용하는 절차(2개의 관련 질문 × 3번의 반복측정 × 1명의 채점자)의 일반화가능도 계수

는 .742로 추정되었다. 바람직한 수준의 일반화가능도 계수인 0.8 이상의 값을 보이는 조건은 반복측정 3회 이상, 관련 질문 2개 이상, 채점자 2명 이상인 것으로 나타났다. 즉, 현재 검찰청에서 사용하고 있는 조건에 채점자를 1명 추가하면 바람직한 수준의 일반화가능도 계수에 도달하는 것으로 나타나, 다수의 채점자가 폴리그라프 검사 차트를 채점하고 다수의 채점자가 산출한 총점을 평균하여 최종 판단에 이용할 것이 권고된다.

기존에 폴리그라프 검사의 채점자간 신뢰도를 추정할 때 사용하였던 방법을 사용하여 채점자간 일치율을 산출한 결과, 검사관이 채점한 점수와 채점자들이 채점한 점수간 상관계수의 평균은 .792인 것으로 나타났으며, 최종판단의 일치율은 .700인 것으로 나타나, 일반적인 채점자간 일치도 수준인 .85~.90[8, 24]보다 낮은 것으로 나타났다. 그러나 이러한 결과는 채점자들 간의 일치도가 낮기 때문에 나타난 결과이기 보다는 실험방법과 실험에 사용한 자료의 특성 때문인 것으로 판단된다.

한국 검찰청에서는 폴리그라프 검사관이 검사 차트를 채점할 때에는 하나의 관련문항에 대한 네 가지 생리적 반응(흉부호흡, 복부호흡, 피부전도수준, 혈압/맥박)을 종합적으로 고려하여 채점하지만[7], 본 연구에서 채점자들은 하나의 관련 질문에 대한 네 가지 생리적 반응들을 각각 독립적으로 채점하였다. 따라서 검사관의 채점결과와 채점자의 채점 결과 간 상관계수가 기존의 연구들보다 낮게 나타난 것은 실험설계상에서 발생한 당연한 결과인 것으로 판단된다. 본 연구에서 검사관의 최종판단과 채점자의 최종판단간의 일치율이 기존 연구들 보다 다소 낮았던 이유도, 위와 같은 실험설계상의 이유와 더불어 본 연구에 사용한 표본의 특성 때문인 것으로 생각된다. 표 1에서 보는 바와 같이, 본 연구에 사용된 표본의 점수분포가 최종판정을 위한 절단점(-11점과 0점) 전후의 사례들이 약 50%가량 포함되어 있어서 검사관의 최종판단과 채점자의 최종판단간 일치율이 낮아진 것으로 여겨진다.

앞서 언급한 바와 같이 한국 검찰청에 있는 폴리그라프 검사관이 검사 차트를 채점하는 방법과 본 연구에서 채점자들이 검사 차트를 채점하는 방법이 서로 달랐다. 본 연구에서 채점자들에게 네 가지 생리적 반응들을 각각 독립적으로 채점하도록 한 이유는 미국 국방성 폴리그라프 연구소(Department of Defense Polygraph Institute)에서 사용하고 있는 수치적 채점 방법[29]으로 신뢰도를 평가하고자 했었기 때문이다. 네 가지 생리적 반응들을 함께 고려하여 채점하는 것이 그렇지 않은 경우보다 내적 일관성 신뢰도를 높여주므로[7], 본 연구에서 산출한 일반화가능도 계수는 한국 검찰청에서 사용하는 폴리그라프 검사절차에 대한 일반화가능도 계수의 최소추정치라고 생각할 수 있다.

본 연구가 일반화가능도 이론을 이용하여 폴리그라프 검사의 신뢰도를 평가한 첫 연구라는 의미를 가지는 반면, 몇 가지 제한점을 가진다. 첫 번째로, 폴리그라프 검사의 신뢰도를 평가한 기존의 연구들은 피검사자가 거짓을 말했는지 혹은 진실을 말했는지에 대한 실체적 진실(ground truth)을 준거로 검사관의 판단과 채점자의 판단이 얼마나 정확했는지를 비교하였다. 그러나 본 연구에서는 분석 자료에 대한 실체적 진실을 파악하지 못하여, 검사관과 채점자의 판단이 얼마나 정확했는지를 평가하지 못하였다. 두 번째로, 검찰청에서 실시한 폴리그라프 검사 중에서 Backster ZCT를 활용한 자료만을 이용하여 일반화가능도 계수를 산출하였으므로, 다른 기법(예; Utah probable-lie test, modified general question test)을 사용한 폴리그라프 검사나 한국의 군수사기관이나 경찰에서 사용하는 폴리그라프 검사의 일반화가능도 계수로 일반화하기 어렵다.

## 참고문헌


[1] 김성숙, 김영분 (2001). 일반화가능도 이론, 교육과학사, 서울.

[2] 박광배 (2002). 법심리학, 학지사, 서울.

[3] 박판규 (2003). 거짓말탐지검사, 삼우사, 서울.

[4] 엄진섭, 지형기, 박광배 (2008). 폴리그라프 검사의 타당도 추정, 한국심리학회지: 사회문제, 14(4), 1-18.

[5] 조은경 (2006). 폴리그라프 검사의 타당성 및 행동분석과의 상관성 연구, 2006 대검찰청 정책연구용역과제 최종보고서.

[6] 지형기, 정재영, 강민국, 김재홍, 김미영, 정규희, 이장한 (2008). 폴리그래프검사의 타당성에 관한 연구, 제4회 심리생리검사 관세미나자료집, 대검찰청.

[7] 한유화, 정재영, 박광배 (2008). Effects of Consistency Criterion for Scoring Polygraph Test for Crime Suspects, 2008 한국심리학회 연차학술대회 논문집, 108-109.

[8] Abrams, S. (1989). The Complete Polygraph Handbook, Lexington Books, Lexington, MA.

[9] Altman, D. G. (1991). Practical statistics for medical research, Chapman and Hall, London.

[10] Ansley, N. (1990). Validity and Reliability of Polygraph decisions in Real Cases. Polygraph, 19(3), 169-181.

[11] Backster, C. (1963). The Backster chart reliability rating method. Law and Order, 1, 63-64.

[12] Bell, B. G., Kircher, J. C., & Bernhardt, P. C. (2008). New measures improve the accuracy of the directed-lie test when detecting deception using a mock crime, Physiology & Behavior, 94, 331-340.

[13] Ben-Shakhar, G. (2002). A Critical Review of the Control Questions Test(CQT). In Murray Kleiner (Ed.), Handbook of Polygraph Testing, Academic Press, San Diego.

[14] Brennan, R. L. (2001). Generalizability Theory, Spriger, Now York.

[15] Crick, J. E., & Brennan, R. L. (1983). manual for GENOVA: A generalized analysis of variance system, ACT Publications, Iowa.

[16] Crocker, L. M., & Algina, James. (1986). Introduction to Classical and Modern Test Theory, Holt, Rinehart and Winston, Inc., New York.

[17] Cronbach, L, J, Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles, Wiley, New York.

[18] Fiedler, K., Schmid, J., & Stahl, T. (2002). What Is the Current Truth About Polygraph Lie Detection?, Basic and Applied Social Psychology, 24, 313-324.

[19] Forensic Research. (1997). Validity and Reliability of Polygraph Testing, Polygraph, 26, 215-239.

[20] Forensic Research. (1997). Validity and Reliability of Polygraph Testing, Polygraph, 26, 215-239.

[21] Honts, C. R., & Raskin, D. C. (1988). A field study of the validity of the directed lie control question. Journal of Police Science and Administration, 16, 56-61.

[22] Honts, C. R., Raskin, D. C., &

Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. Journal of Applied Social psychology, 79, 252-259.

[23] Horvath, F. (1977). The effect of selected variables on the interpretation of polygraph records. Journal of Applied Psychology, 62, 127-136.

[24] Iacono, W. G. (2008). Accuracy of polygraph techniques: Problems using confessions to determine ground truth. Physiology & Behavior, 95, 24-26.

[25] 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, 17-23.

[26] National Research Council. (2003). The Polygraph and Lie Detection, National Academies Press, Washington, D.C.

[27] Patrick, C. J., & Iacono, W. G. (1991). Validity of the control question polygraph test: the problem of sampling bias. Journal of Applied Psychology, 76, 229-238.

[28] Raskin, D. C., & Honts, C. R. (2002). The Comparison Question Test. In Murray Kleiner (Ed.), Handbook of Polygraph Testing, Academic Press, San Diego.

[29] Polygraph Institute (2006). Psychophysiological detection of deception program: psychophysiological detection of deception analysis II --course #503, Department of Defense, USA.

[30] Reid, J. E., & Inbau F. E. (1977). Truth and Deception: The Polygraph ("Lie Detector") Technique, Williams and Wilkins, Baltimore

[31] Shavelson, R. J., & Webb, N. M. (1991). Generalizability Theory: A Primer, Sage, Newbury Park.

[32] Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). Psychological Recording, 2nd Ed., Oxford Press, New York.

[33] Thompson, R. F. (2000). The Brain: A Neuroscience Primer, 3rd Ed., Worth, New York.

[34] U.S. Congress, Office of Technology Assessment. (1983). Scientific Validity of Polygraph Testing: A Research Review and Evaluation - A Technical Memorandum. OTA-TM-H-15, November, U.S. Government Printing Office, Washington D.C.

원고접수 : 08/09/30
수정접수 : 08/11/05
게재확정 : 08/12/08