Journal of KIIT. Vol. 18, No. 4, pp. 77-87, Apr. 30, 2020. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2020.18.4.77

# 컴퓨터 기반 맞춤형 검사의 구현을 위한 웹 기반 플랫폼(LIVECAT)의 정확성 및 효율성 검증

김도경*, 서동기**

# Accuracy and Efficiency of Web-based Assessment Platform (LIVECAT) for Computerized Adaptive Testing

Do-Gyeong Kim*, Dong-Gi Seo**

이 연구는 한림대학교 교내 연구비의 지원으로 수행되었습니다. (기관과제 번호:HRF-201809-001)


## 요 약

최근 컴퓨터 기반 맞춤형 검사에 대한 필요성이 증가하는 데 반해 현장적용 수준은 미미한 편이다. 이에 본 연구는 컴퓨터 기반 맞춤형 검사를 실질적으로 구현할 수 있는 웹 기반 플랫폼, LIVECAT을 개발하고, 개발한 플랫폼의 정확성과 효율성을 검증하였다. 해당 플랫폼은 문항반응이론을 적용한 컴퓨터 기반 맞춤형 검사를 활용하여 문항은행의 제작 및 관리부터 검사의 제작, 검사 시행, 결과 제공까지 가능하도록 개발되었다. H 대학교 전공강의 수업 수강생 141명을 대상으로 해당 플랫폼의 정확성 및 효율성 검증을 위한 연구를 진행하였다. 그 결과, LIVECAT을 통해 추정된 능력모수와 지필검사를 통해 추정한 능력모수의 상관계수가 .779($p<.001$)로 산출되었다. 또한 LIVECAT을 활용한 검사의 길이가 지필검사의 길이보다 약 33% 짧았다. 즉, LIVECAT을 활용할 경우 지필검사보다 더 적은 수의 문항을 통해 지필검사의 결과와 비슷한 수준의 결과를 제공할 수 있음을 검증하였다.

## Abstract

Recently, there has been an increase in the need for computerized adaptive testing(CAT), however its practical use is limited. In this study, we developed a web-based platform, LIVECAT, that could practically implement CAT and verified the accuracy and efficiency of our CAT platform. By using this web-based platform based on item response theory, user enable to produce and manage a item bank, conduct tests and report results of tests easily. 141 students of H University attended on this study to demonstrate the accuracy and efficiency of the LIVECAT. As a result, the correlation coefficient between $\theta$ estimated through the LIVECAT and $\theta$ computed by paper-based test was .779($p<.001$). In addition, the length of the test using LIVECAT was about 33% shorter than the length of the paper-based test. In conclusion, if examinees take the exam by using LIVECAT, they can obtain similar results compared to those of a paper-based test with fewer items.


Keywords
computerized adaptive testing(CAT), item response theory(IRT), rasch model, web-based platform

* ㈜ 더캣코리아 연구원
- ORCID: https://orcid.org/0000-0002-9617-4739
** 한림대학교 심리학과 조교수(교신저자)
- ORCID: https://orcid.org/0000-0002-3049-5556

· Received: Feb. 28, 2020, Revised: Mar. 24, 2020, Accepted: Mar. 27, 2020
· Corresponding Author: Dong-Gi Seo
Dept. of Psychology, Hallym University, Hallymdaehak-gil, Chuncheon-si, Gangwon-do, Republic of Korea
Tel.: +82-33-248-1727, Email: wmotive@hallym.ac.kr

## Ⅰ. 서 론

IT(Information Technology)의 발달로 인해 많은 분야에서 인터넷 및 컴퓨터 등이 활용되고 있다. 검사의 영역 또한 마찬가지이다. 현재 우리나라에서도 운전면허 필기시험이나 기업의 고객만족도 확인을 위한 설문 등 다양한 검사들이 컴퓨터 혹은 모바일 기기를 통해 이루어지고 있다. 현재까지 우리나라에서 IT기기를 통해 실시되는 대부분의 검사들은 지필검사의 내용을 그대로 IT기기 화면으로 옮겨 검사를 진행하는 컴퓨터 기반 검사(CBT, Computer Based Test)이다. 일반적인 지필검사와 비교하였을 때, CBT는 여러 장점이 있다[1]-[4]. 첫 번째로, CBT를 실시할 경우, 지필검사보다 더 많은 정보를 수집할 수 있다. 일반적인 지필검사의 경우 피검자의 응답만을 수집하는 것에 반해 CBT의 경우 문항별 응답시간, 총 소요 시간, 답안 수정 내역 등의 정보를 추가적으로 수집할 수 있다. 두 번째, 여러 가지 색상의 글자나 이미지, 동영상 등의 멀티미디어를 활용한 문항의 제시가 용이하다. 세 번째, 지필검사보다 인력과 경비, 시간 등을 절약할 수 있다. 네 번째, 검사의 시기와 장소에 대한 제약이 적다. 이와 같은 장점 때문에 CBT의 활용도가 점점 높아지는 추세이다[5].

이러한 CBT의 장점을 가장 극대화한 형태의 검사가 바로 컴퓨터 기반 맞춤형 검사(CAT, Computerized Adaptive Testing)이다. CAT은 각 피험자의 능력 수준에 따라 다음 문항을 선별적으로 제시하는 검사방법이다[6]. 이미 많은 선행연구에서 CAT의 효과성에 대한 증거를 제시하였다[7]. CAT은 기존의 고전검사이론(Classical test theory) 기반 검사의 낮은 정확성과 비효율성을 해결하기 위한 대안으로 현장에서 많이 사용되고 있다[8]-[10]. 특히 짧은 시간 내에 더 적은 양의 문항을 통해서 보다 정확하게 피검자의 능력을 측정할 수 있다는 점이 CAT의 가장 큰 장점이다[6]. CAT을 사용할 경우, 일반적인 지필검사보다 평균적으로 50% 짧은 검사 길이에서 동일한 수준의 정확성을 확보할 수 있다[11][12].

국외에서는 세계적인 검사 회사인 ETS (Educational Testing Service)에서 주관하는 다수의 시험(예를 들어, GRE(Graduate Record Examination), TOEFL(Test Of English as a Foreign Language) 등)에 CAT을 적용하여 실시 중이다. 또한 USMLE(United States Medical Licensing Examination), NCLEX-RN (National Council Licensure Examination for Registered Nurses)과 같은 국가시험에도 CAT을 도입하고 있다. 국내에서도 다수의 학자들이 CAT을 도입하기 위한 연구를 진행하였으나[13]-[15], 아직 그 실제적 활용은 미미한 수준이다. 다양한 분야에서 CAT 활용의 필요성이 대두되고 있음에도 실제 국내에서 활용 가능한 CAT 플랫폼이 존재하지 않는 실정이다. 이에 본 연구는 CAT의 구현이 가능한 새로운 CAT 플랫폼을 개발하고, 개발한 플랫폼의 정확성 및 효율성을 검증하였다.

본 논문의 구성을 다음과 같다. 2장은 CAT 관련 연구를 통해 CAT에 관해 간략히 설명하고 국내에서 이루어진 연구에 대해 정리하였다. 3장에서는 CAT 구현을 위해 개발한 플랫폼의 개발 과정 및 플랫폼을 활용한 CAT의 구현 과정에 대해 설명하였다. 4장에서는 개발한 CAT 플랫폼의 정확성과 효율성을 지필검사와 비교하여 검증하였다. 동일한 검사를 기반으로 지필검사와 CAT의 결과를 비교하고 해당 플랫폼을 이용한 피검자들을 대상으로 사용성에 대한 설문조사를 실시하였다. 마지막으로 5장에서는 본 연구의 결론에 관해 기술하고 본 연구가 갖는 한계점 및 향후 연구에 대하여 논의하였다.

## Ⅱ. 관련 연구

CAT은 피검자가 가진 각각의 개별적인 능력 수준에 맞게 적응적으로 평가하는 검사시스템이다. CAT의 가장 큰 목적은 보다 효율적이고 정확하게 피검자의 능력을 측정하는 것이다[16]. 보다 적은 문항의 수를 통해 보다 정확하게 피검자의 능력점수를 산출하는 것이 CAT을 활용하는 이유이다. 피검자의 능력수준에 가장 적절한 문항이 제공되는 경우에 해당 피검자의 능력을 가장 정확히 파악할 수 있다[17]. CAT의 경우 피검자는 자신의 능력에 맞게 선별된 문항을 풀게 되므로 보다 정확하게 능력을 추정할 수 있다. 또한, 자신의 능력 수준보다 너무 어렵거나 너무 쉬운 문항을 풀지 않아도 되며, 각각의 피검자마다 다른 문항을 풀게 된다. 즉,

CAT을 활용할 경우 모든 피검자가 고정된 문항의 시험지에 응시하는 것보다 효율적인 시험의 운영이 가능해진다.

CAT의 실행을 위해서는 다섯 가지의 구성요소가 필수적으로 요구된다[7]. 첫 번째는 문항은행(Item bank)이다. 문항은행은 단순히 검사에 사용되는 문항들의 집합이 아닌 각 문항에 대한 여러 정보들을 체계적으로 관리하기 위한 시스템을 의미한다[18]. CAT를 활용하기 위해서는 아주 큰 규모의 문항은행이 필요하다. 두 번째, 시작 문항의 설정이다. CAT은 각각의 응답에 따라 즉각적으로 피검자의 능력 수준을 산출하여 다음 문항을 선택한다. 그렇다면 아직 해당 피검자의 응답이 없는 경우 어떻게 문항을 선택할 것인가? 이를 지정하는 것이 바로 시작 문항의 설정이다. 시작문항을 선정하는 방법은 중간 수준의 문항을 제공하거나 무작위로 아무 문항이나 제공하는 등 다양하다. 세 번째는 문항 선정 방법이다. CAT에서의 문항 선정 방법은 대부분 문항반응이론(Item response theory)에 기반을 두고 있다. 문항반응이론을 기반으로 하여 산출한 문항의 각 모수를 이용하여 문항정보를 계산하고 주어진 능력수준에서 가장 높은 문항정보를 제공하는 문항을 선정한다. 네 번째, 피검자의 능력모수 추정방법이다. 피검자의 능력모수를 추정하는 방법에는 여러 통계적 기법이 있으며, 연구자는 그중 어떤 방법을 사용할 것인지 결정해야 한다. 마지막으로 종료 규칙(Termination criterion)이 필요하다. 어떤 것을 기준으로 검사를 마칠 것인지 결정하는 것으로 검사의 목적에 따라 달라질 수 있다. 종료 규칙에는 문항의 수를 기준으로 하는 방법, 측정 오차를 기준으로 하는 방법 등 여러 방법이 존재한다.

언급된 다섯 개의 구성요소를 바탕으로 CAT의 알고리즘을 간략히 설명하면 다음과 같다. 먼저 설정된 시작 문항의 규칙에 따라 문항은행에서 첫 번째 문항이 제시가 된다. 그 후 문항 선정 방법에 따라 그다음 문항들이 제시된다. 각 문항에 대한 피검자의 응답이 완료된 즉시 피검자의 능력모수가 추정된다. 피검자의 능력에 맞는 문항이 제시되고, 피검자의 능력모수가 추정되는 단계가 종료 규칙을 만족할 때까지 반복된다.

이 외에도 CAT을 실행하기 위해서는 다양한 고려 대상들이 존재한다. 검사의 제한시간, 실시 장소 등과 같은 일반적인 문제부터 문항은행의 보안, 적절한 시스템 디스플레이와 인터페이스, 데이터베이스 등이 필요하다. 또한 CAT을 구현하기 위한 플랫폼, 시스템의 성능 등도 고려사항이다.

CAT이 지닌 여러 이점으로 인해 CAT의 적용이 확산될 것이라는 전망은 십수 년 전부터 제기되었다[15][19]. 외국에서는 이미 다수의 모의연구 및 경험연구가 진행되었고, CAT을 구현하기 위한 플랫폼도 개발되고 있다[20]. 대표적으로 Concerto 플랫폼[21], CAT-MD[22] 등의 플랫폼이 개발되었다. Concerto는 R 패키지 catR[23]과 HTML-템플릿 기반 drag-and-drop 에디터를 사용하는 상용 플랫폼이다. Concerto는 다양한 문항반응이론 모형과 문항선정 알고리즘을 제공하나 사용이 복잡하고 어렵다. CAT-MD는 아도비 플래시 기술을 이용한 모바일 기반 플랫폼이다. CAT-MD는 선다형 문제와 Rasch 모형만 제공하는 한계가 있어 사용이 제한적이다.

국내에서도 CAT의 활용에 대한 관심이 증가하고 있으며 이에 따라 CAT에 대한 연구도 꾸준히 진행되고 있다. 현재까지 국내에서 진행된 연구들은 대부분 실제 현장에서의 적용과 검증이 포함되지 않은 모의연구로 진행되었다. CAT의 활용에 대한 리뷰 연구[24], 이론적 비교를 위한 시뮬레이션 연구[25], 실제 CAT 프로그램을 탑재하지 못한 채 진행된 연구[26][27], CAT의 필요성 고찰에 대한 연구[28][29] 등이 주를 이루고 있다. 국내에서 개발된 CAT 플랫폼으로 IRT-CAT[15]이 PHP기반으로 사용되고 있으나 의학교육 시험에 적합하게 개발되어 다른 시험에 적용하기에는 유용성이 낮고 모의실험이나 경험연구의 부족으로 타당성이 검정되지 않았다.

## III. CAT 플랫폼(LIVECAT) 개발

### 3.1 LIVECAT 개발

CAT 플랫폼의 개발을 위해 사용된 도구는 표 1과 같다. 개발된 플랫폼은 LIVECAT으로 명명하였고 상업용으로 더캣코리아 홈페이지(http://www.thecatkorea.com)에서 구매가 가능하다.

표 1. 프로그램 개발에 사용된 도구
Table 1. Tools used for programming

| Classification | Tool |
|---|---|
| Web server | apache-tomcat-8.5.40 |
| DBMS | SQL server 2008 R2 |
| Languages | Java-1.8.0-openjdk<br>Highcharts<br>Jquery<br>Javascript<br>Springframework version 4.3.3<br>Springframework.security version 5.0.7 |
| OS | Window 10 |

기존의 개발된 CAT 플랫폼의 한계점을 극복하기 위해 본 연구에서는 편파(Bias)가 가장 낮은 채점(Scoring)방식인 최대우도추정(Maximum likelihood estimation)방식을 적용하였다. 또한 실제 피검자들을 대상으로 실험을 실시하여 알고리즘의 정확성과 효율성을 검증하였다. 기존에 개발된 플랫폼은 모든 문항을 맞추거나 틀릴 경우 채점이 불가능하기 때문에 최대우도추정방식대신 평균사후추정(Expected a posteriori)방식을 사용하였다. 그러나 LIVECAT은 검사진행시 모든 문항을 다 맞추거나 다 틀려도 최대우도추정방식으로 채점이 가능한 알고리즘을 장착하여 더 높은 정확성을 확보하였다. 특히 다양한 문항반응이론 모형의 적용이 가능하고 채점의 결과도 진점수와 T점수 등으로 변환하여 제공할 수 있게 구현되었다.

LIVECAT의 개발에 있어서 가장 중점적으로 고려한 점은 비전문가도 쉽게 CAT을 활용한 검사의 제작 및 시행이 가능하도록 하는 점이다. 이를 위해 관리자와 피검자의 조작이 쉽도록 직관적으로 화면을 구성하였다. 개발된 플랫폼을 활용하여 CAT을 활용한 검사의 제작 및 시행의 과정은 그림 1과 같은 단계로 진행된다.

먼저 검사에 사용할 문항을 중심으로 문항은행을 구축한다. 문항은행의 구축을 위해 플랫폼에 사용할 문항을 등록해야 한다. 이때 해당 문항의 분류, 예비문항 여부, 문항모수, 문항의 형태, 정답 여부 등의 정보를 입력한다. 그림 2는 문항 등록 창의 일부이다. 문항을 등록할 때, 문항의 ID는 해당 문항의 분류 등을 고려하여 시스템에서 자동으로 부여된다. 이는 문항의 ID가 겹치지 않게 하고, 문항의 체계적 관리를 돕기 위한 과정이다.

필요한 문항을 모두 추가하고 나면 하나의 문항은행이 만들어진다. 해당 플랫폼을 활용할 시 문항은행의 구축 및 관리가 보다 용이하다. 문항은행의 모습은 그림 3과 같다. 구축된 문항은행에서 문항의 분류, 문항모수의 추정에 사용된 모형, 난이도 모수 등을 통한 문항의 검색이 가능하다.


그림 2. LIVECAT 화면 중 문항 등록 창
Fig. 2. Screen shot of item registration

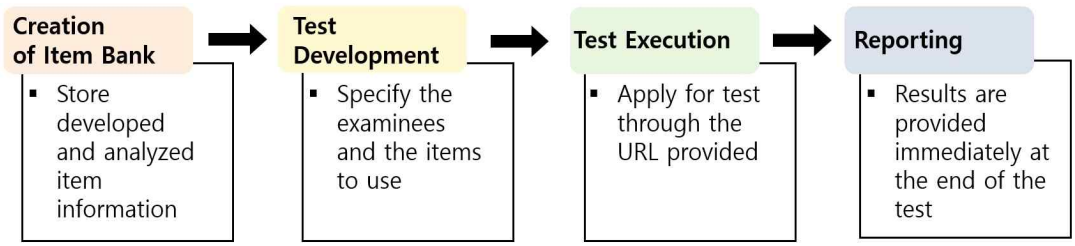


그림 1. LIVECAT을 통한 검사 제작 및 진행 과정
Fig. 1. Process of a test using the LIVECAT


그림 3. LIVECAT 화면 중 문항은행 창
Fig. 3. Screen shot of item bank

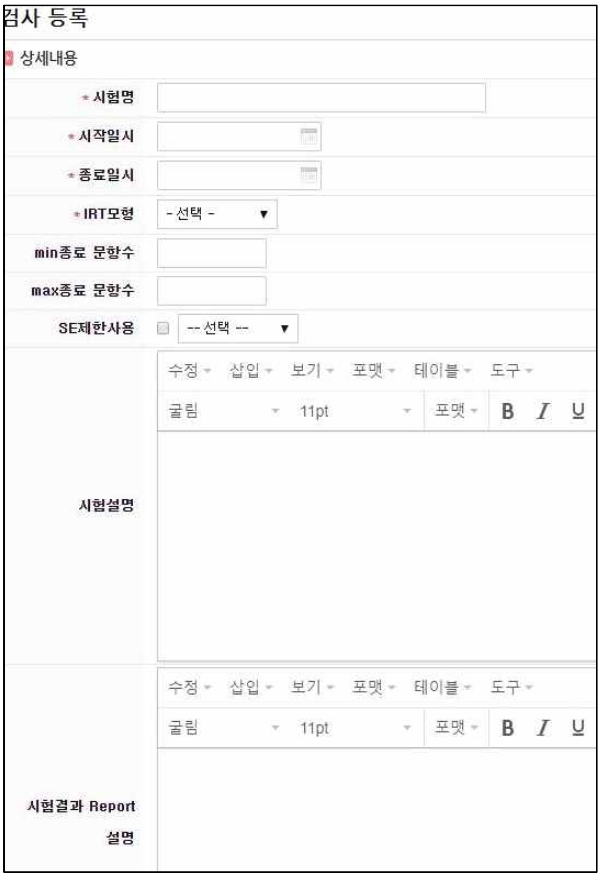


그림 4. LIVECAT 화면 중 검사 등록 창
Fig. 4. Screen shot of test registration

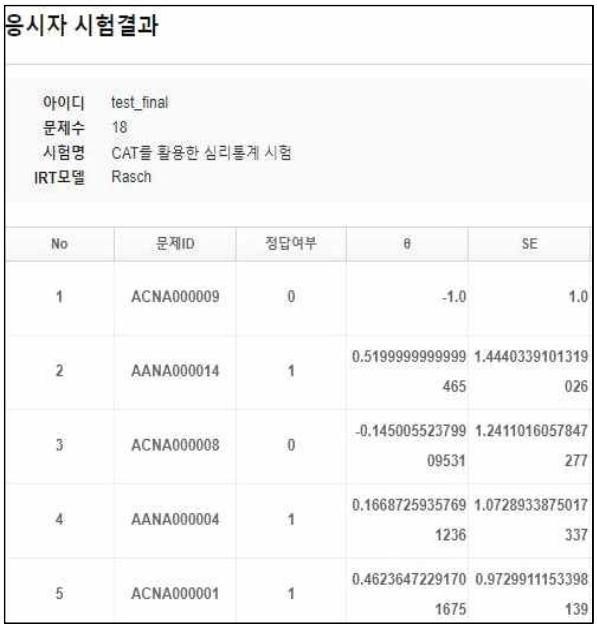


그림 5. LIVECAT 화면 중 개인 피검자 결과 예시
Fig. 5. Screen shot of the individual result page

문항은행의 구축이 완료된 후, 원하는 검사를 제작할 수 있다. 검사의 제작을 위해 해당 검사의 이름, 시작일시 및 종료일시, 검사종료규칙, 검사에 대한 설명, 결과 제공 방법 등의 정보를 입력해야 한다. 또한 해당 검사에 사용할 문항과 응시자격을 부여할 피검자를 지정해준다. 그림 4는 검사 등록 창의 일부이다. 제작된 검사는 연결된 URL을 통해 시작일시부터 종료일시까지 등록된 피검자만 응시가 가능하다.

검사가 종료된 직후, 피검자와 검사 관리자 모두 검사의 결과를 확인할 수 있다. 그림 5는 특정 개인의 검사 결과 예시이다. 관리자는 각 개인의 응답 문항, 정답여부, 능력모수 추정치 등을 확인할 수 있다.

## 3.2 문항반응이론의 적용

개발된 플랫폼은 문항반응이론 모형의 적용이 가능하도록 설계되었다. 문항반응이론이란 검사 이론(Test theory)의 최신 이론으로 고전검사이론의 대안으로 대두되었다[13]. 문항반응이론 모형은 각각의 개별 문항에 대한 피검자들의 반응을 비선형 함수로 나타낸 것이다. 피검자의 능력수준에 따라 문항에 대한 반응의 확률을 함수로 표현하기 때문에 검

사수준의 분석뿐만 아니라 문항수준의 분석이 가능하다.

문항반응이론에서는 각 문항에 대한 난이도(Difficulty), 변별도(Discrimination), 추측도(Guessing) 모수를 추정할 수 있다. 문항에 대한 모수를 몇 가지 추정하느냐에 따라 모형을 구분한다. 문항의 난이도모수는 응답확률이 50%인 지점을 기준으로 결정되는데, 값이 클수록 문항이 어렵다고 해석할 수 있다. 문항의 변별도모수의 경우 응답확률이 50%인 지점의 기울기를 통해 계산된다. 변별도모수가 높을수록, 즉 기울기가 가파를수록 능력수준의 정도에 따른 응답확률이 크게 변화함을 의미한다. 마지막으로 추측도의 경우 피검자의 능력수준과 관계없이 해당 문항을 맞출 확률을 의미한다.

$$P_i(u_i=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad (1)$$

식 (1)은 문항 $i$, 피검자 $j$에 대한 3모수 모형의 문항반응함수이다[30]-[32]. 즉, 문항의 난이도, 변별도, 추측도를 모두 고려하였을 때, $\theta_j$의 능력 수준을 지닌 피검자가 정답을 고를 확률에 대한 수식이다. 여기서 $\theta_j$는 피검자 $j$에 대한 능력모수를 나타내고, $a_i$는 문항 $i$의 변별도, $b_i$는 난이도, $c_i$는 추측도를 의미한다. 해당 식은 정규오자이브모형을 활용한 기존의 식[33]과는 달리 로지스틱 모형을 활용하기 때문에 계산이 더 용이하다. 또한 3모수 모형의 함수를 알고리즘에 적용시켰기 때문에 3모수 모형뿐만 아니라 1모수 모형과 2모수 모형의 적용도 가능하다. 여기서 상수 $D$는 정규오자이브 모형을 로지스틱 모형으로 옮기는 과정에서 생기는 오차를 줄여주기 위해 사용되는 값으로 1.702로 계산한다.

식 (1)을 LIVECAT에 적용할 경우, 모형에서 사람의 능력 $\theta_j$와 문항의 난이도 $b_i$가 같은 척도에서 직접 비교가 가능하기 때문에 적은 문항수로 CAT을 구현할 수 있게 되어 검사의 효율성이 높아진다. 능력모수가 난이도 모수보다 높으면 피험자가 문항을 맞출 확률이 높아지고, 능력모수가 난이도 모수보다 낮으면 문항을 맞출 확률이 낮아지게 된다. 따라서 식 (1)을 기반으로 피험자의 능력이 높아 문항을 맞출 시 즉시 피험자의 능력이 높게 추정되고 다음문항의 난이도는 피험자의 능력모수에 맞는 높은 난이도의 문항이 출제된다. 이러한 과정을 통해 빠른 시간 내에 피험자를 정확하게 측정하고 검사가 중지하게 된다.

$$P_i(u_i=1|\theta_j,b_i)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \quad (2)$$

식 (2)는 문항반응이론 중 $a_i$=1, $c_i$=0인 Rasch 모형[30][34]으로 LIVECAT에 적용한 모형이다. 아래와 같이 Z= ($\theta_j$- $b_i$) 점수를 넣어 식 (3)에 대입하여 로그함수를 취하면 위에서 설명한 것처럼 능력모수와 난이도 모수의 일대일 비교가 가능한 선형함수가 됨이 증명된다.

$$e^{Z_i}=\frac{P_i(\theta_j)}{1-P(\theta_j)}=\frac{P_i(\theta_j)}{Q_i(\theta_j)} \quad (3)$$

$$Z_i=\log\left(\frac{P_i(\theta_j)}{Q_i(\theta_j)}\right)=\theta_j-b_i$$

Rasch 모형은 측정모형 중 하나로 데이터가 모형에 적합다는 가정 하에 사용하고 모수추정값이 안정적(Consistent), 불편차(Unbiased), 그리고 충분통계치이다[34].

## Ⅳ. LIVECAT의 정확성 및 효율성 검증

### 4.1 연구 설계

개발된 CAT 플랫폼(LIVECAT)의 정확성 및 효율성의 검증을 위해 총 두 단계로 이루어진 실험을 진행하였다. 첫 번째 단계에서 피검자는 [심리통계] 쪽지시험과 LIVECAT 사용평가 설문지, 총 두 가지 검사에 참여하게 된다. 여기서 [심리통계] 쪽지시험은 LIVECAT을 통해 진행된다. 첫 시험에 응시하고 2주가 지난 후 지필검사를 통해 동일한 문항으로 이루어진 쪽지시험에 다시 응시한다.

본 실험의 목적은 다음과 같다. 첫 번째, LIVECAT

을 활용한 시험에서의 점수와 지필검사의 점수간의 상관계수의 산출을 통해 LIVECAT의 정확성을 평가한다. 높은 양의 상관계수를 지닐수록 LIVECAT의 정확성이 높다고 판단할 수 있다. 두 번째, LIVECAT을 활용한 시험에서 사용된 문항 수의 평균이 전체 문항 수와 비교하여 얼마나 줄어들었는지 계산하여 LIVECAT의 효율성을 검증한다. 세 번째, LIVECAT을 실제 사용해본 피검자들로부터 수집한 설문조사의 결과를 바탕으로 사용에 대한 만족도를 평가한다.

### 4.2 연구 대상 및 도구

해당 실험은 2019년 2학기 H 대학교 [심리통계] 수강생을 대상으로 실시되었다. 총 154명의 피검자가 참여하였으며, 그중 모든 검사에 응시한 141명의 자료를 분석에 사용하였다. 실험에 참여한 인원 중 60%가 넘는 인원이 이전에 컴퓨터를 통한 시험에 응시해본 경험이 1회 이상 있었다. 1차 검사와 2차 검사 모두 H 대학교 전공 강의실에서 진행되었다. 1차 검사의 경우 각 피검자가 소지하고 있는 스마트기기(노트북, 스마트폰 등)를 활용하여 실시하였다. LIVECAT을 활용한 검사 실시에 앞서 플랫폼 화면의 모습이 기록된 발표 자료를 통해 LIVECAT 사용 방법에 대한 간략한 설명을 진행하였다. 피검자는 쪽지시험이 종료된 직후 제공되는 화면을 통해 자신의 시험 결과를 확인할 수 있다. 또한, 쪽지시험이 종료된 직후 네이버폼을 통해 LIVECAT 사용평가 설문에 대한 응답을 진행하였다. 두 번째 시험은 일정 시간이 지난 후 연구자가 준비한 지필시험지를 통해 다시 실시되었다.

해당 실험에서 활용된 문항반응이론 모형은 Rasch 모형으로, 관련된 모든 분석은 Winsteps 4.0.1 프로그램을 사용하였다. Rasch 모형이란 문항반응이론 모형 중 하나로 식 (1)에서 변별도를 1로, 추측도를 0으로 고정한 상태에서 문항의 난이도모수만을 추정하는 모형이다.

실험에 사용된 문항의 종류는 쪽지시험 문항, 사용평가 설문문항 두 가지이다. 사용평가 설문문항의 경우 13개의 객관식 문항과 두 개의 주관식 문항으로 구성되어 있다. 쪽지시험에 사용된 문항은 총 30개로 [심리통계] 수업 관련 성취도 검사 문항이다. 해당 문항은 2018년 H 대학교 [심리통계] 수강생 162명을 대상으로 수집한 응답을 토대로 Rasch 모형을 적용하여 문항의 난이도를 분석하였다. 문항의 난이도는 최고 1.89에서 최저 -1.96이었으며, 평균 .00, 표준편차 .94이다. 각 문항에 대한 난이도모수의 추정값은 표 2에 기술되어 있다.

첫 번째 쪽지시험에 적용된 LIVECAT 알고리즘은 다음과 같다. 활용되는 문항은행에는 Rasch 모형에 따른 난이도를 문항 특성으로 입력하였고 별도의 내용 균형(Content balancing)은 고려하지 않았다. 문항은행에 저장된 문항은 총 30개이다. 시작 문항의 경우 별도의 알고리즘 없이 무작위로 제공되었으며, 무작위 제공 문항의 수는 다섯 개로 지정하였다. 여섯 번째 문항부터는 직전에 추정된 피검자의 능력수준과 가장 유사한 난이도의 문항이 출제되었다. 피검자의 능력모수는 최대가능도 방법(Maximum likelihood methods)을 통해 추정하였다. 해당 시험은 피검자의 능력 추정에 대한 표준오차의 값이 .5 이하가 되거나, 피검자가 문항은행에 포함된 30개의 문항을 모두 풀게 될 시 종료된다.

표 2. 쪽지시험 문항 난이도
Table 2. B-parameter of items in the quiz

| Item | *b* | Item | *b* | Item | *b* |
|---|---|---|---|---|---|
| 1 | -0.60 | 11 | 0.88 | 21 | -0.93 |
| 2 | 0.72 | 12 | -1.96 | 22 | -0.52 |
| 3 | 1.89 | 13 | -0.74 | 23 | -0.69 |
| 4 | 0.99 | 14 | 0.17 | 24 | 0.43 |
| 5 | 0.17 | 15 | -1.15 | 25 | 0.11 |
| 6 | 0.24 | 16 | 1.84 | 26 | -1.27 |
| 7 | 0.20 | 17 | -0.10 | 27 | -0.32 |
| 8 | 0.66 | 18 | 0.43 | 28 | -1.21 |
| 9 | 0.93 | 19 | -1.54 | 29 | 0.52 |
| 10 | 1.40 | 20 | -0.14 | 30 | -0.40 |

### 4.3 결과

모든 실험이 종료 된 후, 각 쪽지시험 응답을 토대로 사람모수를 산출하였다. 즉, 사람모수의 경우 하나의 동일한 집단에 대해 총 세 가지 방식으로 산출된 값이 존재한다. 첫 번째는 LIVECAT을 통해 계산된 값, 두 번째는 LIVECAT을 통해 수집된 응

답을 바탕으로 다시 산출한 값, 마지막으로 모든 문항에 응답하여야 하는 지필검사를 통해 수집한 응답을 바탕으로 산출한 값이 있다. 사용평가 설문 문항의 경우 빈도분석을 실시하였다. 그 후, 각 조건에 대해 추정된 사람모수 간의 상관계수를 산출하였다.

먼저 개발된 LIVECAT 내의 알고리즘이 능력모수를 제대로 측정하는지 알아보기 위해 LIVECAT이 계산한 능력모수의 값과 해당 응답을 다시 Winsteps 4.0.1을 활용하여 계산한 값 사이의 상관계수를 계산하였다. 그 결과 .993($p<.001$)으로 통계적으로 유의미하게 매우 높은 정적 상관이 있음을 검증하였다. 두 번째로 LIVECAT이 계산한 능력모수와 지필검사 응답을 기반으로 계산된 능력모수 간의 상관계수가 .779($p<.001$)로 유의미하게 높은 정적 상관이 있었다. 또한 지필검사의 총점과 LIVECAT을 통해 산출된 능력모수 간의 상관계수도 .786($p<.001$)로 유의미하게 높은 정적 상관이 있었다. 이를 통해 LIVECAT 내의 능력모수를 추정하는 알고리즘이 높은 정확성을 가지고 있음을 확인하였다.

피검자들은 LIVECAT을 활용한 쪽지시험에서 평균 20.12개의 문항을 풀었으며 첫 문항부터 마지막 문항에 응답하기까지 평균 약 22분 51초가 소요되었다. 또한 능력모수의 추정에 있어 LIVECAT을 활용한 시험에서 지필검사보다 표준오차가 약 .03 낮게 추정되었다. 즉, LIVECAT을 활용할 경우, 전체 시험 길이보다 약 33% 짧은 길이의 시험을 통해 전체 시험 길이와 .779의 상관을 가진 값을, 더 낮은 표준오차로 추정할 수 있음을 확인하였다.

표 3. 점수 간 상관
Table 3. Correlation between each scores

| | A | B | C | D |
|---|---|---|---|---|
| A | 1 | .993** | .786** | .779** |
| B | | 1 | .791** | .790** |
| C | | | 1 | .960** |
| D | | | | 1 |

A: LIVECAT이 추정한 능력모수
B: LIVECAT 응답을 기반으로 재추정한 능력모수
C: 지필검사 총점
D: 지필검사 응답을 기반으로 산출한 능력모수
** $p < .001$

LIVECAT 사용평가 설문 결과, 대부분의 학생이 CAT에 대해 긍정적인 태도를 취하고 있음을 확인할 수 있었다. “일부 과목에서도 컴퓨터 기반 맞춤형 검사와 같은 형태의 시험으로 치렀으면 좋겠다.”에 응답한 피검자가 전체 중 70%가 넘었으며, 약 58%의 피검자가 해당 시험에 응시할 때 불편함이 없었다고 응답하였다. 하지만 대부분의 피검자가 CAT의 원리에 대한 이해의 수준이 낮았다(“나는 오늘 경험한 컴퓨터 기반 맞춤형 검사의 원리에 대해 거의 모른다. 혹은 전혀 모른다.”에 응답한 인원이 약 65%). 그로 인해 약 14%의 피검자가 “이번 컴퓨터 기반 맞춤형 검사는 내 실력을 평가하기에 문제 수가 너무 부족하다고 생각한다.”에 “그렇다.”고 응답하였다. 또한 “나는 이번 컴퓨터 기반 맞춤형 검사에서 기존의 지필검사보다 낮은 점수를 받을 것 같다.”고 응답한 피검자가 약 45%였다. CAT를 활용한 시험에 대해 좋았던 점을 자유롭게 기술하는 주관식 문항에는 시험이 간편하고, 시험 종료와 함께 점수를 확인할 수 있어 좋았다는 점 등의 의견이 많았다. 마지막으로 개선해야 할 점을 자유롭게 기술하는 문항에서는 글씨 크기, 속도, 답안 체크의 불편함 등에 대한 내용이 있었다.

## Ⅴ. 결론 및 향후 과제

IT의 눈부신 발전으로 4차 산업 사회에서는 모든 분야에서 IT를 빼놓고 이야기하기 어려운 실정이다. 이러한 변화에 발맞춰 검사와 교육 현장에서도 더 정확하고 효과적인 측정 방법을 찾기 위해 IT와의 결합방안에 대한 연구가 지속적으로 진행되고 있다. 이러한 현장의 요구로 인해 자연스럽게 CAT에 대한 요구와 필요성도 점점 더 높아지고 있다. 하지만 여전히 국내에서는 CAT을 교육이나 선발 현장에 적용한 사례가 부재하고 적용이 제한적인 분야에 국한되고 있다.

따라서 본 논문에서는 CAT의 이론적인 부분보다는 실용적인 부분에 초점을 맞추어 연구를 진행하였다. 단순히 CAT에 대한 연구를 위한 플랫폼이 아닌 지속적으로 현장에서 사용 가능하고 여러 분야에서 적용이 가능한, 누구나 쉽게 이용할 수 있는 플랫폼을 개발하고자 하였다. 본 연구는 개발 자체

에 머물지 않고 개발한 LIVECAT을 실제 현장에 적용시켜 타당성을 검증하였다. 이를 통해 LIVECAT을 현장에서 사용하더라도 정확성에 문제 없이 CAT이 가지는 장점을 그대로 구현함을 증명하였다.

이전에 실시된 여러 연구에서 이미 CAT의 높은 정확성과 효율성이 검증되어 왔다[7][16]. 그에 비해 본 연구에서 진행한 경험적 결과는 기존의 논문보다 낮은 정확성(CAT 활용 검사와 지필검사 간의 낮은 상관)을 보이는 문제가 있었다. 이는 본 연구에서 사용한 문항은행의 규모가 작고, 사용된 문항의 난이도의 분포가 고르지 않다는 문항은행 자체의 한계점 등 때문이지 알고리즘의 문제는 아니었다. 따라서 다양한 문항을 포함하고 있는 문항은행을 구성하여 LIVECAT을 이용한 경험적 연구가 향후 필요하다. 더불어 본 연구에서 검증한 Rasch 모형을 적용한 이분문항 외에도 다분문항 등을 포함한 다양한 검사에서의 LIVECAT 사용의 검증이 필요할 것으로 보인다. 더 나아가 LIVECAT과 기존의 플랫폼을 동일한 조건(같은 모형, 같은 채점방식)에서의 비교를 통해 LIVECAT의 우수성을 경험적으로 검정할 수도 있을 것이다. 또한 LIVECAT 사용 평가 설문에서 개선할 점으로 지적 받은 부분을 실제적으로 개선하는 것이 본 연구의 향후 과제가 될 것이다.

## References


[1] E. J. Lim et al., "Development of Computer-Based Test (CBT) and Student Recognition Survey on CBT", Korean Journal of Medical Education, Vol. 20, No. 2, pp. 145-154, Jun. 2008.

[2] S. L. Wise and B. S. Plake, "Research on the effects of administering test via computers", Educational Measurement: Issues and Practice, Vol. 3, No. 3, pp. 5-10, Sep. 1989.

[3] T. J. Seong, "CBT(Computer-Based Test) and CAT(Computerized Adaptive Test)", Journal of Education Evaluation, Vol. 5, No. 1, pp. 73-97, Apr. 1992.

[4] S. N. Jung, "Development and Validation of Computer-based Test for Young Children's Computer Application Ability", Doctoral dissertation, The Graduate School, Chonnam University, Feb. 2007.

[5] S. Huh, "Computer-Based Testing and Construction of an Item Bank Database for Medical Education in Korea", Korean Medical Education Review 2014, Vol. 16. No. 1, pp. 11-15, Feb. 2014.

[6] S. G Back and S. H. Chae, "Computerized Adaptive Testing", Wonmisa, Aug. 1998.

[7] D. G. Seo, "Overview and current management of computerized adaptive testing in licensing/ certification examinations", Journal of educational evaluation for health professions, Vol. 14, Jul. 2017. doi: 10.3352/jeehp.2017.14.17.

[8] C. M. Beleckas, A. Padovano, J. Guattery, A. M. Chamberlain, J. D Keener, and R. P. Calfee, "Performance of Patient-Reported Outcomes Measurement Information System (PROMIS) upper extremity (UE) versus physical function (PF) computer adaptive tests (CATs) in upper extremity clinics", The Journal of hand surgery, Vol. 42, No. 11, pp. 867-874, Jul. 2017.

[9] E. M. Gamper et al., "Development of an item bank for the EORTC role functioning computer adaptive test (EORTC RF-CAT)", Health and quality of life outcomes, Vol. 14, No. 1, pp. 72, May 2016.

[10] S. Morris, M. Bass, M. Lee, and R. E. Neapolitan, "Advancing the efficiency and efficacy of patient reported outcomes with multivariate computer adaptive testing", Journal of the American Medical Informatics Association, Vol. 24, No. 5, pp. 897-902, Mar. 2017.

[11] N. A. Thompson and D. A. Weiss, "A framework for the development of computerized adaptive tests", Pract Assess Res Eval 2011, Vol. 16, pp. 1-9, Jan. 2011.

[12] D. J. Weiss, "Computerized adaptive testing for effective and efficient measurement is counseling

and education", Meas Eval Couns Dev 2004, Vol. 37, pp. 70-84, Jul. 2004.

[13] Y. H. Kim, H. T. Jeong, H. J. Cho, J. C. Cha, and J. Y. Lee, "The Devising and Utilizing Method of Web Based Compound CAT Model", Journal of Educational Research, Vol. 14, pp. 97-124, Jan. 2004.

[14] I. C. Choi, "Development and validation of a computer adaptive language testing algorithm", Language research, Vol. 40, No. 1, pp. 227-252, Mar. 2004.

[15] Y. H. Lee, J. H. Park, and I. Y. Park, "Estimation of an examinee's ability in the web-based computerized adaptive testing program IRT-CAT", Journal of Educational Evaluation for Health Professions, Vol. 3, Nov. 2006.

[16] Y. H. Kim, M. Son, and H. T. Jeong, "The Principle and Practice of Computerized Adaptive Testing", Muneumsa, Dec. 2002.

[17] H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, and R. J. Mislevy, "Computerized adaptive testing: A primer", Routledge, Jun. 2000.

[18] F. B. Baker, "Item banking in computer-based instructional systems", Applied psychological measurement, Vol. 10, No. 4, pp. 405-414, Dec. 1986.

[19] M. D. Reckase, "Adaptive testing: The evolution of a good idea", Educational Measurement: Issues and Practice, Vol. 8, No. 3, pp. 11-15, Sep. 1989.

[20] S. Oppl, F. Reisinger, A. Eckmaier, and C. Helm, "A flexible online platform for computerized adaptive testing", International Journal of Educational Technology in Higher Education, Vol. 14, No. 2, Jan. 2017.

[21] K. Scalise and D. D. Allen, "Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform", British Journal of Mathematical and Statistical Psychology, Vol. 68, No. 3, pp. 478-496, Nov. 2015.

[22] E. Triantafillou, E. Georgiadou, and A. A. Economides, "CAT_MD: Computerized adaptive testing on mobile devices", International Journal of Web-Based Learning and Teaching Technologies(IJWLTT), Vol. 3, No. 1, pp. 13-20, May. 2008.

[23] D. Magis and G. Raiche, "Random generation of response patterns under computerized adaptive testing with the R package catR", Journal of Statistical Software, Vol. 48, No. 8, pp. 1-31, May. 2012.

[24] B. S. Choi, "Introducing an Online Measurement System Using Item Response Theory and Computer Adaptive Testing Methods for Measuring the Physical Activity of Community-Dwelling Frail Older Adults", Phys Ther Korean, Vol. 26, No. 3, pp. 106-114, Sep. 2019.

[25] E. Lim, "Comparison of item selection algorithms for test security in computerized adaptive testing", Journal of Education Evaluation, Vol. 27, pp. 955-982, Sep. 2015.

[26] S. W. Na, G. S. Kim, and Y. S. Choi, "Computerized Adaptive Testing Using Bayesian Networks", Jounal of KIISE, Vol. 39, No. 6, pp. 497-506, Jun. 2012.

[27] Y. H. Kim, M. Son, Y. S. Choi, and H. T. Jeong, "Effect of the Time and Confidence Each Test Item on the Decision of SPRT-Typed Computerized Adaptive Testing", Journal of Korean Association for Educational Information and Media, Vol. 7, No. 2, pp. 83-110, Jun. 2001.

[28] S. Huh, "Can computerized adaptive tests be introduced to the Korean Medical Licensing Examination?", Journal of the Korean Medical Association, Vol. 55, No. 2, pp. 124-130, Feb. 2012.

[29] S. Huh, "Application of Computerized Adaptive Testing in Medical Education", Korean J Med Educ, Vol. 21, No. 2, pp. 97-102, Jun. 2009.

[30] F. M. Lord, and. M. R. Novick, "Statistical Theories on Mental Test Scores", IAP, pp.

397-424, Apr. 1968.

[31] F. B. Baker, "The Basics of Item Response Theory", ERIC Clearinghouse on Assessment and Evaluation, Oct. 2001.

[32] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, "Fundamentals of item response theory". Sage, Jul. 1991.

[23] F. Lord, "A theory of test scores", Psychometric monographs, No. 7, 1952.

[34] D. Andrich, "Rasch models for measurement", Sage, Mar. 1988.

## 저자소개

김 도 경 (Do-Gyeong Kim)

2017년 2월 : 한림대학교 심리학과 졸업(문학학사)
2019년 2월 : 한림대학교 일반대학원 심리학과 졸업 (문학석사)
2019년 3월 ~ 2020년 3월 : ㈜더캣코리아 연구원
관심분야 : 심리측정, 컴퓨터 맞춤형 검사, 검사개발, 문항반응이론

서 동 기 (Dong-Gi Seo)

2002년 2월 : 성균관대학교 산업심리학과 졸업(문학학사)
2004년 2월 : Middle Tennessee State University 심리학과 졸업(문학석사)
2010년 2월 : University of Minnesota Twin Cities 심리학과 졸업(철학박사)
2015년 ~ 현재 : 한림대학교 심리학과 조교수
2019년 ~ 현재 : ㈜더캣코리아 대표
관심분야 : 컴퓨터 맞춤형 검사, 문항반응이론, 심리측정 및 검사