

상담성과 측정에서 단계적 적응형 검사의 적용 방안: 검사 설계 및 변화 탐지¹⁾

이진실²⁾ · 백수진³⁾ · 홍유정⁴⁾ · 조은이⁵⁾ · 최영원⁶⁾ · 정하나⁷⁾

요약

본 연구는 상담성과 측정에서 단계적 적응형 검사의 적용 가능성을 탐색하고 측정학적 특성을 평가하고자 하였다. 이를 위해 대표적인 상담성과 측정 도구인 CORE-OM과 OQ-30에 단계적 적응형 검사를 적용하였고, 모의실험을 통해 검사 설계 방식에 따른 측정 정확성을 비교하였다. 또한 단계적 적응형 검사를 활용한 상담성과 측정에서 신뢰할 수 있는 변화를 파악하는 방법을 적용하였다. 연구 결과, 응답자별로 맞춤형 소검사가 제시되는 단계적 적응형 검사는 동일한 문항 수 조건에서 고정형 검사보다 상대적으로 높은 정확성을 보였다. 또한 응답자의 특성이 매우 높거나 낮은 조건에서는 더 적은 수의 문항만으로 고정형 검사에 비해 더 높은 정확도를 보이는 경향이 나타났다. 아울러 단계적 적응형 검사를 적용할 경우 개인별 측정오차를 반영하여 변화 탐지가 가능함을 확인하였다. 이는 단계적 적응형 검사가 상담성과 측정의 효율성과 정확성을 동시에 고려할 수 있는 접근으로서의 가능성을 시사하는 결과로 볼 수 있다. 이러한 결과를 바탕으로 단계적 적응형 검사의 활용 가능성과 실제 현장 적용을 위한 향후 과제를 논의하였다.

주요용어 : 검사 설계, 단계적 적응형 검사, 변화 탐지, 상담성과

1. 서론

상담성과 측정은 상담의 효과성을 과학적으로 검증하고 상담 서비스 질 관리를 위한 정보를 제공할 수 있다는 점에서 중요한 부분이다(왕은자 외, 2018; Choi et al., 2010; Finch et al., 2001). 상담성과 측정의 중요성은 다양한 연구에서 강조되고 있으며, 특히 주기적 결과 모니터링(routine outcome monitoring and feedback) 또는 측정기반 치료(measurement-based care) 접근에서는 지속적인 측정과 피드백 제공이 상담 개입의 효과성을 높인다는 점도 확인되고 있다(Lambert et al., 2018; Lewis et al., 2019; Scott & Lewis, 2015).

1) 이 논문은 2025년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2025S1A5B5A17014042)

2) 제1저자(교신저자): 마음데이터랩, 대표. E-mail: leejinsil326@gmail.com

3) 제2저자: 서울대학교, 강사. E-mail: bsj104@gmail.com

4) 제3저자: 경기도교육연구원, 부연구위원. E-mail: dreamgo5049@gmail.com

5) 제4저자: 선문대학교 상담심리학과, 조교수. E-mail: eycho812@gmail.com

6) 제5저자: Beevelop 상담교육센터, 대표. E-mail: ian.yw.choi@gmail.com

7) 제6저자: Beevelop 상담교육센터, 연구원. E-mail: jeonghn9408@gmail.com

상담성과 측정을 위해 널리 활용되고 있는 대표적인 측정 도구로는 CORE-OM (Barkham et al., 1998; Barkham et al., 2013)과 OQ-30/45(Lambert, 2001)가 있다. CORE-OM은 영국에서 개발되었고 심리상담의 핵심적인 성과를 측정하는 34개 문항으로 구성되어 있다. OQ-30은 미국에서 개발되었으며 내담자의 증상과 함께 기능적인 측면까지 포함하는 30문항으로 구성되어 있다. CORE-OM과 OQ-30은 국내 타당화 연구도 이루어졌으며(김신경, 왕은자, 2016; 손난희, 유성경, 2012), 내담자의 핵심적인 기능이나 증상을 포괄적으로 측정하면서도, 특정한 이론에 한정되어 있지 않고, 양호한 신뢰도와 타당도를 보이는 도구라는 점에서 전 세계적으로 널리 활용되고 있다. 다만 상담 과정에서 주기적으로 활용하기에 문항 수가 적지 않은데, 특히 심리적으로 어려움을 겪고 있는 내담자에게 회기마다 30문항 이상의 검사에 참여하도록 하는 것은 부담으로 작용할 가능성이 있다(Barkham et al., 2013). 이러한 응답 부담은 검사 탈락이나 불성실 응답 가능성을 높여, 결과적으로 상담성과 측정의 질을 저해할 수 있다.

단계적 적응형 검사(이하 MST; multistage adaptive testing)는 맞춤형 문항을 제시함으로써, 모든 응답자에게 동일한 문항이 제시되는 고정형 검사(fixed-form test)에 비해, 더 적은 문항으로도 정확한 측정이 가능하다는 점이 밝혀져 왔다(Mead, 2006; Weiss, 2004). 최근 이러한 장점이 주목받으면서 국내외 다양한 검사들에서 MST를 적극적으로 도입하고 있다. 국가수준 학업성취도 평가에서도 MST 도입 계획을 발표하였고(교육부, 2020), PISA(programme for international student assessment)에서도 2018년부터 MST 방식이 활용되고 있다(조성민 외, 2019). 또한 컴퓨터 기반 적응형 평가가 우울이나 불안 등을 측정하는 데에도 효과적임(예: CES-D에 적응형 검사를 활용)이 과학적으로 검증되고 있으며(Gibbons et al., 2016; Smits et al., 2011; Walter et al., 2007), 심리평가에서도 유용성을 가질 수 있음이 제안되고 있다(Weiss, 2011).

상담성과 측정 도구는 주기적이고 반복적으로 활용되는 특성을 고려할 때 측정의 효율성, 즉 적은 문항으로도 핵심적인 변화를 파악할 수 있어야 한다는 점이 특히 중요하다고 볼 수 있다(안성식 외, 2020). 반복적인 긴 검사 과정은 피로도를 증가시키는 요인이 될 수 있으며(서동기 외, 2019; Ackerman & Kanfer, 2009), 따라서 가능한 적은 문항으로 내담자의 심리적 상태를 신뢰롭게 평가할 수 있도록 측정 도구가 설계될 필요가 있다. 이러한 맥락에서 MST는 기존 방식의 측정학적 한계를 보완하여 측정 정확성을 유지하면서도 문항 수를 줄일 수 있는 하나의 대안으로 고려될 수 있다. MST의 활용은 검사 피로도 감소, 맞춤형 문항을 통한 동기 향상, 심각도가 높거나 낮은 대상에 대한 정확성 개선, 변화 탐지 정밀성 향상 등으로 이어질 수 있으며(Gibbons et al., 2008; Sari et al., 2016; Weiss, 2004), 특히 상담성과 측정 도구의 경우 내담자의 응답 부담을 고려하는 것이 중요하다는 점에서 MST의 적용 가능성을 검토할 필요가 있다.

<그림 1.1> MST의 장점

기존 방식	기존 고정형 검사(Fixed Linear Form) (ex. OQ-30, CORE-OM) → 측정 효율성 낮음	단축형 검사(Shortened Fixed Form) (ex. CORE-10) → 측정 정확성 기존보다 낮아짐
↓		
본 연구의 제안	단계적 적응형 검사(MST) 방식 → 더 적은 문항 수로 측정 정확성 유지	

MST를 효과적으로 적용하기 위해서는 해당 검사 특성에 맞춘 측정학적 검토가 필요하며(박인용 외, 2023), 상담성과 측정에서도 해당 검사의 특성에 적합한 검사 설계를 체계적으로 검증할 필요가 있다. 교육 및 심리검사에 실제로 MST를 적용할 때에는 도입에 앞서 모의실험을 실시하여 측정 정확성 및 효율성, 피험자 특성에 따른 측정오차 등의 측정학적 특성을 비교 평가하는 연구가 지속적으로 수행되고 있으며, 이러한 경험적인 연구결과를 바탕으로 실제 검사 설계 및 활용 방안을 결정하고 있다(구남욱 외, 2021; 박인용 외, 2023; 서민희 외, 2020; 한정아 외, 2022). 따라서 상담성과 측정 도구에서도 모의실험을 통해 적합한 검사 설계와 측정학적 성능을 평가하는 것이 필요하며, 현재 활용되고 있는 고정형 검사에 비해 MST가 어떠한 측정학적 장점을 갖는지 경험적으로 검증할 필요가 있다.

이에 본 연구에서는 상담성과 측정에서 MST를 적용하기 위한 검사 설계 및 변화 탐지 방법을 경험적으로 살펴보고, 그 적용 가능성을 검토하고자 한다. 구체적인 연구 문제는 다음과 같다. 첫째, 상담성과 측정 도구(CORE-OM, OQ-30)에 MST를 적용할 때, 소검사는 어떻게 설계될 수 있는가? 둘째, 상담성과 측정에 MST를 적용하였을 때, 측정 정확성이 어떠한가? 셋째, MST를 활용한 상담성과 측정에서 내담자의 변화는 어떻게 탐지될 수 있는가?

2. 이론적 배경

2.1 상담성과 측정 도구

상담성과란 상담을 통해 내담자가 경험하는 모든 종류의 긍정적 변화를 의미한다(김선경, 왕은자, 2016). 상담성과를 측정하기 위한 도구들은 다양하게 개발되어 왔는데, 크게 상담 만족도를 측정하는 도구(예, SEQ(Stiles & Snow, 1984), CSQ(Larsen et al., 1979)), 특정 증상을 측정하는 도구(예: PHQ-9(Kroenke et al., 2001), GAD-7(Spitzer et al., 2006), 내담자의 기능이나 스트레스를 종합적으로 측정하는 도구(예, CORE-OM, OQ-30/45) 등으로 분류해 볼 수 있다(왕은자 외, 2018; Maruish, 2002).

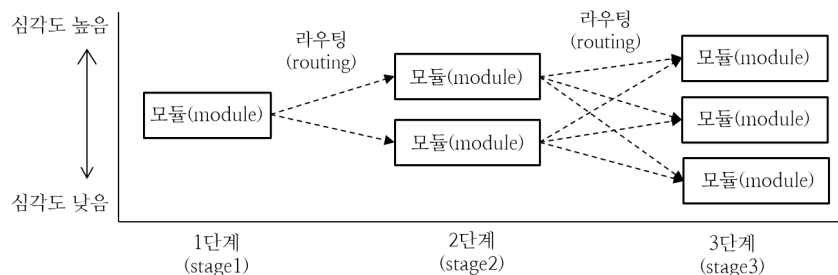
CORE-OM과 OQ-30은 내담자의 증상뿐만 아니라 대인관계와 전반적 기능을 포괄적으로 평가하고자 개발된 대표적인 상담성과 측정 도구이다. CORE-OM은 주관적 안녕감

(subjective well-being), 심리 문제(problems/symptoms), 삶의 기능(life functioning), 위기(risk/harm) 네 영역, 총 34개 문항으로 구성되어 있다. 영국에서 개발되어 유럽을 중심으로 널리 활용되고 있으며, 국내에서도 타당화되어 활용되고 있다(김선경, 왕은자, 2016). OQ-30은 주관적 불편감(subjective distress), 대인관계 기능(interpersonal relationships), 사회적 역할 수행 기능(social role performance) 세 영역, 총 30개 문항으로 구성되어 있다. 미국에서 임상적, 학문적 유용성이 잘 검증되어 널리 활용되고 있으며, 국내에서도 타당화 연구를 통해 활용 가능성이 검증되었다(손난희, 유성경, 2012). CORE-OM과 OQ-30은 점수가 높을수록 심리적 고통 및 기능 저하의 심각도(severity)가 높은 것을 의미하며, 상담을 통한 점수의 감소는 긍정적인 상담성으로 해석될 수 있다.

2.2 단계적 적응형 검사

단계적 적응형 검사(multistage adaptive testing)는 피험자의 심각도/능력 수준에 따라 맞춤형 검사 문항 세트가 단계적으로 제시되는 검사 방법을 말한다. MST의 각 소검사 문항 세트(set of items)는 모듈(module)이라고도 하며, 각 모듈은 특정 심각도/능력 수준에 적합하도록 구성된다. MST는 여러 단계(stage)로 이루어지며, 단계는 피험자에게 문항이 제시되는 순서를 의미한다. 일반적으로 2단계 또는 3단계 설계의 MST가 많이 활용되고 있다(Sari et al., 2016). MST에서는 이전 단계의 응답 결과를 바탕으로 피험자의 심각도/능력을 추정하여 다음 단계의 소검사가 제시된다. 이때 이전 단계 응답을 바탕으로 다음 단계에 제시할 적절한 수준의 모듈이 선택되는 과정을 라우팅(routing)이라고 한다(이하늘, 이용상, 2024; Yan et al., 2014). <그림 2.1>은 1-2-3 설계의 MST의 구성요소를 예시로 제시한 것이다. <그림 2.1>과 같이, 1-2-3 설계에서는 첫 번째 단계에 1개의 공통 모듈, 두 번째 단계에는 2개의 모듈, 세 번째 단계에는 3개의 모듈이 배치되고, 피험자의 이전 단계 응답 결과에 따라 적절한 경로로 라우팅되어 다음 단계에서 적합한 모듈이 제시된다.

<그림 2.1> MST 구성요소



이러한 MST는 컴퓨터 적응형 검사(이하 CAT; computerized adaptive testing)의 특수한 형태로도 볼 수 있다(서민희 외, 2020). CAT에서는 문항 단위로 피험자에게 적합한 문항이 제시된다면, MST는 소검사 단위로 적합한 문항 세트가 여러 단계에 걸쳐 제시

된다. 1단계에서는 모든 피험자에게 동일한 소검사가 공통 모듈로 제시되고, 이후 단계에서는 이전 단계에서의 피험자 심각도/능력 추정 결과를 바탕으로 낮은 수준, 중간 수준, 높은 수준 등 다양한 수준의 여러 소검사 중 가장 적합한 소검사가 선택되어 피험자에게 제공된다(이하늘, 이용상, 2024).

적응형 검사는 개별 피험자의 심각도/능력 수준을 고려하여 문항들이 제시되므로, 모든 피험자에게 동일한 문항이 제시되는 전통적인 고정형 검사에 비해 측정의 정확성 및 효율성이 향상되는 것으로 보고되고 있다(Wang & Chen, 2004; Weiss, 2004). 측정의 정확성은 심각도/능력 추정 시 오차가 적은 정도를 의미하며, 측정의 효율성은 측정을 위해 필요한 문항 수와 관련된다. 고정형 검사는 피험자의 심각도/능력 수준이 평균 수준에 가까울 때에는 비교적 높은 측정 정확성을 보이지만, 양극단의 수준에서는 정확도가 저하되는 한계를 가진다. 이에 비해 적응형 검사는 피험자 수준에 적합한 문항이 제시되기 때문에, 매우 높거나 낮은 수준의 피험자에 대해서도 측정 정확성을 유지할 수 있다는 장점이 알려져 있다(Hendrickson, 2007; Kingsbury & Zara, 1989)

MST와 CAT는 적응형 검사라는 점에서는 유사한 점이 있지만, CAT는 MST에 비해 문항 단위로 피험자 수준에 맞추어 제시되어야 하므로 문항은행의 규모가 특히 더 커야 한다는 부담이 있다. 또한 MST는 시스템 구현 및 운영의 부담을 줄일 수 있고, 검사 설계 및 운영의 측면에서도 특히 장점을 가지는데 MST에서는 모듈 내 문항 구성 등 세부 사항을 검사 개발자가 더 많이 제어할 수 있으며 내용 전문가의 의견을 반영하기 더 용이하다는 측면도 있다(Mead, 2006). 이러한 MST는 미국 대학원 입학자격시험(graduate record examination, GRE), 국제 학업성취도 평가(programme for international student assessment, PISA) 등 최근 다양한 평가 장면에서 적극적으로 적용되고 있다(한정아 외, 2022).

2.3 관련 선행연구

교육 및 심리검사에 적응형 검사를 도입하기 위한 맥락의 관련 선행연구는 크게 두 가지 흐름으로 구분해 볼 수 있다. 첫째, 국내에서는 특히 국가수준 학업성취도 평가에 MST를 도입하기 위해 적용 가능성을 검증하고 측정학적 특성을 확인하는 모의실험 연구가 지속적으로 수행되어 왔다(구남욱 외, 2021; 박인용 외, 2023; 서민희 외, 2020; 한정아 외, 2022). 둘째, 심리검사 분야에서의 국내 연구는 아직 제한적인 반면, 해외에서는 심리검사에 대해서도 적응형 검사의 적용 가능성을 검증하기 위한 연구들이 이루어져 왔다. 예를 들어, CES-D(center for epidemiologic studies depression scale)나 BDI(Beck depression inventory) 등 심리적 증상을 측정하는 검사에 CAT를 적용했을 때 어떤 측정학적 장점을 가지는지 모의실험을 통해 검증한 연구들이 제시되고 있다(Capon et al., 2026; Gardner et al., 2004; Smits et al., 2011).

국가수준 학업성취도 평가에 MST를 적용하기 위한 연구들은 한국교육과정평가원을 중심으로 수행되어 왔으며, 고정형 검사와 적응형 검사의 피험자 능력 추정 정확성 비교, MST 설계 방식(예, 단계별 소검사 수, 라우팅 방식 등)에 따른 측정 성능을 체계적으로 검증해 왔다. 대표적인 연구 결과를 살펴보면, 서민희 외(2020)는 MST의 적용 가능성을

탐색하고자 모의실험을 통해 고정형 검사와 적응형 검사를 비교한 결과, 고정형 검사에 비해 MST에서 피험자의 능력이 보다 정확하게 추정됨을 확인하였다. 또한 고정형 검사는 특정 능력 구간에서는 비교적 큰 오차를 보이기도 하였으나, MST는 전체 능력 범위에서 안정적인 성능을 보였다. 또한 최적의 MST 설계 방안을 탐색하기 위한 모의 실험 연구들도 다각적으로 수행되어 왔다. 한정아 외(2020)는 MST 설계에 따른 성능 비교 결과, 2단계 MST와 3단계 MST에서의 능력 추정 정확성은 비슷하게 나타났으나, 1-2-3 설계에서 최적경로 비율이 가장 높게 나타나 피험자 능력에 맞는 소검사가 가장 잘 배정되었음을 제시하였다. 구남옥 외(2021)는 소검사 배정을 위한 라우팅 방식을 비교하였고, 연구 결과 MFI(maximum fisher information)의 성능이 가장 우수한 것으로 나타났다. 박인용 외(2023)는 피험자 모수 추정 방법에 대한 정확도를 비교한 결과, MLE(maximum likelihood estimation)와 EAP(Expected A Posteriori)가 모두 높은 정확도를 보였고, 극단적인 수준에서의 성취수준 분류는 MLE가, 평균적인 수준에서는 EAP가 다소 더 정확한 경향이 있음을 보고하였다. 학업성취도 평가 맥락에서 1수준(기초학력 미달) 분류 정확도가 중요함을 언급하면서 최종 능력모수 추정 방식으로 MLE를 제안하기도 하였다. 본 연구에서는 이러한 선행연구 결과를 바탕으로 연구를 설계하고 모의실험 조건을 설정하고자 하였다.

한편, 해외 연구에서는 학업성취도 평가뿐만 아니라 심리검사에서도 적응형 검사를 적용하고자 하는 연구들이 수행되어 왔다. 기존에 고정형 검사로 활용되어 온 우울이나 불안 등 심리적 증상 측정 도구에 적응형 검사를 적용했을 때 측정학적 장점을 가지는지 모의실험을 통해 검증해왔다. Smits et al.(2011)은 우울을 측정하는 CES-D(원척도 20 문항)에 적응형 검사를 적용하였고, 연구 결과 적응형 검사에서 평균적으로 문항 수를 약 1/3만 사용했음에도 전체 검사 점수와 0.96 정도의 상관을 보여, 적응형 검사가 검사 효율성을 개선하기 위한 유용한 방법이 될 수 있음을 제시하였다. 또한 Gardner et al.(2004)은 우울을 측정하는 BDI(원척도 21문항)에 적응형 검사를 적용한 결과, 평균적으로 5.6 문항만으로도 BDI 총점과 0.92 정도의 높은 상관을 보여 적응형 검사가 검사 효율성을 크게 향상시키면서도 측정 정확성을 유지할 수 있음을 보였다. Capon et al.(2026)의 연구에서도 정신건강 진료를 받는 과정에서 디지털 건강 평가 도구에 적응형 검사를 적용하였을 때, 평균 문항 수가 69% 정도 감소하고(개인당 49문항→15.3문항), 평균 검사 시간이 10.5분에서 3.3분으로 줄어들었고, 전체 검사와의 일치도가 자살위험성 0.92, 불안 0.92, 알코올 사용 0.91, 심리적 고통 0.88, 기능 수준 0.86 등으로 상당히 높게 나타났음을 제시하였다. 이러한 선행연구들은 심리적 특성을 측정하는 측정 도구에서도 적응형 검사가 측정학적 장점을 가질 수 있음을 보여준다. 또한 우울과 같은 단일 증상뿐만 아니라 여러 표준화된 척도 점수를 활용하여 다양한 정신건강 영역을 동시에 평가하는 상황에서도 적응형 검사가 유용할 수 있음을 시사한다. 그리고 이러한 결과는 상담성과 측정 도구인 CORE-OM과 OQ-30에서도 적응형 검사를 적용한다면, 고정형 검사보다 더 적은 문항으로도 측정 정확도를 유지할 수 있을 가능성을 보여준다. 다만, 이러한 선행연구들은 적응형 검사 중 CAT를 주로 적용하였는데, 본 연구에서는 MST가 가지는 검사 제어 및 운영의 측면에서의 장점을 고려하여 상담성과 측정 도구에 대한 MST의 적용 가능성을 검증하고자 하였다.

3. 연구방법

본 연구는 MST 검사 설계, 성능 평가 모의실험, 변화 탐지 적용의 세 단계로 수행되었으며, 분석에는 R의 MASS, mirt, mstR 패키지가 활용되었다.

<그림 3.1> 연구 절차



3.1 MST 검사 설계

상당성과 측정 도구의 문항 모수 추정을 위해, CORE-OM과 OQ-30의 국내 타당화 연구(김선경, 2012; 김선경, 왕은자, 2016; 손난희, 유성경, 2012)에 보고된 문항별 평균과 표준편차, 요인부하량 및 요인 간 상관을 기반으로 응답 분포를 재현한 데이터를 생성하였다. 그리고 다분 문항에 대한 문항반응이론(이하 IRT; item response theory) 모형인 등급반응모형(graded response model)을 적용하여 문항 모수를 추정하였다. 그런 다음 자동 검사 구성(이하 ATA; automated test assembly)을 적용하여 소검사별 목표 심각도(target Θ)에서 검사정보함수(이하 TIF; test information function) 정보량이 최대가 되도록 소검사 문항 조합을 설계하였다.

적용된 연구방법과 관련하여 보다 구체적으로 기술하면, Aybek(2023)은 고전검사이론(이하 CTT; classical test theory) 추정치를 활용해 IRT 모수를 간접적으로 산출한 결과 간접 추정된 값과 실제 문항 모수 간의 상관이 매우 높았고 적응형 검사에서의 심각도/능력 모수 추정도 안정적임을 제시하였다. 또한 McDonald(1999)는 요인분석과 문항반응이론을 통합된 수학적 틀로 제시하여 모수 변환의 이론적 근거를 제공하였고, Stochl et al.(2016)은 McDonald(1999)의 접근을 적응형 검사를 위한 시뮬레이션 연구에 적용하여 활용한 바 있다. 본 연구에서도 현실적으로 원자료 확보가 어려운 상황에서 적응형 검사를 위한 문항 수준의 정보를 추정하기 위한 접근으로 근사 자료에 기반한 간접 추정 방식이 활용되었다. 문항 모수 추정 시 참고한 CORE-OM과 OQ-30의 국내 타당화 연구의 데이터 특징은 <표 3.1>과 같다.

<표 3.1> CORE-OM, OQ-30의 국내 타당화 연구의 데이터 특징

척도	출처	대상	표본 크기	표본 구성		신뢰도 (a)
				성별	연령	
CORE-OM	김선경, 왕은자 (2016)	성인 내담자	n=600	남 36.4%, 여 63.6%	20대 35.5%, 30대 35%, 40대 17.3%, 50대 이상 10.8%, 무응답 1.4%	0.92
OQ-30	손난희, 유성경 (2012)	대학생 및 성인 내담자	n=350	남 25.7%, 여 74.3%	평균연령 26.69세 (SD = 8.09) (18~55세)	0.94

덧붙여, 다분 문항을 위한 대표적인 IRT 모형으로는 등급반응모형(Samejima, 1969)과 부분점수모형(Muraki, 1992)이 있는데, Smits et al.(2011)은 이 두 모형이 거의 동일한 결과를 제공하지만 등급반응모형은 이해하기 쉽고 해석 가능한 모수를 제공(Mellenbergh, 1995)한다는 장점이 있음을 강조하였다. 이에 본 연구에서도 등급반응모형을 활용하였다. 또한 IRT 분석 전에 일차원성 가정과 지역독립성 가정을 만족하는지 확인하였을 때, 주성분분석에서 첫 번째 요인의 전체 문항분산에 대한 설명량이 Reckase(1979)의 기준 (20%)을 충족하였고, 문항 간 잔차 상관의 평균이 0.1보다 작아, IRT 분석을 위한 가정을 충족하는 것으로 판단하였다.

ATA 알고리즘은 Lord(1971)와 Samejima(1977)가 제시한 TIF 중심의 검사 조립 원리를 토대로 하였으며, 본 연구에서는 소규모 문항은행의 특성을 고려하여 문항은행 내 가능한 문항 조합을 생성하고 TIF의 최대 정보 위치와 그 지점에서의 정보량(test information at θ)을 기준으로 필터링하는 방식을 적용하였다. 아래 식에서 $TIF(\theta^*)$ 는 목표 심각도 수준에서의 검사정보함수, $I_i(\theta^*)$ 는 목표 심각도 수준에서 문항 i 의 정보량, $S_i^{(k)}$ 는 k 단계 I 번째 모듈에 선택된 문항 집합, n_k 는 k 단계의 각 모듈 문항 수 제약을 나타낸다.

$$\begin{aligned} & \text{Maximize } TIF(\theta^*) = \sum_{i \in S} I_i(\theta^*) \quad \text{단, 공통 모듈에서는 } \text{Maximize } \int_{\theta_L}^{\theta_U} TIF(\theta) d\theta \quad (3.1) \\ & \text{subject to } \sum_{i=1}^I S_i^{(k)} = n_k, \quad |\theta_{target} - \theta^*| \leq tolerance, \quad S_i^{(k)} \cap S_j^{(k)} = \emptyset \end{aligned}$$

ATA 적용 과정으로는, 먼저 전체 TIF를 기반으로 평균 심각도를 산출하고 이를 중심으로 ± 1 표준편차에 해당하는 지점을 고려하여 목표 심각도 수준(θ^*)을 설정하였다. 다음으로 정보량이 가장 큰 최적의 문항 세트를 탐색하였고, 공통 모듈은 특정 심각도 수준이 아닌 검사정보함수 전체 구간에서의 총 정보량(AUC)이 최대화하도록 문항 세트가 구성되었다. 중복 제시를 방지하기 위해 한 단계에서 선택된 문항은 이후 단계의 문항 풀에서 제외되었다.

3.2 성능 평가 모의실험

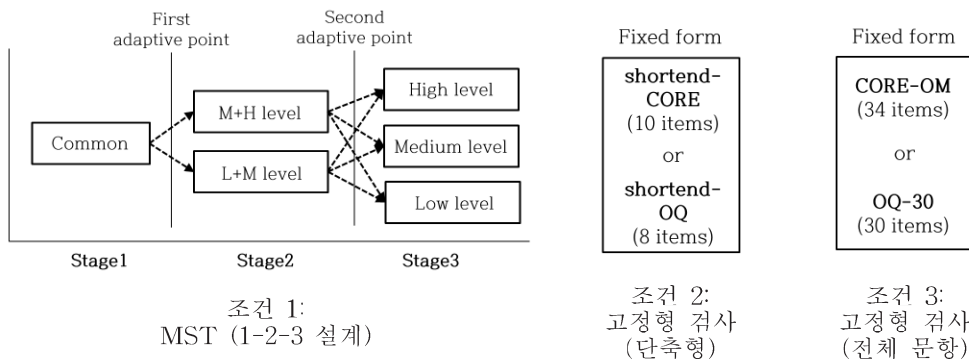
모의실험(Monte Carlo simulation)을 통해 MST 기반 검사와 고정형 검사의 측정 정확성(accuracy)을 비교 평가하였다. MST의 모의실험 조건은 관련 선행연구(구남욱 외, 2021; 박인용 외, 2023; 서민희 외, 2020; 한정아 외, 2022 등)를 참고하여 설계하였다. MST 구성은 정확성과 최적 경로 비율이 우수한 것으로 보고된 1-2-3 설계를 적용하였다. 1-2-3 설계는 PISA 및 국가수준 학업성취도 평가에도 적용되고 있다. 또한 라우팅 방식으로는 대표적으로 많이 활용되는 MFI를 적용하였고, MFI는 다음 단계의 모듈을 선택할 때 감정적인 심각도/능력 수준에서 정보량이 가장 높은 모듈을 선택하는 방식이다(Sari et al., 2016). 피험자 모수 추정은 MLE를 적용하였으며, MLE는 각 문항 점수를 단순히 합하는 방식과는 달리, 피험자의 응답 패턴과 문항 특성에 대한 정보를 결합하여 관찰된 응답 패턴을 생성할 가능성이 가장 높은 θ 값을 찾아 피험자 모수를 추정한다(Embretson & Reise, 2000).

MST와의 비교를 위한 고정형 검사의 조건으로는 전체 문항을 활용한 검사(CORE-OM 34문항, OQ-30 30문항)와 단축형 검사(CORE-10의 10문항, OQ-10 문항 중 OQ-30과 겹치는 8문항)를 포함하였다. 즉 모의실험에서는 MST(1-2-3 design), 단축형 고정형 검사, 전체 문항에 대한 고정형 검사의 세 가지 조건을 비교하였다. 그리고 문항 수가 동일한 조건에서 MST와 고정형 검사의 측정 정확성을 비교하기 위해, MST의 문항 수는 단축형 검사의 문항 수와 동일하게 설계되었다.

<표 3.2> 검사 설계 방식에 따른 문항 수 조건

검사 설계 방식	문항 수 조건	
	CORE-OM	OQ-30
조건 1: 단계적 적응형 검사 (1-2-3 설계)	10 (3-3-4)	8 (2-3-3)
조건 2: 고정형 검사 (단축형)	10	8
조건 3: 고정형 검사 (전체 문항)	34	30

<그림 3.1> 검사 설계 방식에 따른 모의실험 조건



모의실험을 위한 자료 생성을 위해 사용한 피험자 모수(θ)는 표준정규분포를 활용하여 5,000명의 θ 를 생성하였고, 이를 참값으로 활용하였다. 문항 모수는 앞서 추정된 CORE-OM, OQ-30의 각 문항 모수 값을 그대로 반영하였다. 통계적 분포로부터 n 명의 피험자 모수를 생성하여 참값으로 활용하는 방법은 MST 모의실험에서 일반적으로 많이 활용되고 있으며(구남욱 외, 2021; 박인용 외, 2023; 서민희 외, 2020; 한정아 외, 2022 등), 본 연구에서도 이러한 방법을 적용하였다.

성능 평가를 위한 지표로는 θ 에 대한 참값과 추정치 간 상관, 전체 편향(overall bias), MAE(mean absolute error), RMSE(root mean squared error)를 활용하였다. 또한 θ 구간별 조건부 편향(conditional bias) 통계치도 비교하여 심각도 수준별 측정의 정확성도 함께 평가하였다. 아래 식에서 θ_i 는 피험자 모수의 참 값을 $\hat{\theta}_i$ 은 피험자 모수의 추정치를 나타낸다. 전체 편향은 추정치가 참값보다 전반적으로 높거나 낮게 추정되는 정도를 나타내고, MAE는 평균적으로 참값에서 얼마나 벗어나 있는지를 보여준다. RMSE는 추정치와 참값 간 오차의 크기를 반영하며 큰 오차에 더 큰 가중치를 부여하여 추정 성능을 평가하는 지표이다.

$$overall\ Bias = \frac{1}{N} \sum (\hat{\theta}_i - \theta_i) \quad (3.2)$$

$$MAE = \frac{1}{N} \sum |\hat{\theta}_i - \theta_i| \quad (3.3)$$

$$RMSE = \sqrt{\frac{1}{N} \sum (\hat{\theta}_i - \theta_i)^2} \quad (3.4)$$

참값과 추정치 간 상관은 1에 가까울수록, 편향의 절댓값, MAE, RMSE는 0에 가까울수록 측정 정확성이 높다고 해석할 수 있다.

3.3 변화 탐지

상당성과 평가를 위한 방법으로 Jacobson과 Truax(1991)는 신뢰할 수 있는 변화(이하 RC; reliable change)의 개념을 제안하였다. RC란 내담자의 변화가 우연한 오차에 의한 것이 아니라 실제적인 변화로 해석할 수 있는지를 의미한다(왕은자 외, 2014; Jacobson & Truax, 1991). RC 지수 및 95% 신뢰구간은 다음과 같이 구해질 수 있으며(Christensen & Mendoza, 1986), RC 지수의 절댓값이 1.96을 넘거나, 95% 신뢰구간에 0을 포함하지 않으면 신뢰할 수 있는 변화에 해당한다고 해석할 수 있다.

$$RC\ index = \frac{change}{SE_{change}} \quad (3.5)$$

$$95\%\ C.I. = change \pm 1.96 \times SE_{change} \quad (3.6)$$

SE_{change} 는 변화에 대한 측정오차를 나타내며, CTT 기반 방법과 IRT 기반 방법에서는 SE_{change} 를 다른 방식으로 추정한다. 현재 실무에서는 CTT 기반 변화 탐지 방법이 주로 활용되고 있으며, OQ-30에서는 약 10점 이상의 변화를 신뢰로운 변화(RC)로 해석할 수 있다고 제안하였고(Lambert et al., 2005), CORE-OM에서는 자료의 신뢰도와 표준편차를 이용하여 RC 기준을 산출할 것을 권장하고 있다(Evans, 2025).

$$CTT: SE_{change} = SD_{pre} \sqrt{2} \sqrt{1 - reliability} \quad (3.7)$$

$$IRT: SE_{change, i} = \sqrt{SE_{pre, i}^2 + SE_{post, i}^2} \quad (3.8)$$

하지만 이러한 CTT 기반의 방법은 모든 내담자의 측정오차가 동일하다는 다소 강한 가정을 바탕으로 하고 있다. 최근에는 이를 보완하기 위해 IRT 기반의 변화 탐지와 같은 보다 정교한 통계적 방법이 제안되고 있다(Blampied, 2022). CTT 기반의 방법에서는 검사 전체의 신뢰도와 표준편차를 활용하여 내담자의 측정오차를 모두 동일한 값으로 산출하는 반면, IRT 기반의 변화 탐지 방법에서는 각 내담자별 문항 반응을 바탕으로 개인별 측정오차를 추정할 수 있고 이를 변화 탐지에 반영한다(Hays et al., 2021; Maydeu-Olivares, 2021). IRT 기반의 변화 탐지는 CTT 기반의 변화 탐지에 비해 더 우수한 정확도를 보이는 것으로 보고되고 있다(Keller & Alexandrowicz, 2024; Tang et al., 2023). 본 연구에서는 IRT 기반 RC 탐지 방법을 적용하여, MST를 활용한 상담성과 측정에서 내담자 변화 탐지 방법이 어떻게 활용될 수 있는지를 제시하였다.

4. 연구결과

4.1 단계적 적응형 검사의 소검사 설계

CORE-OM과 OQ-30에 대해 MST 소검사를 설계한 결과는 다음과 같다. <표 4.1>에 제시된 각 소검사의 최대 정보 위치(peak θ)를 통해 해당 소검사에서 가장 잘 측정하는 심각도 수준을 확인할 수 있으며, 각 단계 소검사는 목표 심각도 수준을 전반적으로 잘 반영하는 것으로 나타났다. 3단계 Low, Medium, High 소검사에 주목해보면, CORE-OM MST에서는 각각 $\theta = -0.6, 0.3, 1.1$ (정규분포 기준 하위 27%, 상위 38%, 상위 14% 지점)에서 정보량이 가장 높았고, OQ-30 MST에서도 각각 $\theta = -1.0, -0.2, 0.7$ (하위 16%, 하위 42%, 상위 24%)에서 정보량이 가장 높았다. 이는 각 소검사가 서로 다른 심각도 수준을 구분하여 측정할 수 있게 설계되었음을 보여준다. 덧붙여, CORE-OM MST는 OQ-30 MST에 비해 상대적으로 더 높은 심각도 수준에서 정보량이 높은 경향을 보였는데, CORE-OM 원칙도에 위기 영역 문항들도 포함되어 있어 보이는 특성으로 해석할 수 있다.

<표 4.1> MST 소검사별 특성: CORE-OM, OQ-30

		CORE-OM			OQ-30		
		문항 수	Peak θ	Σ TIF	문항 수	Peak θ	Σ TIF
1단계	공통 모듈	3	0.10	5.2	2	-0.10	6.7
2단계	L+M 수준	3	-0.05	4.7	3	-0.45	9.7
	M+H 수준	3	0.65	9.0	3	0.25	10.8
3단계	Low 수준	4	-0.60	11.3	3	-1.00	6.8
	Medium 수준	4	0.30	13.4	3	-0.20	12.6
	High 수준	4	1.10	14.7	3	0.70	11.1

<표 4.2> ATA를 활용한 MST 소검사 설계 결과: CORE-OM

		소검사 문항 구성
1단계	공통 모듈	· CORE-OM 문항 10번 (다른 사람에게 말하는 것이 어렵게 느껴진 적이 있다) · CORE-OM 문항 12번 (그동안 해 온 일에 만족하고 있다(역)) · CORE-OM 문항 31번 (나의 미래는 낙관적이라고 생각한다(역))
2단계	L+M 수준	· CORE-OM 문항 4번 (나는 괜찮은 사람이라고 생각한다(역)) · CORE-OM 문항 17번 (내 문제로 인해 압도당하는 느낌이다) · CORE-OM 문항 30번 (나의 문제와 어려움에 대한 책임은 나에게 있다고 생각한 적이 있다)
	M+H 수준	· CORE-OM 문항 15번 (심한 공포나 두려움을 느낀 적이 있다) · CORE-OM 문항 17번 (내 문제로 인해 압도당하는 느낌이다) · CORE-OM 문항 25번 (다른 사람들이 나를 비난한다고 느낀 적이 있다)
3단계	Low 수준	· CORE-OM 문항 2번 (긴장하거나 불안감을 느낀 적이 있다) · CORE-OM 문항 13번 (원치 않는 생각이나 느낌 때문에 힘든 적이 있다) · CORE-OM 문항 29번 (다른 사람들과 있을 때 짜증이 난 적이 있다) · CORE-OM 문항 32번 (나는 내가 원하는 것을 이루었다(역))
	Medium 수준	· CORE-OM 문항 1번 (완전히 혼자이고, 소외되었다고 느낀 적이 있다) · CORE-OM 문항 11번 (긴장이나 걱정 때문에 중요한 일을 못한 적이 있다) · CORE-OM 문항 14번 (울고 싶은 적이 있었다) · CORE-OM 문항 28번 (원치 않는 이미지나 기억이 떠올라서 괴로웠던 적이 있다)
	High 수준	· CORE-OM 문항 16번 (삶을 마감할 계획을 세워 놓았다) · CORE-OM 문항 23번 (절망적이고 희망이 없는 것처럼 느껴진 적이 있다) · CORE-OM 문항 26번 (친구가 없다고 생각한 적이 있다) · CORE-OM 문항 27번 (불행하다고 느낀 적이 있다)

<표 4.3> ATA를 활용한 MST 소검사 설계 결과: OQ-30

		소검사 문항 구성
1단계	공통 모듈	· OQ-30 문항 15번 (집중이 잘 되지 않는다) · OQ-30 문항 25번 (긴장된다)
2단계	L+M 수준	· OQ-30 문항 6번 (허약한 것 같다) · OQ-30 문항 26번 (직장이나 학교생활을 잘 못해내고 있는 것 같다) · OQ-30 문항 27번 (내 마음에 뭔가 문제가 있다고 느낀다)
	M+H 수준	· OQ-30 문항 4번 (짜증난다) · OQ-30 문항 5번 (인생을 끝내고 싶다는 생각을 한다) · OQ-30 문항 10번 (나는 행복한 사람이다(역))
3단계	Low 수준	· OQ-30 문항 2번 (직장이나 학교 혹은 다른 일상 활동에서 스트레스를 받는다) · OQ-30 문항 13번 (외롭다) · OQ-30 문항 22번 (내 인생에 만족한다(역))
	Medium 수준	· OQ-30 문항 7번 (두려운 느낌이 든다) · OQ-30 문항 17번 (괴로운 생각들이 머릿속에서 떠나지 않는다) · OQ-30 문항 29번 (슬프다)
	High 수준	· OQ-30 문항 1번 (매사에 관심이 없다) · OQ-30 문항 11번 (내가 쓸모없다고 느낀다) · OQ-30 문항 24번 (뭔가 나쁜 일이 생길 것 같다)

4.2 검사 설계 방식에 따른 측정 정확성

검사 설계 방식에 따른 측정 정확성을 비교한 결과, MST에서 참값과 추정치 간 상관이 0.92로 높게 나타나 우수한 측정 정확성을 보였다. 전체 문항을 활용하는 고정형 검사에 비해서는 소폭 낮은 수준이었으나, 이는 문항 수의 차이(전체 문항: 34문항/30 문항, MST: 10문항/8문항)로 인한 현상으로 볼 수 있으며, 1/3~1/4 수준의 적은 문항 수로도 전체 문항에 근접한 수준의 측정 정확성을 보였다는 점에서 MST의 효율성을 보여주는 결과로도 볼 수 있다. 또한 동일한 문항 수 조건에서 비교했을 때에는, MST는 고정형 검사에 비해 더 높은 측정 정확도를 보였다. 참값과 추정치 간 상관을 살펴보면, 단축형 고정형 검사에서 0.88~0.89인데 비해 MST에서는 0.92로 더 높은 값을 보였고, MAE도 MST가 단축형 고정형 검사보다 조금 더 낮게 나타났다.

전체 편향은 MST와 고정형 검사 모두 0에 가깝게 나타났으나, θ 구간별로 살펴 보았을 때, 단축형 고정형 검사는 심각도가 매우 낮은 구간($\theta \leq -2$)에서는 체계적 과대 추정, 심각도가 매우 높은 구간($\theta > 2$)에서는 체계적 과소추정 경향을 보였다. 또한 심각도가 매우 높거나 낮은 조건에서는, 더 많은 문항을 사용하는 전체 문항 고정형 검사보다도 더 적은 수의 맞춤형 문항을 사용하는 MST에서 편향이 더 작은 것으로 나타났다. 이는 MST는 각 응답자의 심각도 수준에 적합한 문항을 제시함으로써, θ 의 전 범위에서 안정적인 정확성을 유지할 수 있음을 보여주며, 특히 심각도가 높거나 낮은 응답자의 경우 MST를 활용하면 추정의 정확성이 더 높아질 수 있음을 시사한다.

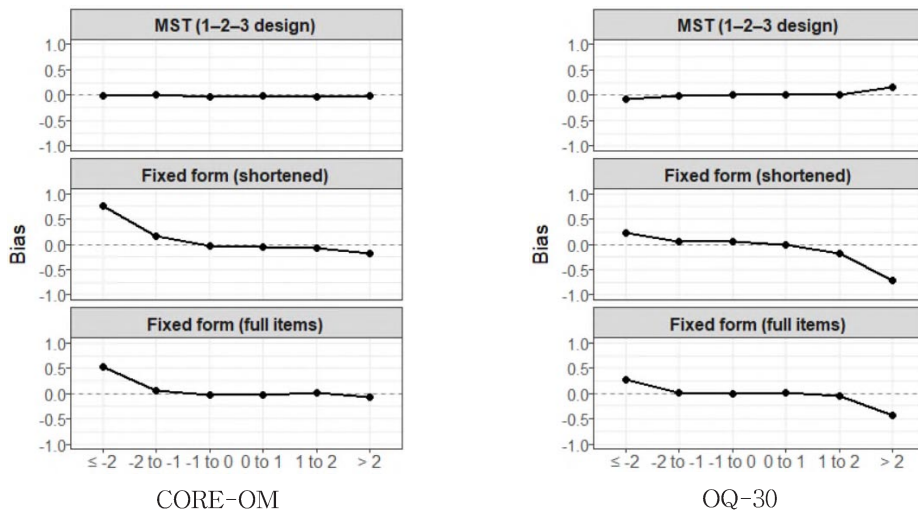
<표 4.4> 검사 설계에 따른 측정 정확성 비교

		Items	$\text{Corr}(\theta, \hat{\theta})$	Bias	MAE	RMSE
CORE-OM	단계적 적응형 검사 (1-2-3 설계)	10	0.92	-0.02	0.34	0.43
	고정형 검사 (단축형)	10	0.89	-0.01	0.38	0.47
	고정형 검사 (전체 문항)	34	0.95	-0.01	0.25	0.32
OQ-30	단계적 적응형 검사 (1-2-3 설계)	8	0.92	0.00	0.34	0.42
	고정형 검사 (단축형)	8	0.88	-0.01	0.35	0.48
	고정형 검사 (전체 문항)	30	0.95	-0.01	0.24	0.30

<표 4.5> 검사 설계에 따른 θ 구간별 편향 비교

		θ 구간별 편향					
		$\theta \leq -2$	$-2 < \theta \leq -1$	$-1 < \theta \leq 0$	$0 < \theta \leq 1$	$1 < \theta \leq 2$	$2 < \theta$
CORE-OM	단계적 적응형 검사 (1-2-3 설계)	-0.02	0.00	-0.03	-0.01	-0.04	-0.01
	고정형 검사 (단축형)	0.75	0.17	-0.03	-0.06	-0.07	-0.18
	고정형 검사 (전체 문항)	0.53	0.07	-0.04	-0.03	0.00	-0.07
OQ-30	단계적 적응형 검사 (1-2-3 설계)	-0.07	-0.01	0.02	0.01	0.01	0.15
	고정형 검사 (단축형)	0.23	0.05	0.05	0.00	-0.19	-0.71
	고정형 검사 (전체 문항)	0.26	0.00	-0.02	0.01	-0.05	-0.43

<그림 4.1> 검사 설계 방식에 따른 θ 구간별 편향 비교

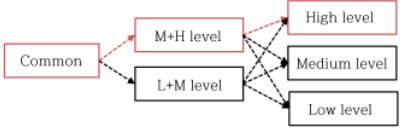
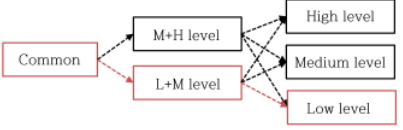
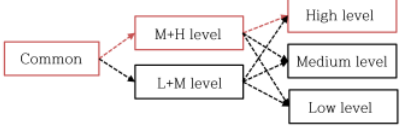
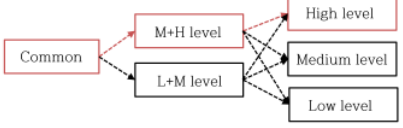
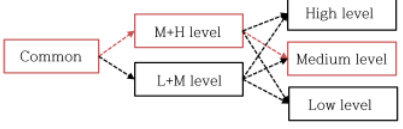
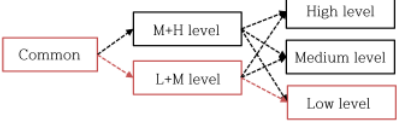


4.3 MST를 적용한 상담성과 측정에서의 변화 탐지

<표 4.6>에는 가상 내담자에 대한 변화 탐지 예시를 제시하였다. CORE-OM MST를 적용했을 때, 사례 1에서는 치료 전 $\Theta_{Pre}=2.10(SE_{Pre}=0.42)$, 치료 후 $\Theta_{post}=-1.82(SE_{post}=0.52)$ 로, 치료 전보다 치료 후에 심리적 고통 및 기능 저하의 심각도(Θ)가 감소한 것으로 나타났다. 이러한 변화가 RC에 해당하는지 확인한 결과, 관찰된 변화에 대한 95% 신뢰구간이 $[-5.23, -2.61]$ 로 0이 포함되지 않았다. 이는 측정 시마다 우연한 오차에 의해 측정값이 다소 달라질 수 있음을 고려하더라도, 실제 변화량이 0이 아닐 것으로 추정됨을 의미한다. 따라서 사례 1의 내담자의 관찰된 변화는 측정오차에 의한 변동이 아니라 실제 발생한 신뢰로운 변화로 해석할 수 있다. 사례 2에서는 치료 전 $\Theta_{Pre}=1.11(SE_{Pre}=0.41)$, 치료 후 $\Theta_{post}=0.96(SE_{post}=0.41)$ 으로 심각도가 소폭 감소하였다. 이러한 변화가 RC에 해당하는지 확인해보면, 관찰된 변화의 95% 신뢰구간 $[-1.28, 0.97]$ 에 0이 포함되어 있어, 측정오차를 고려할 때 신뢰로운 변화로 보기는 어렵다고 판단할 수 있다. 사례 3에서는 치료 전($\Theta_{Pre}=-1.62(SE_{Pre}=0.46)$)과 치료 후($\Theta_{post}=-4.00(SE_{post}=0.88)$)의 변화가 RC에 해당하는지 확인했을 때 95% 신뢰구간이 $[-4.34, -0.41]$ 로 나타나 신뢰로운 변화에 해당한다고 해석할 수 있다.

이러한 결과를 기존의 고정형 검사에서 활용되어 온 변화 탐지 방법의 결과와 비교해보았다. 고정형 검사에서는 신뢰도와 표준편차를 활용하여 RC에 해당하는 변화 기준 점수(cutoff)를 산출할 수 있다. 실제로 CORE-OM에서는 수집한 자료의 신뢰도와 표준편차를 입력하여 RC 기준점수를 산출할 수 있는 온라인 계산 도구(PSYCTC RCI Calculator)를 안내하고 있다(Evans, 2025). 본 연구에서는 CORE-OM 국내 타당화 연구에서 보고된 신뢰도($\alpha=0.92$)와 표준편차($SD=0.61$)를 적용했을 때, RC 기준점수는 0.48점으로 산출되었다. 이 기준을 적용한 결과, 사례 1은 신뢰로운 변화에 해당하였고, 사례 2와 사례 3은 신뢰로운 변화에 해당하지 않는 것으로 나타났다. 사례 1과 사례 2에서는 MST의 IRT 기반 변화 탐지와 고정형 검사의 CTT 기반 변화 탐지 결과가 동일하게 나타났다. 사례 3에서는 두 방법의 변화 탐지 결과가 차이를 보였는데, 고정형 검사의 CTT 기반 RC 평가에서 사례 3의 변화 점수는 0.47점으로 RC 기준점수인 0.48점에 매우 근접하였으나 충족하지는 못하였고, MST의 IRT 기반의 RC 평가에서는 신뢰로운 변화에 해당하는 것으로 탐지되었다. 또한 두 접근의 변화 탐지 방법을 비교하여 살펴보면, 고정형 검사의 CTT 기반 변화 탐지 방법에서는 모든 내담자에게 동일한 측정오차($SE_{Change} = 0.61 \times \sqrt{2} \sqrt{(1-0.92)}$)를 적용하는 반면, MST의 IRT 기반 변화 탐지 방법에서는 각 내담자별 측정오차(사례 1: $SE_{Change}=0.67$, 사례 2: $SE_{Change}=0.58$, 사례 3: $SE_{Change}=0.99$)를 반영하여 평가한다는 점에서 차이가 있다.

<표 4.6> MST를 활용한 상담성과 측정에서 변화 탐지 예시

		고정형 검사 (기존 방법)	MST (제안하는 방법)
사례 1	상담 전 측정	Score= 2.29점	$\Theta_{Pre}=2.10$ ($SE_{Pre}=0.42$) 
	상담 후 측정	Score= 0.94점	$\Theta_{post}=-1.82$ ($SE_{post}=0.52$) 
	상담 전-후 변화	Δ Score= -1.35점	$\Delta\Theta=-3.92$ 95% C.I.=[-5.23, -2.61]
	RC 평가	Yes	Yes (신뢰할 수 있는 변화에 해당)
	문항 수	34	10
사례 2	상담 전 측정	Score= 2.62점	$\Theta_{Pre}=1.11$ ($SE_{Pre}=0.41$) 
	상담 후 측정	Score= 2.21점	$\Theta_{post}=0.96$ ($SE_{post}=0.41$) 
	상담 전-후 변화	Δ Score= -0.41점	$\Delta\Theta=-0.15$ 95% C.I.=[-1.28, 0.97]
	RC 평가	No	No (신뢰할 수 있는 변화에 해당하지 않음)
	문항 수	34	10
사례 3	상담 전 측정	Score= 0.65점	$\Theta_{Pre}=-1.62$ ($SE_{Pre}=0.46$) 
	상담 후 측정	Score= 0.18점	$\Theta_{post}=-4.00$ ($SE_{post}=0.88$) 
	상담 전-후 변화	Δ Score= -0.47점	$\Delta\Theta=-2.38$ 95% C.I.=[-4.34, -0.41]
	RC 평가	No	Yes (신뢰할 수 있는 변화에 해당)
	문항 수	34	10

5. 결론 및 논의

이 연구는 상담성과 측정에서 단계적 적응형 검사의 적용 가능성을 탐색하고, 시뮬레이션을 통해 그 성능을 평가하였다. 이를 위해 대표적인 상담성과 측정 도구인 CORE-OM과 OQ-30에 MST를 적용하였고, 모의실험을 통해 검사 설계 방식에 따른 측정 정확성을 비교하였다. 또한 MST를 적용한 상담성과 측정에서 내담자의 변화 탐지의 예시를 제시하였다.

연구 결과, 첫째 CORE-OM과 OQ-30에 MST를 적용한 결과, ATA를 활용하여 소검사를 설계했을 때 각 소검사는 목표한 심각도 수준을 전반적으로 잘 반영하는 것으로 나타났다. Zheng et al.(2013)은 바람직한 MST 설계의 특징으로 각 소검사가 서로 다른 능력 수준에서 높은 정보량을 제공하여 정보함수가 구분되어야 하며, 소검사 경로가 다르더라도 동등한 측정이 가능해야 하고, 내용 균형 등 다양한 설계 제약을 충족하는 것이 필요함을 언급하였다. 본 연구에서 제안한 CORE-OM과 OQ-30의 MST 소검사들은 서로 다른 소검사가 다양한 심각도 수준을 측정할 수 있도록 설계되어야 한다는 원칙을 충족하는 것으로 볼 수 있다. 다만 소검사별 최대 정보량에서는 차이가 있었는데, 문항은행의 규모가 제한적인 경우 다양한 설계 제약을 동시에 충족하는 것이 어렵다는 점을 고려할 때(Zheng et al., 2018), 제한된 문항은행 내에서 목표한 심각도 수준을 우선적으로 고려하는 과정에서 가지는 한계로 볼 수 있다.

둘째, 검사 설계 방식에 따른 측정 정확성을 비교한 결과, MST는 동일한 문항 수의 단축형 고정형 검사보다 더 높은 측정 정확성을 보이는 것으로 나타났다. 또한 심각도 수준이 매우 높거나($\Theta > 2$) 낮은($\Theta \leq -2$) 구간에서는 MST에서 전체 문항 고정형 검사보다도 더 높은 정확성을 보였다. 이러한 결과는 MST가 고정형 검사보다 능력 전체 범위에서 안정적인 성능을 보인다는 선행연구(서민희 외, 2020; Weiss, 2011)와 일치하는 결과이다. 또한 CES-D, BDI, 디지털 정신건강 평가 도구에 CAT를 적용하였을 때, 검사 효율성을 높이면서도 높은 측정 정확성을 유지하였음을 제시한 선행연구(Capon et al., 2026; Gardner et al., 2004; Smits et al., 2011)와도 일치하는 결과이다. 본 연구에서 CORE-OM은 34문항에서 10문항으로, OQ-30은 30문항에서 8문항으로 줄었음에도 참값과 추정치의 상관이 0.9 이상으로 높게 나타났다. 이는 Smits et al.(2011)이 CES-D에서 전체 문항의 약 1/3만으로 전체 검사와 약 0.96의 상관, Gardner et al.(2004)이 BDI에서 전체 문항의 약 1/4만으로 전체 검사와 약 0.92의 상관을 보인 결과와 유사한 수치이다. Gardner et al.(2004)은 문항 수를 줄이는 것이 단순히 검사 시간을 단축하는데 그치는 것이 아니라, 임상 장면에서 다른 문제를 함께 평가할 시간을 확보할 수 있고, 특히 반복적인 모니터링과 같이 응답자의 부담이 중요한 상황에서 적응형 검사의 유용성이 더욱 크다고 강조하기도 하였다. 이러한 결과는 적응형 검사가 상담성과 측정에서 검사 효율성을 개선하기 위한 유용한 방법이 될 수 있는 가능성을 시사하는 결과로 볼 수 있다.

덧붙여, 본 연구에서 확인된 MST의 측정학적 장점은 CORE-OM과 OQ-30 모두에서 유사한 경향을 보였는데, 이는 두 도구 모두 상담성과를 측정하기 위한 목적으로 심리적

증상과 기능 수준 등 유사한 개념을 측정하고 있다는 점에서 유사한 특성을 생각해 볼 수 있다. 더불어 MST 설계 시 적용된 문항 모수 추정 방법, ATA 알고리즘, MST의 단계 수 및 단계별 모듈 수 등 MST 설계 조건 또한 유사했기 때문에 나타난 결과로 해석된다. 한편, 두 도구의 특성에 따른 차이도 일부 확인되었는데, CORE-OM은 특히 매우 낮은 심각도($\theta \leq -2$) 구간에서, OQ-30은 특히 매우 높은 심각도($\theta > 2$) 구간에서 편향이 상대적으로 더 크게 나타나는 경향이 있었다. CORE-OM에는 “삶을 마감할 계획을 세워 놓았다”와 같이 매우 높은 수준의 심리적 고통을 측정하는 문항들이 일부 포함되어 있고, OQ-30에는 “직장이나 학교 혹은 다른 일상 활동에서 스트레스를 받는다”와 같이 비교적 경미한 스트레스와 관련된 문항들이 함께 포함되어 있는데, 이처럼 각 도구가 더 많은 정보량을 가지는 심각도 수준의 차이가 반영된 결과로 해석할 수 있다.

셋째, MST를 활용한 상담성과 변화 탐지 방법을 적용하였을 때, 가상 내담자 사례에 MST를 활용한 변화 탐지가 적용될 수 있음을 확인하였다. MST의 IRT 기반 변화 탐지와 기존 고정형 검사의 CTT 기반 변화 탐지 결과를 비교했을 때, IRT 기반 변화 탐지에서는 개인별 측정오차를 고려하여 변화 판단이 가능하다는 장점이 있었다. 또한 두 방법의 변화 탐지 결과는 대체로 유사한 결과를 보였으나, 기준점에 근접한 사례에서는 두 방법 간 차이를 보이기도 하였다. 이러한 결과는 선행연구에서 두 방법 간 변화 탐지의 결과가 전반적으로 크지 않으나(Jabrayilov et al., 2016), 일부 조건에서 IRT 기반 접근이 CTT 기반 접근보다 개인의 변화를 보다 민감하고 정확하게 탐지할 수 있다는 결과(Tang et al., 2023)와 일치하는 결과로 볼 수 있다. 이는 MST를 활용한 상담성과 측정에서 내담자 변화 탐지가 실제로 구현되고 활용될 수 있는 가능성을 보여주는 결과로 볼 수 있다. 이처럼 실제 상담 장면에서도 <표 5.1>과 같은 예시처럼 MST를 활용하여 내담자별 상담성과 측정 및 신뢰로운 변화 평가가 가능할 것으로 기대된다.

<표 5.1> 상담성과 측정에서 MST 활용 방안 예시

상담성과 측정에서 MST 적용 과정 및 변화 평가 결과 제공 예시(안)	
MST	○ 상담 전: 상담성과 MST 실시 (내담자별 맞춤형 1-2-3단계 문항 제시) → 상담 전 심각도 추정
적용 과정	○ 상담 후: 상담성과 MST 실시 (내담자별 맞춤형 1-2-3단계 문항 제시) → 상담 후 심각도 추정, 상담 전-후 변화 탐지
<상담 전-후 변화 평가 결과>	
	→ 상담 전 심각도: $\theta_{Pre}=2.10$ (백분위 98) 상담 후 심각도: $\theta_{post}=-1.82$ (백분위 3)
변화 평가	→ 상담 전-후 변화의 95% 신뢰구간: $[-5.23, -2.61]$
결과 제공	신뢰할 수 있는 변화 여부: ☒ Reliable improvement ☐ Reliable deterioration ☐ No reliable change
“상담 후 심리적 어려움의 심각도가 상담 전에 비해 감소하였고, 이러한 변화는 측정오차를 고려하더라도 신뢰할 수 있는 변화(reliable change)로 판단됩니다.”	

종합하면 본 연구의 결과 MST는 기존 고정형 검사와 유사한 수준의 측정 정확도를 유지하면서 검사 문항 수를 일부 줄일 가능성을 보여주었다. 상담성과 측정에서의 MST 활용은 상담성과를 더 적은 시간과 적은 문항 부담으로도 안정적으로 측정할 수 있는 가능성을 보여주며, 극단적인 심각도 수준에서 정확성이 향상될 수 있다는 점은 실제 상담 장면에서 민감하게 주의를 기울여야 하는 위험군 평가 및 모니터링에도 활용될 수 있는 가능성을 보여준다. 더 나아가 검사 시간의 단축은 단순히 행정적인 효율성을 넘어, 상담 과정에서 내담자와의 상호작용 및 개입에 더 많은 시간을 할애할 수 있는 환경 조성에 기여할 수 있을 것으로 생각된다. 또한 특히 우울이나 불안 증상이 심하여 피로감이나 무기력감이 크거나, ADHD를 비롯한 주의집중의 어려움 또는 일시적인 혼란으로 인해 30문항 이상의 자기보고식 심리검사에 끝까지 응답하기 어려운 내담자에게는 MST의 효율성이 더욱 큰 장점으로 작용할 것으로 기대된다.

적응형 검사의 측정학적 장점은 여러 선행연구들에서 꾸준히 보고되어 왔지만 (Mead, 2006; Weiss, 2011), 국내에서는 아직 활용이 제한적인 편이다. 이러한 배경에는 여러 요인이 있을 수 있는데, 적응형 검사를 위해서는 문항은행 구축, 시스템 개발 등 초기 구축 비용과 기술적 부담이 크다는 점과, 기존의 고정형 검사 중심의 평가 방식이 실무 현장에서 오랫동안 유지되어 온 점 등을 생각해 볼 수 있다. 그러나 최근에는 교육 분야에서 PISA 등 대규모 평가에 적응형 검사가 도입되고 있으며(조성민 외, 2019), 해외의 정신건강 분야에서도 CAT-MH(Computerized Adaptive Testing for Mental Health)와 PROMIS CAT(Patient-Reported Outcomes Measurement Information System Computerized Adaptive Test) 등이 심리평가 도구로 활용되고 있다(Cella et al., 2007; Gibbons et al., 2016). 국내에서도 심리검사가 지필 방식에서 온라인 방식으로 점차 전환되고 있는 점을 고려할 때, MST 역시 초기의 도입 부담과 저항은 있을 수 있으나 향후 상담성과 측정에서의 활용 가능성을 기대해 볼 수 있다.

또한 상담성과 측정을 포함한 심리평가 장면에서는 CAT에 비해 MST가 가지는 실무적 장점도 생각해 볼 수 있다. CAT에서는 피험자 심각도/능력에 따라 주로 통계적 기준에 따라 문항을 제시하는 반면, MST에서는 단계별 모듈을 사전에 구성할 수 있기에 핵심적인 문항이나 임상적으로 중요한 내용을 미리 포함할 수 있다. 예를 들어 우울을 평가하고자 할 때, 진단적 중요성이 높은 핵심 증상과 관련된 문항이 공통적으로 포함되도록 미리 설계할 수 있다. 이처럼 MST는 시스템 구축과 운영의 용이성뿐만 아니라, 적응형 검사의 효율성을 유지하면서도 내용적인 타당성도 체계적으로 확보할 수 있다는 점에서 중요한 장점을 가진다고 볼 수 있다.

다만 본 연구는 시뮬레이션에 기반한 분석이므로, 이러한 결과를 실제 상담 장면의 효과로 바로 일반화하기에는 제한이 있다. 이에 본 연구에서 확인한 MST의 성능은 탐색적인 수준에서 바라볼 필요가 있으며, 후속 연구에서 실제 상담 장면에서의 응답자료를 바탕으로 교차타당화가 이루어질 필요가 있다. 또한 본 연구에서 사용한 CORE-OM과 OQ-30은 적응형 검사를 위해 개발된 대규모 문항은행이 아니므로(CORE-OM은 전체 34문항, OQ-30은 전체 30문항), 소검사 간 정보량과 내용 균형을 동시에 최적화하는 데에는 한계가 있었다. MST는 충분한 규모의 문항은행을 전제로 설계될 때 그 장점이

극대화될 수 있기에, 향후 상담성과 측정 도구의 문항은행을 확장할 수 있다면 보다 정교한 MST 설계가 가능할 것으로 기대된다. 또한 본 연구에서는 상담성과를 하나의 잠재특성으로 가정하여 MST를 구성하였으나, 실제 상담성과는 여러 다차원적인 하위 영역으로 바라볼 수도 있으며, 향후에는 다차원 IRT 등을 통해 MST 설계 시 상담성과의 다양한 측면을 보다 충실히 반영하는 것도 필요할 수 있겠다. 그럼에도 불구하고 본 연구는 상담성과 측정 분야에서 MST의 적용 가능성을 실증적으로 탐색한 연구라는 점에서 의의가 있다.

MST를 활용한 맞춤형 검사는 기존의 지필형 검사의 긴 역사에 비해 국내에서는 도입 초기 단계에 해당하며, 다양한 후속 연구가 지속적으로 필요할 것으로 생각된다. IT 개발자와의 협업을 통해 MST 기반 상담성과 측정 소프트웨어를 실제로 구현하고 이를 상담 현장에 적용하여 안정성을 검증하는 것도 중요한 후속 과제가 될 수 있다. 또 다른 측면으로, MST를 활용하면 응답자의 부담은 줄어들 수 있지만, 검사자는 MST 시스템 구현과 유지를 위한 추가적인 비용이 발생하게 된다. 따라서 향후 적응형 검사 실제 현장 적용을 위해서는 기술적, 제도적 지원 및 기반 마련이 장기적으로 중요한 과제가 될 것으로 생각된다.

(2026년 6월 27일 접수, 2026년 8월 3일 수정, 2026년 8월 13일 채택)

참고문헌

- 교육부 (2020). 국가수준 학업성취도 컴퓨터 기반 평가 구축 계획(안). 교육부 교육 기회보장과.
- 구남욱, 김미림, 이소라, 박민호, 김동호, 한경택, 김지혜. (2021). 국가수준 학업성취도 평가의 컴퓨터 기반 평가 체제 구축 방안 (연구보고 RRE 2021-1). 한국교육과정평가원.
- 김선경 (2012). 상담서비스 공급자/ 제공자를 위한 상담 성과관리체계 개발: 핵심 상담 성과도구(CORE-OM)를 기반으로. [신진연구자지원사업 결과보고서].
- 김선경, 왕은자 (2016). 한국판 핵심 상담성과도구(The Korean Inventory of Clinical Outcomes in Routine Evaluation-Outcome Measure; K-CORE-OM)의 타당화. <상담학연구>, 17(3), 105-127.
- 김우림, 김창대 (2019). 상담 서비스 이용 의사결정 과정: 원형-의향 모형을 중심으로. <상담학연구>, 20(1), 21-37.
- 박인용, 김미림, 박민호, 박민호, 강혜진, 박도영, 김완수, 임황규 (2023). <국가수준 학업 성취도 평가의 단계적 적응형 평가 체제 구축: 검사 설계 및 척도화> (연구 보고 RRE 2023-8). 한국교육과정평가원.
- 상경아, 김미림, 김완수, 이소라, 박민호, 임황규 (2022). <컴퓨터 기반 국가수준 학업 성취도 평가의 척도 연계 및 문제은행 설계> (연구보고 RRE 2022-1). 한국 교육과정평가원.
- 서동기, 이순목, 김종남, 최승원, 채정민, 정선호, 조성경, 김명기, Chad, E. (2019). 단축형 심리검사 개발의 측정학적 방법과 타당화: 한국형 역기능우울척도를 중심으로. <한국심리학회지: 일반>, 38(1), 75-102.
- 서민희, 김미림, 김완수, 이소라, 한정아, 김성훈, 서영숙 (2020). <컴퓨터 기반 국가수준 학업성취도 평가 도입을 위한 검사 설계 방안> (연구보고 RRE 2020-3). 한국교육과정평가원.
- 손난희, 유성경 (2012). 상담성과 측정을 위한 상담성과척도(Outcome Questionnaire-30: OQ-30)의 타당화. <상담학연구>, 13(1), 1-15.
- 안성식, 심은정, 박승민 (2020). 상담현장에서 활용가능한 초간단 척도(Ultra-brief scale)인 상담성과척도(Outcome Rating Scale: ORS)의 국내 타당화 연구, <상담학연구>, 21(4), 27-46.
- 왕은자, 김선경, 박경희, 김계현 (2014). 상담성과평가와 신뢰할 수 있고 임상적으로 의의있는 지수의 활용가능성. <상담학연구>, 15(3), 1085-1100.
- 왕은자, 유정이, 김선경 (2018). 국내 개인상담 성과의 측정 및 평가에 대한 분석: 학술지 <상담학연구>와 <한국심리학회지: 상담 및 심리치료>를 중심으로. <상담학 연구>, 19(5), 1-23.
- 이하늘, 이용상 (2024). 단계적 적응형 검사(MST)에서 피험자 능력 추정 방법의 정확성 비교, 교육평가연구, 37(2), 325-350.
- 조성민, 구남욱, 김현정, 이소연, 이인화. (2019). <OECD 국제 학업성취도 평가 연구:

- PISA 2018 결과 보고서> (연구보고 RRE 2019-11). 한국교육과정평가원.
- 한정아, 한경택, 서민희 (2022). 국가수준 학업성취도 평가에서 단계적 검사의 적용 가능성 탐색. <교육평가연구>, 35(1), 49-74.
- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, 15(2), 163-181.
- Aybek, E. C. (2023). The relation of item difficulty between classical test theory and item response theory: Computerized adaptive test perspective. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Journal of Measurement and Evaluation in Education and Psychology)*, 14(2), 118-127.
- Barkham, M., Bewick, B., Mullin, T., Gilbody, S., Connell, J., Cahill, J., Mellor-Clark, J., Richards, D., Unsworth, G., & Evans, C. F. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling and Psychotherapy Research*, 13(1), 3-13.
- Barkham, M., Evans, C., Margison, F., Mcgrath, G. Mellor-Clark, J., Milne, D., & Connell, J. (1998). The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health*, 7(1), 35-47.
- Blampied, N. M. (2022). Reliable change and the reliable change index: still useful after all these years?. *The Cognitive Behaviour Therapist*, 15, e50.
- Capon, W., Hickie, I. B., Varidel, M., LaMonica, H. M., Borgnolo, L. J., Crouse, J. J., Scott, E. M., & Iorfino, F. (2026). Validating an adaptive digital assessment of youth mental health needs: A cross-sectional study. *npj Digital Medicine*, 9(1), e173.
- Cella, D., Yount, S.E., Rothrock, N.E., Gershon, R.C., Cook, K.F., Reeve, B.B., Ader, D.N., Fries, J.F., Bruce, B., & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. *Medical Care*, 45, S3-S11.
- Choi, K.-H., Buskey, W., & Johnson, B. (2010). Evaluation of counseling outcomes at a university counseling center: the impact of clinically significant change on problem resolution and academic functioning. *Journal of Counseling Psychology*, 57(4), 297-303.
- Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, 17, 305 - 308.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.
- Evans, C. (2025, October 15). FAQ: How do I interpret LD-CORE scores? Clinical Outcomes in Routine Evaluation (CORE System Trust).
- <https://www.coresystemtrust.org.uk/home/faq/how-do-i-interpret-ld-core-scores/>

- Finch, A. E., Lambert, M. J., & Schaalje, B. J. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology and Psychotherapy*, 8, 231-242.
- Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, 4, 13.
- Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, 12, 83 - 104.
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59(4), 361-368.
- Hays, R. D., Spritzer, K. L., & Reise, S. P. (2021). Using item response theory to identify responders to treatment: examples with the patient-reported outcomes measurement information system (PROMIS®) physical function scale and emotional distress composite. *Psychometrika*, 86, 781-792.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44-52.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive test. *Applied Measurement in Education*, 2(4), 359-375.
- Jacobson, N., & Truax, P. (1991). Clinical Significance: A statistical approach to defining meaningful change in psychotherapy Research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19.
- Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40(8), 559-572.
- Keller, F., & Alexandrowicz, R. W. (2024). Assessing individual change: A comparison of reliable change indices based on classical test theory and various item response theory models. In M. Stemmler, W. Wiedermann, & F. L. Huang (Eds.), *Dependent Data in Social Sciences Research* (pp. 735 - 765). Springer.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606-613.
- Lambert, M. (2001). Psychotherapy outcome and quality improvement: introduction to the special section on patient-focused research. *Journal of Consulting*

- and Clinical Psychology*, 69(2), 147-149.
- Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N., & Vermeersch, D. A. (2005). *The Outcome Questionnaire-30.2*. OQ Measures.
- Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, 55(4), 520-537.
- Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., & Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Evaluation and Program Planning*, 2(3), 197-207.
- Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H., Hoffman, M., Scott, K., Lyon, A., Douglas, S., Simon, G., & Kroenke, K. (2019). Implementing measurement-based care in behavioral health: A review. *JAMA Psychiatry*, 76(3), 324-335.
- Lord, F. M. (1971). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66(336), 707-711.
- Maruish, M. E. (2002). *Psychological testing in the age of managed behavioral health care*. Lawrence Erlbaum Associates.
- Maydeu-Olivares, A. (2021). Assessing the accuracy of errors of measurement: Implications for assessing reliable change in clinical settings. *Psychometrika*, 86(3), 793 - 799.
- McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates.
- Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, 19(3), 185-187.
- Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19(1), 91 - 100.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.
- Reckase, M. D., (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, 4(3), 207-230.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(1), 1-97.
- Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 42(2), 193-198.
- Sari, H. I., Yahsi-Sari, H., & Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 388-406.
- Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice*, 22(1), 49-59.

- Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147-155.
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092-1097.
- Stiles, W. B., & Snow, J. S. (1984). Counseling session impact as viewed by novice counselors and their clients. *Journal of Counseling Psychology*, 31(1), 3-12.
- Stochl, J., Böhnke, J. R., Pickett, K. E., & Croudace, T. J. (2016). An evaluation of computerized adaptive testing for general psychological distress: Combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Medical Research Methodology*, 16, e58.
- Tang, X., Schalet, B. D., Peipert, J. D., & Cella, D. (2023). Does scoring method impact estimation of significant individual changes assessed by patient-reported outcome measures? Comparing classical test theory versus item response theory. *Value in Health*, 26(10), 1518 - 1524
- Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and Evaluation of a Computer Adaptive Test for 'Anxiety' (Anxiety-CAT). *Quality of Life Research*, 16(1), 143-155.
- Wang, W-C., & Chen, P-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28(5), 295-316.
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70-84.
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1-27.
- Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: theory and applications*, (pp. 3-20). Boca Raton, FL: CRC Press.
- Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2018). Overview of test assembly methods in multistage testing. In D. J. Weiss & J. Yao (Eds.), *Multistage Testing* (pp. 115 - 134). Routledge.

Application of multistage adaptive testing in measuring counseling outcomes: Test design and change detection⁸⁾

Jinsil Lee⁹⁾ · Sujin Baeg¹⁰⁾ · Yujung Hong¹¹⁾ · Eun-Yi Cho¹²⁾
· Youngwon Choi¹³⁾ · Ha-Na Jeong¹⁴⁾

Abstract

This study examined the applicability of multistage adaptive testing (MST) for counseling outcome assessment. MST was applied to the CORE-OM and OQ-30, and measurement accuracy was evaluated using Monte Carlo simulations. Results indicated that, under equivalent item-length conditions, MST, which adaptively selects item modules based on respondent severity, tended to show higher measurement accuracy than traditional fixed-form tests where all respondents receive an identical set of items. Notably, for respondents at extreme severity levels, MST showed smaller bias despite using substantially fewer items. In addition, reliable change can be evaluated within the MST accounting for individual measurement error, supporting its potential utility in counseling outcome measurement.

Key words : Change detection. Counseling outcomes. Multistage adaptive testing. Test design

8) This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of (NRF-2025S1A5B5A17014042)

9) First Author and Corresponding author, Chief Representative, M.D.
E-mail: leejinsil326@gmail.com

10) Second Author, Lecturer, Dept. of Education, Seoul National University. Gwanak-ro 1, Seoul, Korea. E-mail: bsj104@gmail.com

11) Third Author, Research Fellow, Gyeonggi Institute of Education.
E-mail: dreamgo5049@gmail.com

12) Fourth Author, Assistant Professor, Dept. of Counseling Psychology, Sun Moon University.
E-mail: eycho812@gmail.com

13) Fifth Author, CEO, Beevelop. E-mail: ian.yw.choi@gmail.com

14) Sixth Author, Researcher, Beevelop. E-mail: jeonghn9408@gmail.com