KOREAN JOURNAL OF PSYCHOLOGY: GENERAL

Vol. 43 NO. 3 2024. 9.

한국심리학회지 일반

ISSN 2734-1127(Online)

한국심리학회지
일반

43권 3호 (2024년 9월)

KOREAN JOURNAL OF PSYCHOLOGY: GENERAL

43권 3호

Contents

Mixed Methods Research on the Effectiveness of Counseling on Youth who are in Readiness for Self-Reliance ..... HaeYoun Choi · JeeSung Baek

Overall evaluation of structural equation models and reflection on effect size and continuity ..... So-Hyun Yoo · Su-Young Kim

How to analyze group difference in change: Comparing difference score model and analysis of covariance model ..... Youngsoo Lee · Hye Won Suk

A Systematic Review of Translating and Adapting Psychological Test: Practices and Recommendations ..... Mirim Kim · Yeji Im

Published by
THE KOREAN PSYCHOLOGICAL ASSOCIATION

목 차

자립준비청년 상담 효과에 관한 혼합연구 ..... 최해연 · 백지성

구조방정식 모형의 전반적인 평가 및 효과크기와 연속성에 대한 숙고 ..... 유소현 · 김수영

반복측정 자료에 기반한 변화의 집단차 분석 방법: 차이점수 모형과 공분산분석 모형 비교 ..... 이영수 · 석혜원

번안 심리검사 타당화 작업에 대한 체계적 검토: 검수와 보고 관행에 대한 검토와 제언 ..... 김미림 · 임예지

한국심리학회

한국심리학회지: 일반 2024년 제43권 제3호

발 행 처: 한국심리학회 인쇄일: 2024년 9월 25일
발 행 인: 최훈석(성균관대학교) 발행일: 2024년 9월 25일
주 소: (04778) 서울시 성동구 뚝섬로 1길 25 서울숲 한라에코밸리 906호 제작처: 책과공간(02-725-9371)
전 화: 02-567-0102 팩 스: 02-738-0104
홈 페 이 지: https://www.koreanpsychology.or.kr

'한국심리학회지: 일반'은 한국심리학회에서 발간하는 학술지로서 연 4회(3월 25일, 6월 25일, 9월 25일, 12월 25일) 간행되며, 심리학 분야의 창의적인 이론연구, 논쟁을 정리하는 개관연구, 심리학의 여러 하위분야의 공통적 관심이 될 수 있는 실증연구, 그리고 측정 및 연구방법론의 논문을 게재합니다. '한국심리학회지: 일반'은 무료로 배포합니다.

Vol. 43, No. 3. **September 25, 2024**

Korean Journal of Psychology: General
by The Korean Psychological Association

Korean Journal of Psychology: General, issued four times a year, publishes theoretical papers, empirical research crossing subdisciplines, and measurement and research methodology. Inquiries concerning the subscription for the journal and the submission of manuscripts should be sent by e-mail to the Editor, Korean Journal of Psychology: General, edit@kpsy.or.kr

The Korean Psychological Association
#906, Seoul Forest Halla Eco Valley, 25 Ttukseom-ro 1-gil, Seongdong-gu, Seoul, S. Korea

이 학술지는 2023년 대한민국 교육부와 한국연구재단의 학술지 지원사업의 지원을 받아 발간되었음(NRF-2023-2023S1A8A1097597)

「한국심리학회지: 일반」 투고논문 작성 안내

(1) 학회지 게재논문의 성격: 본 지에서는 심리학의 발전에 기여하는 창의적인 이론연구, 논쟁을 정리하는 개관연구, 측정 및 연구방법론 논문, 또는 실증연구를 게재할 수 있다. 실증연구의 경우에는 심리학의 여러 하위분야의 학자들에게 공통적인 관심이 될 수 있는 실증연구들로 게재를 한정한다. 특히, 자기보고라는 단일 방법에 의한 1회성 설문조사 자료에 기반한 연구(single source, cross-sectional, self-report survey design 연구) 논문은 특별한 경우가 아니라면 심사 대상에서 제외한다. 아울러 다양한 분과학회에 걸쳐 공통 관심사가 되는 주제(예, 학회가 개최하는 학술심포지움의 주제)를 가지고 특집을 꾸밀 수도 있다. 끝으로 해당호에서 논쟁적인(controversial) 글에 대하여 편집위원회 주관하에 논평을 받고 그에 대한 저자의 반론을 싣는 것을 시도할 수 있다.

(2) 논문작성의 언어: 한글 논문을 원칙으로 하나 영어 논문도 게재 가능하다.

(3) 논문작성 소프트웨어는 반드시 HWP를 사용한다. 논문작성의 상세 양식은 한국심리학회 저술 <학술논문 작성 및 출판 지침 2판(2012, 박영사)> 과 APA논문 작성 스타일을 따라 작성하며, 2020년 6월부터 투고되는 논문은 참고문헌과 본문 안의 참고문헌 인용 표기를 모두 로마자로 하며 APA 표기법을 따른다.

(4) 모든 연구논문은 150단어(600자) 안팎의 본문앞 초록과 참고문헌뒤 초록, 5개 이내의 주요어를 포함해야 한다. 본문 앞 초록은 본문과 같은 언어를, 참고문헌 뒤 초록은 본문과 다른 언어를 사용한다. 예로서, 본문이 영어논문이면 참고문헌뒤 초록은 한글로 한다. 영문초록은 미국심리학회의 논문 데이터베이스인 'PsycINFO'에 실리므로 미국심리학회(APA) 출판 규정에 맞게 쓰여야 한다.

(5) 논문의 길이는 15-20페이지 이내를 권장한다.

(6) 본문은 휴먼명조체 10호 크기로 하고, 장평 95, 자간 -10, 줄간격 160으로 하여 작성한다. 본 학회지의 한 페이지에는 한글로 약 1,800자, 영문으로 약 3,700자(약 500단어)가 들어감을 고려하여 원고를 작성한다.

(7) 그림이나 표가 있는 경우 HWP file에서 본문과 함께 바로 열릴 수 있는 형식으로 작성한다 그림이나 표는 각 페이지의 상단 또는 하단에 밀착한다. 그림은 흑백으로 작성하여 명료하게 인쇄될 수 있어야 하며, 흐린 선, 가는 점선, 계조 흑백(예를 들어, 회색), 색채 등은 인쇄상의 문제가 있으니 피하기로 한다.

(8) 논문의 접수: 한국심리학회지: ACOMS+ 투고 시스템
(https://acoms.accesson.kr/kpageneral/oprs/main/jrnlMain.do)에 회원가입 및 로그인하여 투고한다.

(9) 문의사항 접수: 논문 양식과 관련한 질문이 있을 경우에는 편집위원회 이메일로 연락한다.
(E-mail: edit@kpsy.or.kr)

한국심리학회지

일 반

제 43 권 제 3 호

자립준비청년 상담 효과에 관한 혼합연구
최해연 · 백지성 ····· 173

구조방정식 모형의 전반적인 평가 및 효과크기와 연속성에 대한 숙고
유소현 · 김수영 ····· 199

반복측정 자료에 기반한 변화의 집단차 분석 방법:
차이점수 모형과 공분산분석 모형 비교
이영수 · 석혜원 ····· 231

번안 심리검사 타당화 작업에 대한 체계적 검토:
검수와 보고 관행에 대한 검토와 제언
김미림 · 임예지 ····· 261

부록 1. 논문게재 관련서류 ····· i
부록 2. 논문작성양식 ····· iii
부록 3. 임원진 ····· vi

한 국 심 리 학 회

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 173-197
http://dx.doi.org/10.22257/kjp.2024.9.43.3.173

자립준비청년 상담 효과에 관한 혼합연구*

최 해 연　　백 지 성†

충북대학교 심리학과/교수

자립준비청년의 정신건강을 위한 근거기반 개입과 체계적인 정책 마련을 위한 실증적 근거가 요구되는바, 본 연구는 자립준비청년을 위한 전문적인 상담의 필요성과 효과성을 검토하였다. 연구 1에서는 자립준비청년 31명(평균 나이 21.81세, 여성 18명)이 상담심리사 1급 보유 상담자로부터 평균 9.90(SD=2.13)회 상담을 받은 전후의 변화를 검증하였다. 상담 후 자립준비청년의 심리증상이 유의하게 감소하였고, 자아정체감과 자아존중감이 유의하게 증가하였다. 주요 호소 문제 수준이 유의하게 감소하고 높은 상담만족도를 보고하였다. 연구 2에서는 자립준비청년의 상담 경험을 질적 연구하여 상담 효과의 임상적 의미를 탐색하였다. 연구 1에 참가한 자립준비청년 5명의 상담 경험을 분석하여 추출한 20개 개념을 공감적 관계에서 솔직함, 나를 알아가는 시간, 비합리적 신념에서 벗어남, 긍정성의 증가의 4개 범주로 구조화하였다. 공감적 상담 관계는 정서적 안정과 있는 그대로 자기표현을 가능하게 하고, 이는 알아차림, 의식적인 성찰, 수용과 통합을 촉진함으로써 심리증상 완화와 주체적인 자기개념을 강화하였다. 상담 효과의 핵심 요인으로서 상담 인력의 전문성 확보의 필요성을 논의하였다.

주요어 : 자립준비청년, 보호종료아동, 상담 효과, 상담인력 전문성 기준, 상담 요인.

---

* 이 논문은 2022학년도 충북대학교 학술연구지원사업의 연구비 지원을 받아 수행된 연구이며, 연구 1은 교신저자의 석사학위 논문 자료를 사용하였음.

† 교신저자: 백지성, 성신여자대학교 성신인권센터 인권상담소 전문위원, (02844) 서울 성북구 보문로34다길 2, 성신여자대학교 수정캠퍼스 성신별관 102호, E-mail: jsbaek131@gmail.com

보호자의 부재나 학대, 경제적 어려움 등으로 가정 내 보호를 받지 못하는 아동은 가정 외 보호 조치를 받게 되고, 만 18세에 보호가 종료되어 자립을 시작해야 한다. 이들은 보호종료아동 등으로 불리다 2021년부터 자립준비청년으로 명명되었다(보건복지부, 2021). 보호종료가 이루어지는 후기청소년기의 이들은 학업, 진로, 취업, 대인관계 등 수많은 삶의 과제를 스스로 결정하고 책임지는 적응 과정에서 많은 스트레스를 경험한다(Arnett, 2015). 특히 일반 가정에서 성장한 청년과 비교할 때 자립준비청년들은 초기 외상 및 취약한 삶의 조건으로 인해 더욱 극심한 스트레스를 겪을 수 있다(Cunningham & Diversi, 2013).

자립준비청년들은 보호종료 이후 자립 과정에서 주거, 학업, 취업, 경제적 문제로 생활고를 겪거나(정익중, 2007) 홀로 삶을 책임지는 부담, 자립에 실패할 수 있다는 불안과 우울을 경험한다(홍예영, 김유숙, 2020). 외로움은 이들이 직면하는 또다른 큰 어려움이다. 다양한 생활 문제를 겪어내는 과정에서 자립준비청년은 불안, 우울, 자살사고 등 심리정서 상태의 취약성을 나타내지만, 시설 및 가족과의 관계가 단절되어, 취약성을 완화할 사회적 지지체계는 부족하다(이상정, 김지민, 안은미, 김무현, 2020).

한편, 원가족과 미해결 문제는 이들의 적응을 저해하는 위험요인이다(장혜림, 정익중, 2017; 장정은, 전종설, 2018). 보호아동 대부분은 부모의 빈곤, 이혼, 사망 등 고통스러운 가정환경을 경험하였고, 최근에는 학대나 가정폭력으로 인해 보호를 받게 된 경우가 더욱 증가하였다(통계청, 2020). 원가족과 관련된 심리적 외상, 부모로부터 버려졌다는 박탈감과 분리 경험으로 인한 정서적인 결핍, 사회적 편견과 반복된 대인관계 상처 등이 치유되지 못한 채 방치된 경우가 많으며(고수안, 2023; 박신애, 최옥채, 2018; 이정우, 이소연, 2023; 황수연, 2018) 이는 심리적 문제로 나타나고 건강한 성인기로의 전환을 방해한다. 자립준비청년은 자신이 자라온 환경을 숨기려는 등 심리적 위축을 나타내고, 정신건강의 위험성이 높다(정익중, 김주현, 2019; Reilly, 2003). 이러한 시기에 반복적인 외상 경험은 신체화, 해리와 같은 심각한 심리증상, 정체성의 병리적 변화에 영향을 미칠 수 있다(Herman, 1992). 자립준비청년의 자살사고가 일반 청소년보다 3배 높다(이상정 외, 2021)는 결과는 이들의 환경적, 심리적 취약성을 가늠하게 한다.

이러한 자립준비청년의 취약한 현실에 대한 국가적 관심이 쏠리며, 최근 자립준비청년 지원을 위한 법적 근거가 마련되고 보호기간 연장, 자립정착금 및 자립수당 확대, 주거 안정 등 지원 범위도 확대되고 있다(보건복지부, 2022a; 보건복지부, 2022b). 경제・주거・교육・일자리 분야의 지원이 강화된 것은 바람직하나, 한편으로 지원 정책들이 경제적 지원에 치우쳐 있어, 성인기 전이 과정을 조력할 심리서비스의 지원이 부족하다(이상정 외, 2019; 제철웅, 장영인, 2019)는 지적은 여전히 유효하다. 경제적 자립을 위해서도 심리적 역량은 뒷받침되어야 한다(김예성, 이경상, 2015). 심리적 문제를 효과적으로 다루지 못할 경우, 진로탐색과 취업의 어려움은 물론, 일자리 유지나 자산관리에 어려움을 겪을 수 있다. 또한, 심리적 문제의 방치는 정신장애의 중증화와 만성화를 초래하며, 이는 개인의 고통을 넘어 사회적 비용을 초래한다. 이에 자립준비청년의 건강한 성인기 전환을 위해서 심리치료가 필

요하다는 점은 지속적으로 강조되어 온 바이다(이정애, 2018; 장윤정, 2013).

최근 들어 정신건강의 중요성이 인식되며, 2022년부터는 청년마음건강지원사업(보건복지부, 2022b) 등, 청년의 정신건강을 지원하는 정책들이 제안되기 시작하였다. 국가 차원의 심리서비스 제공에 있어 근거기반 개입(Barkham, Hardy, & Mellor-Clark, 2010)은 필수적일 것이다. 상담이 심리증상을 개선하는 등 유의미한 효과를 나타내는지, 어떠한 경험을 통해 상담효과가 발생하는지를 밝히는 실증연구는 실효성 있는 정책개발과 전문적인 서비스개입의 근거가 될 것이다. 이러한 맥락에서 본 연구에서는 2020년 민관협력 사업으로 수행되었던 자립준비청년 상담의 효과를 검증하고자 한다.

본 연구는 내담자 경험을 중심으로 객관적이고 신뢰로운 자료 수집, 상담 개입의 전문성 확보, 타당화된 측정도구 사용, 상담효과에 대한 통계적 분석을 통해 상담 효과를 실증적으로 확인하고, 질적연구를 병행하여 상담 효과의 함의를 구체화함으로써 상담 개입의 근거를 제시하고자 하였다.

첫째, 일반적으로 상담의 효과는 상담성과를 측정함으로써 파악된다. 상담성과는 상담회기 중 나타나는 내담자의 체험, 통찰, 감정표현 등의 변화뿐만 아니라 상담 진행 중이나 종결 후에 나타나는 내담자의 정서, 행동, 신념 및 사고의 변화 등을 포함한다(전용오, 2000). 그런데 상담의 효과는 직접 관찰이 어려우므로 상담자의 임상적 평가로 추정하는 경우가 많았다(손난희, 유성경, 2012). 그러나 상담자의 임상적 평가는 내담자 증상 악화를 정확하게 예측하지 못하거나(Lambert, 2007) 내담자가 평정한 상담 성과와 상관이 낮은(Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001) 등 정확성에 한계가 있었다. 이에 본 연구에서는 상담의 효과 검증을 위해 내담자 경험과 보고를 측정하였다. 또한, 자립준비청년이라는 동질 조건의 연구참가자를 대상으로 하며, 통계적 검증이 가능한 수의 자료를 수집하였다.

둘째, 상담 연구에서 전문성을 갖춘 상담자에 의한 전문적 상담을 제공하는 것은 무엇보다 중요하다(김계현, 2002). 상담과 같은 심리서비스는 관련 인력의 전문 역량이 그 질을 결정하며(OECD, 2015) 고도의 전문성을 요구한다. 상담에서의 전문성은 제도적으로 공인된 상담교육과 검정 과정을 거쳐 자격을 취득한 상담자가 학문적 기초와 실증적 이론에 근거하며, 효과적인 절차와 방법을 활용하고, 윤리적인 규범을 준수하며 상담목표를 달성해 나가는 과정(김인규, 2021)이라 할 수 있다. 그런데 상담자 발달에 관한 수많은 연구에서 이러한 역량과 전문성을 구분하는 기준은 아직 명확하지 않다(금명자, 정상화, 2021). 따라서 상담 관련 교육 및 학력 검증과 함께 상담, 검사, 수퍼비전 등 다양한 수련 요소의 수준을 검증하는 제도적 장치인 자격증(윤희섭, 정현희, 2010)이 상담자의 전문성을 객관적으로 판단하는 대안으로 사용되고 있다. 국내 공신력이 인정된 상담 관련 자격증은 국가자격인 보건복지부 정신건강임상심리사와 여성가족부 청소년상담사와 더불어, 민간자격으로 (사)한국심리학회/한국상담심리학회에서 발급하는 상담심리사가 대표적이다(강창욱, 이동환, 정진영, 박장군, 2022). 상담심리사 자격제도는 상담심리학 유관 전공 석사학위 취득 후 최소 3년간 실무수련, 최소 400시간의 개인상담과 이에 대한 수퍼비전, 집단상담, 심

리평가 등 실무요건이 자세하게 명시되고 엄격한 자격증의 질 관리(안성희, 성현모, 김보람, 이상민, 2022)를 통해 상담자의 전문성을 확보하고 있다. 본 연구에서는 (사)한국심리학회/한국상담심리학회의 상담심리사 1급 자격증 취득자로 상담자 자격을 제한하여, 상담 처치의 전문성을 확보하였다.

셋째, 상담으로 인한 변화를 과학적으로 증명하기 위해서는, 스트레스 및 증상의 감소, 기능 수준의 향상 등과 같은 내담자의 실제적 변화를 개념화할 수 있는 객관적인 측정도구를 사용해야 한다(왕은자, 유정이, 김선경, 2018). 이에 본 연구는 체계적인 문헌검토를 통하여 자립준비청년의 적응과 밀접한 관련이 있는 심리적 변인을 선별하고, 신뢰도와 타당도가 검증된 측정도구를 사용하여 이를 측정하였다.

연구 1에서는 상담이 실제적인 효과가 있었는지를 검증하기 위하여, 상담만족도 및 호소문제 수준의 변화와 함께 심리증상, 자아정체감, 자아존중감의 변화를 확인하였다. 이어 연구 2에서는 질적연구를 실시하여 상담 과정에서 겪은 경험과 변화를 탐색하고 자립준비청년이 보고한 상담 효과의 구체적인 의미를 밝히고자 하였다. 상담이 자립준비청년에게 실제적인 도움을 주는지, 어떻게 도움이 되는지에 대한 깊이 있는 이해는 효과적인 정책 수립에 필수적일 뿐 아니라, 심리서비스를 제공하는 전문가에게도 구체적인 참고자료가 될 것이다.

연구 1. 자립준비청년의 상담 효과 검증: 심리증상, 자아정체감, 자아존중감의 변화

(사)한국상담심리학회는 2019년 아동권리보장원과 업무협약을 맺고, 자립준비청년 심리상담 지원 사업을 3년간 진행하였다. 사업 첫 해는 15명의 자립준비청년이 상담을 지원받았다. 이후 자립준비청년을 위한 상담이 필요하다는 인식이 확대되어 2020년 57명이 상담 지원을 받았으며, 2021년은 국가 예산이 마련되며 102명의 자립준비청년이 상담을 받았다. 본 연구는 2020년 상담을 받았던 자립준비청년들을 대상으로 상담 경험과 변화를 확인하였다.

상담으로 인한 변화를 과학적으로 증명하기 위해서는 내담자의 변화를 개념화할 수 있는 변인과 이에 대한 객관적 측정도구가 필요하다(왕은자, 유정이, 김선경, 2018). 본 연구에서는 프로그램 효과 검증 연구들이 일반적으로 사용하는 상담만족도나 호소문제 수준의 변화와 함께, 성공적인 자립의 핵심적인 변수인 심리증상의 감소, 그리고 중장기 발달의 기초가 되는 자기개념의 변화를 측정하였다.

심리증상의 감소는 효과적인 상담의 성과로 일관되게 보고되는 지표(김계현, 2002)이며 자립준비청년 상담의 주요한 목표이다. 자립준비청년은 가족관계에서의 외상, 결손가정 출신이라는 낙인감, 장기보호 과정에서의 심리적 취약성과 자립 과정에서의 어려움(이상정 외, 2020)으로 외로움, 분노, 불안, 우울, 자신감 상실, 자살사고 등 다양한 심리증상을 나타낸다(이태연, 최은숙, 이세정, 2019; 장혜림, 정익중, 2017). 이러한 심리증상은 그 자체로 부적응을 나타내며 성공적인 자립을 저해하는 위험요소이다(이정애, 2018; 장윤정, 2013). 이에 심리증상의 개선은 상담 효과성을 의미할 것이다.

대조적으로 자기개념을 대표하는 자아존중감과 자아정체감은 자립준비청년의 회복탄력

성(regilience)과 밀접하게 관련이 있고(강현아, 2010), 자립준비에 긍정적인 영향을 미치는 것으로 알려져 있다(김예성, 이경상, 2015; 이정애, 이화조, 정익중, 2017; 정선욱, 2015). 자아존중감은 자신에 대한 가치판단으로, 개인의 발달에 중요한 영향을 미친다(Coopersmith, 1981). 보호아동들의 불안정, 행동과잉, 집중력 결여, 관계문제, 학업성취 저조, 두려움, 과도한 애정 갈망 등의 심리행동 문제들은 낮은 자아존중감과 같은 취약한 자기개념과 연결된다(이미혜, 2002). 불확실한 자아정체감 역시 역할이나 가치관 혼란으로 인해 불안이나 우울 등 부적응과 연결된다(Erikson, 1968). 청소년기 자아정체감 형성은 현재의 적응뿐 아니라, 이후 성격발달과 성인기 사회적응에 영향을 미치는 중요한 변수이다(박아청, 2002).

이에 본 연구는 자립준비청년 자립의 위험요인인 심리증상과 함께 보호요인인 긍정적 자기개념 변인의 변화를 통해 상담의 효과를 검증하였다. 더불어 상담의 주요한 목표가 내담자가 호소하는 문제와 고통의 감소임으로 내담자의 호소문제들을 반복 측정하여 호소문제 수준의 변화를 확인하고, 선행연구(강유임, 서선아, 2023; 김명찬, 이현진, 2016)에서 상담효과성 지표로 사용되어온 상담만족도를 포함하였다. 2020년 상담을 받은 자립준비청년을 대상으로 연구참가자를 모집하는 목적적 표집방법(purposive sampling)을 사용하였고, 단일집단 사전사후설계(one group pretest - posttest design)로 상담 효과를 검증하였다.

방 법

연구참가자

내담자

최초 57명 상담 신청자 중 35명(61.40%)이 10회기 만기종결, 5명(8.77%)은 합의 조기 종결하였다. 이 중 31명의 자립준비청년이 상담 사전, 사후 설문에 모두 응답하여 연구에 참여하였다. 상담 시작 전, 연구목적 및 연구참가자의 권리를 설명하고 연구참가동의를 받았다. 본 연구는 대학 기관윤리위원회의 심의를 거쳤다(20-3-R-12). 연구참가자의 평균 연령은 만 21.81세(*SD*=2.58)이고, 18세부터 30세까지 분포하였다. 여성이 18명이고 남성이 13명이었다. 직업은 학생 11명, 무직 9명, 시간제 근무 7명, 전일 근무가 4명 순이었다. 학력은 대학교 졸업 이상 11명, 대학교 재학 11명이 가장 많았고 고등학교 졸업 11명, 중학교 졸업이 1명으로 나타났다. 상담회기는 8-10회기를 진행한 이가 22명, 4-6회기 조기종결이 5명, 11-15회기 참여한 이가 4명이었다.

상담자

본 연구에서 상담은 공인된 전문가에 의해 제공되었다. (사)한국심리학회/한국상담심리학회 상담심리사 1급 자격 취득자를 대상으로 자립준비청년 심리지원 상담자 모집공고를 하였고 42명이 상담을 진행하였다. 본 연구에 참여한 상담자는 26명이었다. 상담자들은 상담심리사 1급 자격 취득 이후 평균 6.43년(*SD*=4.62)의 상담경력을 보유하였다. 상담자 나이는 평균 46.04세(*SD*=6.66)이고 여자가 24명(92%)이었다. 상담자에게 사업 및 연구와 관련된 설명문을 보내고, 상담경력 및 인구통계학적 정보에 대한 이용 동의를 받았다. 연구자는 상담 사업에 참여하여 연구참가자의 특성과 경험에 대한 민감성을 높였다. 연구자가 상담한 내담자의 자료는 연구에 포함되지

않았다.

측정 도구

심리증상

심리증상을 측정하기 위해 BSI-18(Brief Symptoms Inventory-18)을 사용하였다. Derogatis (2001)가 개발한 간이정신진단검사에서 18개 문항을 선별하여 국내 대학생을 대상으로 타당화한 단축본(박기쁨, 우상우, 장문선, 2012)이다. 측정하는 심리증상은 공황 3문항, 신체화 5문항, 불안 4문항, 우울 6문항으로 구성되어 있고, 5점 Likert 척도로 평정한다. 본 연구에서 내적합치도(Cronbach's α)는 .94이고. 하위요인의 내적합치도는 .71 ~ .80이다.

자아정체감

송현옥(2008)이 개발한 자아정체감 척도를 한국청소년정책연구원(2014)에서 보완한 수정본을 사용하였다. 자기수용성, 미래확신성, 목표지향성, 주도성, 친밀성을 측정한다. 총 8개 문항으로 구성되었으며, 5점 Likert 척도이다. 송현옥(2008)의 연구에서 내적합치도는 .93이었고, 본 연구에서는 .74이었다.

자아존중감

Rosenberg(1965)가 개발하였고 이훈진과 원호택(1995)이 번안한 한국판 자아존중감 척도를 사용하였다. 문항은 '나는 내가 적어도 다른 사람들만큼 가치 있는 사람이라고 생각한다' 등을 포함한 10문항으로 구성되고, 5점 Likert 척도이다. 본 연구에서는 내적합치도는 .92이었다.

주요 영역 호소문제

상담을 받기를 원하는 주호소 문제를 영역별로 나누어 그 수준을 측정하였다. 대인관계, 가족, 이성・성, 정서, 성격, 행동・습관, 학업・진로・일, 건강, 경제 영역에 현재 스스로 문제가 있다고 느끼는 정도를 표시한다. 5점 Likert 척도(1 = '아주 좋음', 2 = '좋음', 3 = '보통', 4 = '나쁨', 5 = '매우 나쁨')이며, 본 연구에서 내적합치도는 .84이었다.

상담만족도

상담 종결 후 상담에 대한 전반적 만족도를 측정하였다. 본 연구를 위해 제작한 '상담에 대해 얼마나 만족하십니까?' 단일 문항으로, 5점 리컬트 척도('1=매우 불만족', '5=매우 만족')이다.

분석 방법

SPSS 21.0을 사용하여 자료를 분석하였다. 연구참가자의 인구사회학적 특성을 파악한 후 심리증상, 자아정체감 그리고 자아존중감이 심리상담 전후 차이가 있는지를 확인하기 위해 비모수 검정(Wilcoxon Signed Rank)을 실시하였다. 단일집단 사전사후 설계로 통제집단이 없기 때문에, 개입으로 인한 실질적 변화를 파악하기 위해 Cohen이 제안한 효과크기(Effect Size; *ES, d'*)[1]를 구하였다. 효과크기가 .80이상이면 효과가 매우 큰 것이며, .51-.79이면 중등도 유의성으로 해석한다 (Cohen, 1988).

---

1) *ES(d')*=(사후점수-사전점수)/(사전사후점수 차이의 표준편차)

결 과

연구참가자는 평균 9.42회기(*SD*=2.28) 상담을 받았고, 평균 0.48회기(*SD*=.89) 사전연락 없는 결석을 하였다. 상담만족도는 평균 4.80점(*SD*=.40)으로 높게 나타났다. 진행된 상담회기 및 상담만족도는 표 1과 같다.

표본수가 50명 미만이므로 정규성 검증을 위해 Sapiro-Wilk 검증을 적용하였다(D'Agostino & Stephens, 1986). 그 결과, 사전검사에서 심리증상, 자아정체감, 자아존중감 척도 모두에서 정규성이 충족되지 않았다. 이에 비모수검정을 실시하였다.

상담 사전검사에서의 각 변인의 평균 및 상관을 살펴보았다. 상관분석 결과 참가자의 나이는 어떤 변인과도 유의미한 상관을 나타내지 않았다. 심리증상은 내담자의 호소문제 수준(r=.81, p<.001)과 유의한 정적 상관을 나타냈고, 자아정체감(r=-.54, p<.001) 및 자아존중감(r=-.75, p<.001)과 강한 부적 상관을 보였다. 자아정체감은 자아존중감(r=.78, p<.001)과 강한 정적 상관을 나타냈다. 심리증상을 구성하는 불안, 우울, 공황, 신체화의 하위요인들 사이에 .71~.82의 강한 상관관계가 존재하였다. 변인 간 상관관계는 표 2에 제시하였다.

표 1. 연구 참가자의 상담회기 및 상담만족도

(N = 31)

구분	연구참가자	상담만족도 M(SD)
만기종결	27	4.77(.42)
조기종결	4	5.00(.00)
전체	31	4.80(.40)

상담의 효과를 나타내는 사전사후 변화 정도를 구체적으로 살펴보았다. 사전검사에서 심리증상은 평균점수 2.52(*SD*=.93)였으나, 사후검사의 평균은 1.93(*SD*=.89)로 나타나 심리증상의 감소가 유의하였다(Z=-4.19, p<.001). 상담 전후 자아정체감의 변화(Z=-3.74, p<.001)와 자아존중감의 변화(Z=-3.52, p<.001)도 유의하였다. 모든 상담 효과 지표에서 효과크기는 매우 크게 나타났다(d= .83~.97). 상담 전 · 후 차이 검증 결과는 표 3에 제시하였다.

심리증상의 4가지 하위요인인 공황, 신체화, 불안, 우울에 대해 살펴본 결과, 상담 전후 공황의 변화(Z=-3.31, p<.001)와 신체화의 변화(Z=-2.84, p<.01), 불안의 변화(Z=-3.37, p<.001), 우울의 변화(Z=-4.26, p<.001)가 모두 유의하였다. 상담받기를 원했던 영역별 문제 수준도 유의하게 감소하였다(Z=-4.21, p<.001). 각각의 효과 크기 또한 크게 나타났다(d=

표 2. 변인 간 상관관계(N = 31)

변수	1.	2.	3.	4.	5.
1. 나이	-				
2. 심리증상	-.12	-			
3. 자아정체감	.05	-.54**	-		
4. 자아존중감	.28	-.75***	.78***	-	
5. 주요 영역 문제인식	-.19	.81***	-.68***	-.79***	-

주. **p<.01, ***p<.001

표 3. 상담 전·후 효과 검증(N = 31)

영역	사전		사후		*Z*	*ES(d‘)*
	M	*SD*	*M*	*SD*		
심리증상	2.52	.93	1.93	.89	-4.19***	-.91
자아정체감	2.43	.69	2.72	.71	-3.74***	.83
자아존중감	3.07	1.00	3.59	1.00	-3.78***	.83
주요 영역 문제인식	3.25	.69	2.75	.69	-4.21***	-.97

주. ***p<.001

표 4. 심리증상 하위요인의 상담 전·후 차이(N = 31)

영역	사전		사후		*Z*	*ES(d‘)*
	M	*SD*	*M*	*SD*		
공황	2.71	1.22	2.08	1.10	-3.31***	-.69
신체화	1.97	.83	1.61	.80	-2.84**	-.57
불안	2.48	1.08	1.95	1.05	-3.37***	-.69
우울	2.91	1.02	2.12	1.07	-4.26***	-1.01

주. **p<.01, ***p<.001

표 5. 영역별 문제 수준의 상담 전·후 차이(N = 31)

영역	사전		사후		*Z*	*ES(d‘)*
	M	*SD*	*M*	*SD*		
대인관계	2.77	.99	2.39	.88	-2.56*	-.51
가족	3.45	1.21	3.03	1.14	-2.70**	-.55
이성·성	3.19	.98	3.03	.84	-.68	-.14
정서	3.42	1.15	2.71	1.10	-3.25**	-.71
성격	2.97	1.11	2.42	1.06	-2.87**	-.59
행동습관	3.32	.98	2.74	1.00	-2.52*	-.51
학업·진로·일	3.58	1.06	2.68	.91	-3.96***	-1.04
건강	3.03	1.02	2.65	1.14	-2.22*	-.42
경제	3.55	.93	3.13	1.12	-2.39*	-.45
전체	3.25	.69	2.75	.69	-4.21***	-.97

주. **p<.01 ***p<.001

.69~1.04). 심리증상 하위요인의 상담 전·후 차이는 표 4에, 삶의 주요 영역에서 내담자가 느끼는 문제의 상담 전·후 차이 비교는 표 5에 제시하였다. 연구참가자들이 호소하였던 영역별 문제 수준은 대부분 상담 이후 유의하게 감소하였다. 상담 전 연구참가자들은 학업·진로·일, 경제, 가족, 정서, 행동습관에 상대적으로 문제가 많고 어려움을 느낀다고 보고하였다. 상담 후, 학업·진로·일 영역에서 문제가 있다고 느끼는 정도가 가장 많이 감소하였고 가장 큰 효과크기를 나타냈다. 정서 영역에서도 큰 폭으로 문제의 감소를 보고하였고 그 외 성격, 가족, 행동습관에서 유의한 문제의 감소를 보고하였다. 대인관계, 행동습관, 건강, 경제 영역에서는 유의하지만 약한 문제의 감소를 보고하였다. 이성·성 영역에서는 변화가 유의하지 않았다.

연구 2. 자립준비청년이 경험한 상담 과정에서의 변화

연구 1은 양적 연구설계를 통해, 상담으로 인한 자립준비청년의 변화가 유효한지 검증하였다. 평균 9.42회의 상담을 받은 자립준비청년은 상담에 대한 높은 만족감, 주요 영역에서 문제 감소, 심리증상 감소, 그리고 자아정체감 및 자아존중감의 증가를 보고하였다.

상담효과 연구는 대부분 이처럼 상담 전후 또는 집단 비교의 연구설계를 사용하여 그 효과를 검증한다. 이는 상담효과를 입증하는 강력한 방법이지만, 한편으로 한계 역시 갖는다(Lambert & Vermeersch, 2008). 예를 들어, 상담을 받은 개별 참가자가 나타내는 반응의 다양성을 드러내거나, 상담 이후 나타난 효과의 임상적 의의를 제공하지 못한다(왕은자, 유정이, 김선경, 2018). 이에 양적 연구설계 이외에도 임상적 관찰, 질적 연구, 체계적 사례 연구, 메타분석 등 다양한 연구법이 적용될 필요가 제기되어 왔다(임민경 외, 2013; Wampold, 2007). 따라서 연구 2에서는 연구 1에서 확인된 변화의 구체적인 내용, 그 임상적 의미를 질적 연구를 통해 확인하고자 한다.

상담의 효과가 무엇인지 깊이 있게 탐구하기 위해서는, 연구참가자의 경험을 직접 듣고 이해할 필요가 있다. 이에 연구 2에서는 연구 1의 연구참가자를 대상으로 상담 경험과 변화에 대한 심층면담(in-depth interview)을 실시하고, 연구주제에 대한 탐색적 이해에 적합한 귀납적 분석(Patton, 1990) 방식으로 면담자료를 분석하였다. 이를 통해 자립준비청년은 상담 과정에서 어떠한 경험을 하였으며, 어떤 인지, 정서, 행동상의 변화 과정을 통해 상담의 효과가 발생하였는지 구체적으로 이해할 수 있을 것이다.

방 법

연구참가자

연구 1에 참여하였던 자립준비청년 31명을 대상으로 연구참가자를 모집하였다. 연구목적, 연구참가자 조건, 연구 목적 및 절차, 사례에 대해 안내문을 개별 발송하였다. 연구 참여 조건은 '자립준비청년 심리지원 사업 참가, 상담자와 내담자의 합의 종결한 상태'이었다. 연구 참가 의사를 밝힌 9명 중 최종 5명을 대상으로 연구가 진행되었다. 연구참가자는 남자 3명, 여자 2명이었고 학력은 대학 재학 4

표 6. 질적연구 참가자 특성

구분	성별	연령	학력	지역	응답한 문제 영역	상담기간/회수
A	여	21	대학 재학	경기	무응답	11주 10회기
B	남	24	대졸 이상	경기	가족, 학업・진로・일, 경제	13주 10회기
C	남	21	대학 재학	경기	학업・진로・일, 경제	17주 10회기
D	여	20	대학 재학	서울	건강, 대인관계, 성격	21주 5회기
E	남	21	대학 재학	부산	이성・성	9주 10회기

명, 대학 졸업 이상 1명이었다. 이들은 9주에서 21주 기간 동안 평균 9회(SD=2.24)의 상담을 받았다. 연구참가동의 후 면담이 진행되었다. 상담 종결 후 2-3개월 경과 시점에 면담이 진행되었다. 연구참가자에 대한 세부정보는 표 6에 제시하였다.

연구면담자

면담의 일관성과 전문성을 위하여 1명의 연구자가 면담을 진행하였다. 면담자는 심리학박사이며 상담심리사 1급 자격을 보유하였다. 면담자는 청소년 정신건강 및 자립 관련 연구 경험이 있으며, 자립준비청년의 실태에 관한 문헌연구 및 상담을 통해 연구참가자에 대한 민감성을 높였다.

연구절차

자료 수집 기간은 코로나19로 인한 강력한 방역 정책이 시행되고 있었기 때문에, 안전을 위해 화상 면담을 진행하였다. 면담의 일관성을 위하여 사전 준비된 질문을 사용하여 반구조화 인터뷰를 하였다. “상담을 받기 전과 후 무엇이 달라졌나요?”, “어느 시점에, 어떤 계기로 변화를 느낄 수 있었나요?”, “변화하는데(또는 변화하지 않은데) 영향을 미친 요인은 무엇인가요?”가 핵심질문이었다. 면담은 최소 34분에서 최대 1시간 44분에 걸쳐 진행되었다. 면담은 녹음 후 축어록으로 기록되었다.

자료분석

반복적 비교분석(유기웅, 정종원, 김영석, 김한별, 2012) 및 구체적인 관찰을 통해 일반적인 패턴을 찾아가는 귀납적 분석(inductive analysis)(Patton, 1990) 방법을 사용하였다. 분석을 위해 연구참가자 5명의 면담 축어록을 반복적으로 읽어 나가며 전체적으로 자료를 파악하였다. 자립준비청년의 상담 경험과 변화에 초점을 두고, 그 경험의 본질을 드러내는 진술문을 추출하였다. 유사한 진술문들을 개념을 가장 잘 드러내는 진술문 하나로 대표하였다. 추출한 진술문들의 공통점과 차이점을 파악하여 진술문들을 배치하고 재배치하는 지속적 비교과정을 거치며 범주화 작업을 하였다. 기본 의미단위인 개념을 구성하는 진술문들을 추출하였다. 개념들을 유사성에 따라 주제로 묶고, 주제들을 묶어 범주를 구성하였다. 분석 과정과 결과에 대한 객관성 및 다각적인 관점을 확보하고자 질적 연구자인 심리학 전

공 박사 1명에게 주제와 범주의 적절성을 중심으로 감수를 받았다.

결 과

본 연구는 자립준비청년의 상담 경험과 그 효과를 이해하는 데 목적이 있다. 연구참여자들의 상담 경험을 분석한 결과, 최종적으로 20개의 개념을 추출하였고, 9개의 주제가 나타났으며, 이를 공감적 관계에서 솔직함, 나를 알아가는 시간, 비합리적 신념의 변화, 긍정성의 증가 4개 범주로 구조화하였다. 자립준비청년이 상담 과정에서 경험한 내용과 상담으로 인한 효과를 분석한 결과는 표 7과 같다.

범주 1. 공감적 관계에서 솔직함

자립준비청년은 대부분 마음 놓고 이야기할 곳이 없어, 힘든 일이 있으면 혼자서 삭히

표 7. 자립준비청년의 상담 경험과 효과

범주	주제	개념
공감적 관계에서 솔직함	나누고 공감받는 관계의 힘	공감받지 못하는 고통에서 자유로워짐
		대화를 나눌 사람이 있다는 편안함
	있는 그대로 나를 드러내는 편안함	혼자 삭히고 묵히기보다 털어놓는 홀가분함
		남이 어떻게 볼지에 대한 두려움의 변화
나를 알아가는 시간	알아차림의 증가	자신의 몸과 감정에 대한 자각 증가
		자기 생각과 행동에 대한 이해 증가
	차분한 성찰의 증가	불안한 반추에서 차분한 성찰로
		무기력한 방치에서 능동적인 대응으로
	정체성의 재정립	자기에 대해 관찰하고 알아가는 과정
		과거를 돌아보며 자신의 행동, 행동 배경 이해
		자기 관찰과 상담사의 피드백에서 정체성 성찰
비합리적 신념의 변화	불완전함에 대한 수용과 이해	결핍이 부끄러워 회피하던 나로부터 변화
		완벽주의와 집착 대신 여유로 편안해짐
	부정적 정체성의 변화	두려워하는 자아상과 현실의 자기를 구분
		완벽하지 않아도 비정상이 아님을 알게 됨
긍정성의 증가	관점의 전환	불행에도 불구하고 잘 살아온 나
		자신의 강점에 주의하고 인정
		자신을 인정하지 않는 무력감에서 안정감으로
	효능감과 주체성의 증가	지지 속에서 문제해결, 효능감과 활력
		눈치 보기보다 내가 좋아하고 원하는 것을 선택

고 해소하지 못하였다. 그러나 상담에서는 터놓고 이야기하고, 공감받고, 마음의 짐을 내려놓는 경험을 하였다. 점차 부끄럽거나 두렵게 느껴져도, 있는 그대로 자신을 드러내고, 이에 따르는 편안함과 안정감을 경험하였다.

나누고 공감받는 관계의 힘.

"가장 힘들었을 때가 언제였는가를 뒤돌아보면 대부분 공감받지 못했을 때 심리적으로 가장 불안했었고, 힘들었고, 뭔가 크게 상처받았다는 생각이 들어요. (중략) 아, 나는 공감을 되게 중요하게 여기는 사람이구나." (내담자 A)

"상담을 곱씹어보면서 옛날과 비교해 지금이 행복하다는 걸 느껴요. 혼자는 해도 쉽지 않잖아요. 그런데 대화를 나눌 수 있는 사람이 있다는 게 속이 편하더라고요." (내담자 E)

있는 그대로 나를 드러내는 편안함.

"예전의 저는 힘든 일이 있으면 혼자서 삭히고 그랬지만 이제는 주변 사람들한테 조금씩 더 얘기하려고 해요. 아무리 부끄러워도 한 번쯤 더 이야기하면서 조금씩 풀어가는 것 같아요. 이야기를 터놓는 것이, 마음이 그렇게 홀가분할 수 있는 거구나를 심어준 계기가 됐어요. 나에게 그리고 누구에게도 솔직한 사람이 되었기 때문에 마음의 짐들을 항상 묵혀 놓고 살지 않는 계기가 되었어요." (내담자 B)

"제가 사람 앞에 나가는 거를 두려워했는데, (상담에서) 얘기가 잘 되어서 남에게 그렇게 보여도 상관이 없다 이렇게 마인드가 바뀌었어요." (내담자 E)

범주 2. 나를 알아가는 시간

자립준비청년들은 스트레스를 받을 때 자신의 신체감각이나 생각을 잘 의식하지 못하고, 신체화, 무기력, 대인기피 등의 증상을 나타냈음을 보고하였다. 그러나 상담에서 보다 안정된 상태가 되어, 문제 상황과 자신의 반응을 주의깊게 관찰하고 표현하였다. 이를 통해 자신의 문제, 스트레스 반응, 부정적 사고, 행동의 이유나 배경 등을 더 잘 알아차리고, 성찰할 수 있게 되었다.

알아차림의 증가.

"스트레스에 무감각한 성격이어서 스트레스 받는 동안은 모르다가 마지막에 스트레스를 풀로 받았을 때 그때야 몸에 이상이 생기는 타입(이거든요). 상담을 받고 나서는 내가 스트레스를 받는 중이구나, 자아성찰 같은 걸 하게 되어요." (내담자 A)

"대인기피증 있었거든요. 왜 내가 (대인기피증이) 있는 줄 몰랐는데, 같이 얘기하다 보니, 옛날엔 좋은 옷 이런 거 없어서 자존감이 낮아졌다고 생각했는데 저가 남들보다 약간 못났다는 (생각해서), 저 자체를 보이는 게 부끄럽다 이런 생각이 들었던 것 같아요." (내담자 E)

차분한 성찰의 증가.

"저를 불안하게 만드는 결정 같은 것들

을 계속 생각하고 되풀이하고, 그게 불안으로 이어지니 일상생활이 잘 안 되더라고요. 어떻게 의논해야 하는질 전혀 몰랐어요. 그러다 보니 내가 했던 행동들도 믿음이 안 가고 앞으로 내가 해야 하는 행동들도 믿음이 안 서는 거예요. 그런데 상담을 하고 나서 제가 했던 행동들에 대해서 그때 왜 그렇게 했는지에 대해서 좀 더 이성적으로 생각할 수 있게 된 것 같아요. 가장 크게 문제 삼았던 건 인간관계였는데 그 문제에 대해서 조금 더 차분하게 대할 수 있게 됐고, 전체적으로 이제 불안했던 생각을 너무 안고 있지 않게 되었어요." (내담자 D)

"무기력한 상황이어서 생각을 정리할 시간이 선생님과 얘기하는 그 시간밖에 없었거든요. 너무 무기력해서 그냥 있었는데 (이제는) 문제도 파악해 보고, 앞으로 어떻게 행동 할 것인가 (생각해요). 그런 자세가 아예 없었던 마음에서 조금 해보자는 생각으로. 내 삶을 더 좋아지게 할 수 있는 것들에 대해서 더 적극적으로 생각하게 변화했어요." (내담자 C)

정체성의 재정립.

"나의 행동에 어떤 배경이 있었는지를 이해할 수 있었고, 내가 평소에 어떤 걸 했는지 다시 생각할 수 있는 그런 시간이었던 것 같아요. 이야기를 들어준 것만으로도 다시 어떻게 살았는지 생각할 수 있게 된 거죠. 내가 뭘 얘기했고 다시 생각이 들고, 저를 발견하는, 저의 다른 부분을 발견하는 부분에서 설레는 게 생겼어요." (내담자 E)

"장단점 이런 걸 정리하고, 그 재료들을 꺼내놓고 말하는 시간이 많다 보니까 왜 이런 현상이 일어났는지 여러 가지 견해를 들어볼 수 있고, 그런 재료를 놓고 선생님의 피드백에서 나온 답변을 하나의 가설과 증거로 삼고? 정체성? 정체성을 좀 더 확인하고 검증하는 시간이 된 것이 컸던 거 같아요." (내담자 C)

범주 3. 비합리적 신념의 변화

연구참가자들은 상담 과정에서 다양한 비합리적 사고, 특히 자신의 정체감과 관련된 비합리적 사고와 두려움을 인식하고 이에 도전하였다. 자신이 결핍되고 비정상적인 존재라는 인식, 이로 인한 수치심과 두려움으로 인해 회피하고, 강박관념에 몰두하느라 상황을 다각적으로 보지 못함을 자각하였다. 상담 과정에서 그러한 사고의 비합리성을 인식하고 대안적 사고를 하였으며, 자신의 불완전함에 대해 수용적 태도가 강화되었다.

불완전함에 대한 수용과 이해.

"남들보다 약간 못났다, 좋은 옷 이런 거 없어서 자존감이 낮아졌다, 밖에 나가면 나를 보이는 게 부끄럽다 이런 생각이 들었던 것 같아요. 근데 상담하고 없어졌어요. 앞으로는 피하거나 부끄러워하지 말자 이렇게 하죠." (내담자 E)

"인간관계는 완벽해야 한다는 생각이 있거든요. 인간관계가 완벽해야 하고 그만큼 다른 일들도 완벽하게 잘해야 한다고

생각을 했거든요. 그런데, 하나에 집착을 버리니까, 나머지도 잘 안 돼도 그럴 수도 있지 하게 됐고 전체적으로 사람들한테도 여유를 많이 가지게 된 것 같아요. 그냥 편해요. 뭘 해야 한다고 막 시달리지도 않고 그냥 덜 괴로워서 좋아요." (내담자 D)

부정적 정체성의 변화.

"진짜 오리지날 고민을 이야기하게 되더라고요. 제가 아빠를 닮아갈까 무서워했던 그 순간에 관해서 이야기했는데, 상담자가'그러면 스스로에 대한 장점을 한 번 이야기해 보시는 게 어때요?'라고 했고, ' 나는 책임감이 강한 사람이구나'를 처음으로 깨닫게 되었어요." (내담자 A)

"아, 나는 이제 비정상이 아니구나. 사람이 아무리 완벽해지려 해도 완벽하지가 않잖아요? 완벽해지려고 왜왜왜를 파 봐도 파고들수록 행복도가 막 낮아지는 느낌? (상담 후에는) 어떤 행동을 하면서, 잘 안 되면 그냥 그럴 수 있구나 하고 그냥 상황을 이해하고 받아들이는 습관을 들이려고 막 노력을 하고 있어요. 장점은 챙기고, 어떻게 더 단점을 보완할까? 생각을 해보게 된 거 같아요." (내담자 C)

범주 4. 긍정성의 증가

연구참가자들은 결핍이 있다는 자기인식을 가졌고, 이 때문에 자신의 실수나 부족한 점에 과민하고, 완벽해지려 과도하게 노력하며, 그렇지 못할 때 스스로 비난하며 무기력해지는 패턴을 나타내었다. 그러나 상담 과정에서, 불행이나 결함 이면에 이를 극복하는 노력과 강점, 걱정과 달리 잘 살아낸 자신의 모습, 완벽하지 않다고 비정상은 아니라는 인식, 실수와 같은 불완전함을 수용하며 대처해갈 수 있다는 인식이 생겨났다고 보고하였다.

관점의 전환.

"제가 참 불행하고 재미없는 삶을 살았다고 이야기했거든요. 그런데 상담 선생님이 그 상황에서 나쁜 범죄 안 하고 비행청소년 안 되고 산 것에 칭찬을 해주셨거든요. 그래서 내가 잘살았구나!" (내담자 E)

"책임감이 강하다는 아주 좋은 말이 있는데도 불구하고 강박관념이라는 생각에 시달렸어요. 이제는 스스로한테 책임감이 강해서 그래(라고 생각해요. 책임감은) 좋은 점이잖아요" (내담자 A)

"제가 느꼈던 불안, 실수, 생각 그리고 그 외의 것들도 생각하게 되고 그걸 가지고 다른 사람도 이해하게 된 것 같아요. 실수나 이런 거를 비판하거나 걱정하는 마음에서 좀 이해하는 마음으로 바뀌었어요." (내담자 D)

효능감과 주체성의 증가.

"상담에서 인생계획을 한 번 짰었는데, 제가 인생계획을 얘기할 때 힘들어하는 모습이 있으면, (상담자가) 같이 옆에서 이런 다른 방법도 있다고 같이 설명해주고. 바로 실행 안 해도 1년 단위씩 계획이라서 부담 없이 할 수 있었어요. 벌써 한 개 해결했거든요. 더 빨리 목표를 이루었을 때

더 일찍 계획을 당겨보자 이런 욕심도 있고, 자신감이 생긴 것 같고 좀 더 열심히 하려고 하는 것 같아요. 앞으로 어떻게 해야 할지 막막했는데, 상담하면서 좀 더 틀을 잡게 됐어요." (내담자 E)

"결핍도 있다 보니, 껍데기를 쫓아가는 것, 스스로 부정하고 있었는데 상담을 통해서 알게 되더라고요. '외적인 시선보다 나를 위해서 살아야겠다' 그런 생각을 할 수 있는 시간이었어요. 나다운 게 뭔가에 대해서 생각할 수 있었던 좋은 시간. 내가 진짜 좋아하는 것을 하는 게 저의 행복, 제 마음의 안정에 훨씬 더 도움이 된다고 생각이 들어요." (내담자 B)

전체 결과를 정리하면 다음과 같다. 연구참가자 상당수는 자신이 다른 사람들에 비해 결핍되었고 비정상의 존재라는 인식을 했다. 이 때문에 자신의 실수나 부족한 점에 과민하고, 남들의 시선에 민감하며, 과도하게 걱정하거나 무기력해지는 패턴을 보고하였다. 이는 우울, 불안과 초조함, 두려움과 강박관념, 무감각, 무기력과 회피, 신체화 등의 심리증상 및 낮은 자존감과 함께 부정적이고 혼란스러운 자기정체감과 연결되었다.

상담이라는 공감적 관계에서 연구참가자들은 터놓을 곳이 있다는 안정감을 가지고, 점차 자신을 있는 그대로 자신을 드러내게 되었다. 점차 외부 시선이나 고정관념에 집착하기보다 자신에 대한 조절된 관찰과 알아차림이 일어났다. 상담자와 주고받는 상호소통과 피드백의 과정에서 자신의 비합리적 신념을 깨닫고 관점을 넓혔다. 자신이 살아온 과정, 자신의 행동 배경과 이유, 생각과 감정, 장단점 등에 관해 이야기하며 자신에 대한 이해를 넓혀갔다. 자신이 겪은 불행이나 결함 이면에 이를 극복하는 노력을 가치 있게 생각하게 되고, 불완전함을 수용하였다. 자신의 욕구를 알아차리고 존중하는 마음이 생겨남에 따라 주체성을 강화되었다. 문제해결자로서 자기, 강점에 대한 인식도 증가하였다. 이러한 인식들은 긍정적 자아정체감으로 통합되고, 자기 존중감을 높이는 역할을 하였다.

종합논의

본 연구는 자립준비청년 대상 상담이 실효성이 있는지, 이러한 효과는 어떠한 경험으로 인한 것인지를 밝히고자 하였다. 연구 1에서는 (사)한국상담심리학회 상담심리사 1급으로부터 평균 9.9회 상담을 받은 31명의 자립준비청년의 상담 전후 변화를 비교하였다. 연구참가자들은 상담에 대해 높은 만족도를 보고하였고, 상담 전후 삶의 주요 영역에서 호소문제가 유의하게 감소하였음을 보고하였다. 상담 이후, 우울과 불안을 비롯한 심리증상이 유의하게 감소하고, 자아존중감과 자아정체감의 자기개념 역시 유의한 개선을 나타내었다. 이 중 5명의 자립준비청년을 대상으로 실시된 질적연구는 연구 1에서 검증된 심리증상 완화 및 자기개념 변화의 의미를 구체화하였다. 연구1과 연구2의 결과를 종합하여 살펴보면 다음과 같다.

첫째 상담은 심리증상을 개선한다. 연구 1에서 상담 전과 비교할 때 상담 후, 심리증상은 큰 효과크기로 개선되었다. 상담전후 결과를 통제집단과 비교할 수 없다는 본 연구의 한계를 보완하는 방안으로, 상담 후 변화된

심리증상 수준을 같은 척도를 사용한 선행연구 결과와 비교하였다. 불안과 우울을 비롯한 심리증상의 경우 본 연구에서 상담 후 1.93점으로 감소하였는데, 이는 같은 척도를 사용한 심리문제가 없는 일반 대학생 집단의 평균점수 1.68-2.05점(김은정, 김진숙, 2020; 봉은주, 하윤주, 2013)와 유사하다. 즉, 상담의 효과로 심리증상으로 인한 고통과 기능 저하가 개선되었다 할 수 있다. 또한, 연구 1의 양적연구 결과와 일치하게, 연구 2의 질적연구 결과도 상담이 우울, 불안과 초조함, 두려움과 강박관념, 무감각, 무기력과 회피, 신체화 등 다양한 심리증상을 개선함을 재확인하였다. 이러한 심리증상의 감소는 성공적 자립을 저해하는 위험요인이 감소하였음을 의미하며, 상담의 효과성을 확인하는 일반적 지표이다(김계현, 2002; 이정애, 2018).

둘째, 상담은 자립준비청년의 심리적 고통을 완화할 뿐 아니라, 스트레스에 기능적으로 대처하고 건강한 성인기의 기반이 되는 자아기능을 강화함을 확인하였다. 연구참가자의 자아존중감과 자아정체감의 자기개념은 상담 전후 모두 큰 효과 크기로 변화하였다. 자아존중감은 보호시설 퇴소 후, 발달과 적응에 긍정적인 영향을 미치고(김예성, 이경상, 2015) 자아정체감 역시 대인관계에서의 안정감, 자아존중감, 스트레스 저항력과 관련된다(Marcia, 1980). 자아정체감과 자아존중감의 발달은 자립 과정에서의 적응력과 깊이 관련되며, 특히 청소년기에는 성격발달의 기초이며, 삶의 방향을 결정하는 중요한 과제이다(박아청, 2002; 원재순, 2018; 최정호, 한영주, 2020). 연구 2는 상담으로 인한 연구참여자들의 자기개념 변화 과정을 구체화하였다. 연구참가자들은 상담자와 상호작용 속에서 자신의 경험을 탐색하고, 의식적으로 자각하고, 재해석하였다. 자기와 관련한 낙인과 비합리적 사고 및 그 영향을 더욱 명료히 알아차리며, 이를 현실적인 자료를 바탕으로 수정해 나갔다. 자신이 비정상이라는 비합리적 신념으로 빠져들며 무기력감으로 현실을 회피하는 대신, 불완전함과 고통을 수용하였다. 즉, 불완전한 자기와 함께 긍정적 자기의 측면도 자각하며 자기의 요소들을 보다 유연하게 통합해 나갔다. 이러한 자기정체성의 재인식 과정은 자기에 대한 인정과 존중을 강화하였다. 상담자의 존중적 태도와 긍정적 반응 역시 내담자의 자존감을 자극하고 강화하는 것으로 보인다. 더불어 새로운 발견과 성장의 느낌은 활력과 긍정성을 초래하고 이 역시 자존감을 높이는 역할을 하였다.

셋째, 본 연구에서는 상담 효과를 가져온 상담 요인들을 탐색할 수 있었다. 본 연구에서 나타난 상담의 효과에는 내담자 요인과 상담자 요인이 모두 작용한 것으로 보인다. 상담을 신청한 자립준비청년들은 문제의식을 느끼며 스스로 도움을 요청하는 자발성(Dean, 1958)을 가졌다. 그런데도 연구 2의 분석결과를 살펴보면, 상담에 대한 긍정적인 기대와 변화에 대한 희망(Frank & Frank, 1991)은 상담자와의 상호작용을 통해 유지되고 강화된 것으로 보인다. 자립준비청년은 상담 초기에는 자기를 숨기거나 상담효과에 대한 의구심을 가진 경우가 많았지만, 상담자의 공감적이면서도 전문적 개입에 신뢰감을 형성한 것으로 보인다. 상담에서 자립준비청년의 변화를 촉진하는 핵심적인 요인은 상담자와의 관계였다. 본 연구에서 상담이라는 공감적 관계가 주어지자 연구참가자들은 안정감을 경험하게 되고, 점차로 자신을 숨기거나 꾸미기보다 있는 그대로 표현할 수 있었다. 이러한 개방성

은 자기에 대한 조절된 관찰과 알아차림으로 이어지고 자기인식을 확장할 수 있게 하였다. 상담자의 경청, 전문적인 질문과 피드백은 내담자들이 자신의 경험에 집중함으로서, 비합리적 사고를 다룰 뿐 아니라 주의하지 않았던 자신의 강점과 긍정적인 정체성을 발견하고 통합하도록 도왔다. 또한, 삶의 어려운 문제들은 새로운 방식으로 접근하며 단계적으로 풀어가는 방법을 익혀가도록 도왔다. 지지와 공감적 태도로 대안적 관점과 행동을 촉진하는 개입, 알아차림과 자기수용을 촉진하는 개입, 문제와 패턴 이해를 위한 개입 등 효과적인 상담자 요인(김은하, 김현준, 김이윤, 김주영, 2018)을 본 연구결과에서 확인할 수 있었다. 교류 대상이 없고 정서적 지지를 받지 못하여 자신감이 저하되고 불안과 두려움을 경험하는 이에게 상담자는 강력한 보호 요인이다(김미연, 2010). 홀로 세상에 적응해가는 자립준비청년에게 경제적 지원만큼이나 애착을 형성하고 지지를 제공하는 심리적, 사회적 안전망이 필요할 것이다. 본 연구는 상담자가 자립준비청년에게 이러한 사회적 안전망의 하나로 작동하는 동시에, 방치되었던 외상의 치유와 성격의 발달을 촉진함을 확인하였다.

상담의 실효성을 확인한 본 연구결과는 상담 역량이 검증된 전문인력의 개입이 전제되었다는 점을 다시 한번 강조할 필요가 있다. 내담자의 상태를 정확히 평가하고 긍정적인 치료 결과를 가져올 수 있는 임상적 전문성(APA, 2006)은 상담 효과를 결정하는 핵심 조건이다(홍은택, 김현진, 박수현, 최기홍, 2023). 이는 국가 심리서비스에 전문인력이 투입되어야 함을 강력히 시사한다. 그러나 국내 심리서비스 관련 법과 제도를 살펴보면 상담을 비롯한 심리서비스의 정의와 전문성을 규정하는 기본 모법이 부재하고, 이로 인해 정책중복, 사각지대, 비전문성으로 서비스의 비효율성이 존재한다. 또한, 국내 상담 수요가 폭발적으로 증가하며 비전문가의 서비스 제공 역시 기하급수적으로 급증하고(상담인적자원개발위원회, 2019), 이에 따른 피해 역시 속출하고 있다(강창욱, 이동환, 정진영, 박장군, 2022). 전문성이 결여된 상담은 단순히 효과 없음을 넘어, 내담자의 문제를 악화시키고 또 다른 심리적 외상을 만들어낼 수 있다. 심리서비스가 급증하는 현 상황에서, 실효성 있는 서비스가 가능하도록 상담 인력의 전문성 기준과 관련 법령의 정비가 필요하다(원성두, 장은진, 2022; 홍은택, 박수현, 최기홍, 2023).

본 연구의 의의와 한계를 정리하면 다음과 같다. 증거기반 개입이 강조됨에 따라 상담 개입의 효과를 검증하는 연구들이 다양하게 이루어지고 있다. 그러나 국내 연구들은 집단상담 프로그램을 개발하고 그 효과를 입증하려는 시도가 다수이며, 개인상담의 효과 검증 연구는 드물다. 표준화된 처치의 어려움, 상담 효과 측정의 합의 결여, 충분한 자료 수집의 어려움이 그 현실적인 이유이다(김동민, 2021). 이러한 맥락에서 본 연구는 엄격한 상담자 자격 기준을 설정하여 상담 처치의 타당성을 확보하고, 타당화된 측정도구를 사용하고, 자립준비청년이라는 동질성을 기반으로 통계적 검증이 가능한 수의 연구참가자를 확보하고, 양적 연구에서 검증된 상담 효과를 질적 연구를 통해 그 임상적 함의를 구체화했다는 점에서 근거기반 개입을 위한 상담 연구로서 의의가 있을 것이다.

그러나 본 연구는 상담 효과 검증 연구로서 설계상 한계 역시 가진다. 첫째, 본 연구는 단일집단 사전사후 설계로 설계되었다. 상담

전후 변화의 효과크기가 크고, 질적연구 결과가 양적연구 결과와 부합하는 등의 요소를 고려할 때 연구의 내적타당도를 확보하였다 볼 수 있으나, 그럼에도 통제집단이 설계되지 않았기 때문에 성숙요인이나 검사요인 등이 작용하였을 가능성을 완전히 배제할 수 없다. 또한, 연구 기간 내 발생가능한 외생변수를 영향 역시 완전히 배제하지 못한다. 추후 연구에서는 상담 미신청 또는 대기 조건의 통제집단을 계획함으로써 연구의 내적타당도를 높이는 노력이 필요할 것이다.

둘째, 본 연구는 구체적으로 어떠한 상담 개입이 이루어졌는지 확인할 수 있는 상담자 요인이 변인으로 포함되지 않았다. 질적연구를 통해 상담 효과에 상담자 요인이 작용하였음을 간접적으로 확인할 수 있지만, 이를 직접 검증한 것은 아니다. 이에 상담 개입의 구체적인 특성이나 상담 효과가 나타난 정확한 기전을 파악하기 어렵다. 향후 연구에서는 구체적으로 어떠한 상담 이론이나 기법의 처치가 적용되었는지, 또는 처치 외 성과에 영향을 미치는 요인이 무엇인지(김동민, 2021)를 세심하게 측정할 필요가 있을 것이다.

셋째, 본 연구는 상담이 비교적 성공적으로 진행되고 합의 종결한 내담자의 경험을 담고 있다. 본 연구에는 상담의 취소, 중도탈락 등으로 합의 종결하지 않은 사례들이 포함되지 않았기 때문에, 탈락률이나 악화율과 같은 성과지표(Castonguay et al., 2013)를 확인할 수 없다. 자립준비청년은 보호종료 5년 이내 20.2%가 연락이 끊어지고 사후관리에서 누락되는 특성이 있다(아동권리보장원, 2022). 즉 다양한 자립지원이 마련되어도, 일부 자립준비청년에게는 여전히 지원을 활용할 접근성, 동기, 생활기술이 부족할 수 있다. 또는 상담에 편견을 갖거나 그 유용성에 대한 기대가 낮아, 상담을 받지 않을 수 있다. 이에 상담을 원하지 않거나 중단한 사례, 효과가 없거나 역기능적으로 작용한 사례들에 대한 파악과 연구가 필요하다. 이러한 이해는 자립준비청년이 상담을 이용하지 않는 상담서비스갭을 줄이고, 상담 효과를 높이는 데 도움이 될 것이다.

마지막으로 보호아동과 자립준비청년의 건강한 발달을 위하여, 상담 등 심리서비스가 전문화, 다각화, 다양화될 필요성을 다시 한번 강조하고자 한다. 자립준비청년은 부모와 분리 외에도 다양한 심리적 외상에 노출되었을 위험이 있다(황수연, 2018). 아동 및 청소년기의 심리적 외상은 심리증상뿐만 아니라 성격에도 부정적인 변화를 일으킨다(김은희, 이인혜, 2016; Herman, 1992). 성격은 오랜 시간에 걸쳐 형성되는 것처럼, 변화 역시 오랜 시간이 걸릴 수 있다. 자립준비청년의 호소문제와 상담자의 전문적 평가를 바탕으로 장기 상담 등 다양한 형태의 지원책이 마련될 필요가 있다. 또한, 자립준비청년의 어려움은 단순히 퇴소 전・후에 발생하는 것이 아니라 의지할 곳을 찾지 못한 총체적 문제가 누적된 것이다. 따라서 예방적 관점에서 보호 과정 전 단계에서 시의적절하게 전문적 상담 지원이 이루어질 필요가 있다.

참고문헌

강유임, 서선아 (2023). 청소년을 위한 언택트 스트레스 관리 집단상담 프로그램 효과성 연구. 청소년상담연구, 31(1), 183-211. http://dx.doi.org/10.35151/kyci.2023.31.1.009

강창욱, 이동환, 정진영, 박장군 (2022. 5. 23.)

"무조건 합격이세요" 엉터리 심리상담사, 기자도 땄다[이슈&탐사]. 국민일보.
https://www.kmib.co.kr/article/view.asp?arcid=0017103881

강현아 (2010). 시설 퇴소청소년의 레질리언스에 영향을 미치는 요인. 청소년학연구, 17(2), 155-179.
https://kiss.kstudy.com/Detail/Ar?key=2825762

고수안 (2023). 자립준비청년의 시설 퇴소 후 자립 과정에서의 어려움에 대한 질적연구: 정책 수혜자의 입장에서. 서울대학교 대학원 석사학위논문.
https://www.riss.kr/link?id=T16865579

금명자, 정상화 (2021). 국내 상담자 발달 연구 동향. 교육문화연구, 27(2), 483-508.
https://doi.org/10.24159/JOEC.2021.27.2.483

김계현 (2002). 교육상담에서의 효과성 연구와 메타분석. 아시아교육연구.
https://hdl.handle.net/10371/88945

김동민 (2021). 무엇이 상담효과를 산출하는가?: 상담의 효과성에 관한 경험적 증거의 함의. 상담학연구, 22(5), 59-75.
https://doi.org/10.15703/kjc.22.5.202110.59

김명찬, 이현진 (2016). 정서중심 집단상담 프로그램이 청소년의 감정표현불능증, 우울, 신체화 증상 및 대인관계에 미치는 영향. 상담학연구, 17(6), 223-239.
http://dx.doi.org/10.15703/kjc.17.6.201612.223

김미연 (2010). 그룹홈 청소년의 자립의지에 영향을 미치는 요인에 관한 연구. 청소년문화포럼, 24(1), 7-38.
https://kiss.kstudy.com/DetailOa/Ar?key=51097628

김예성, 이경상 (2015). 시설청소년의 자립준비 정도에 영향을 미치는 요인에 관한 연구. 청소년문화포럼, 42, 7-32.
https://doi.org/10.17854/ffyc.2015.04.42.7

김은정, 김진숙 (2020). 아동기 정서적 외상경험과 성인기 정신병리 및 대인관계 문제의 관계에서 정서조절곤란의 매개효과: 정서적 학대와 정서적 방임의 차별적 경로를 중심으로. 상담학연구, 21(3), 23-44.
https://doi.org/10.15703/kjc.21.3.202006.23

김은희, 이인혜 (2016). 아동 청소년기 외상경험과 경계선 성격 특성의 관계: 대상적 자기 손상의 매개효과. 한국심리학회지: 상담 및 심리치료, 28(4), 1003-1022.
https://accesson.kr/kcpa/assets/pdf/20892/journal-28-4-1003.pdf

김은하, 김현준, 김이윤, 김주영 (2018). 내담자가 인식하는 효과적인 상담자의 행동, 반응, 태도에 관한 개념도 연구: 대학상담센터를 중심으로. 상담학연구, 19(1), 1-21.
https://doi.org/10.15703/kjc.19.1.201802.1

김인규 (2021). 한국 상담이 나아갈 방향. 상담학연구, 22(4), 29-38.
https://doi.org/10.15703/kjc.22.4.202108.29

박기쁨, 우상우, 장문선 (2012). 대학생 집단을 통한 단축형 간이정신진단검사-18(BSI-18)의 타당화 연구. *Korean Journal of Clinical Psychology, 31*(2), 507-521.
https://doi.org/10.15842/kjcp.2012.31.2.006

박신애, 최옥채 (2018). 아동양육시설 퇴소 청소년의 외상후성장 경험. 사회과학연구, 34(2), 127-153.
https://doi.org/10.18859/ssrr.2018.05.34.2.127

박아청 (2002). 정체감 교섭 과정과 정신적 건강과의 관련에 관한 연구. 교육심리연구, 16(4), 207-228.
https://www.dbpia.co.kr/journal/articleDetail?nod

eId=NODE06763365
보건복지부 (2021). 자립준비청년 자립의 길, 따뜻한 포용정책으로 동행 [보도자료]. https://www.mohw.go.kr/board.es?mid=a10503010100&bid=0027&act=view&list_no=366425&tag=&nPage=207
보건복지부 (2022a). 부모의 마음으로, 따뜻하게 동행하겠습니다 [보도자료]. https://www.korea.kr/docViewer/skin/doc.html?fn=c09a05f59ae985449289 98d566b865b0&rs=/docViewer/result/2023.02/13/c09a05f59ae98544928998d566b865b0
보건복지부 (2022b). 2022년 청년마음건강지원사업 실시 [보도자료]. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=1&CONT_SEQ=371023
봉은주, 하윤주 (2013). 대학생의 인터넷중독과 성인아이 성향, 정신건강과의 관계. 한국산학기술학회 논문지, 14(10), 5037-5047. https://doi.org/10.5762/KAIS.2013.14.10.5037
손난희, 유성경 (2012). 상담성과 측정을 위한 상담성과 척도(Outcome Questionnaire-30: OQ-30)의 타당화. 상담학연구, 13(1), 1-15. https://doi.org/10.15703/kjc.13.1.201202.1
송현옥 (2008). 청소년기의 자아정체감에 영향을 미치는 관련변인 간의 구조분석. 계명대학교대학원 박사학위논문. https://www.riss.kr/link?id=T11598966
아동권리보장원 (2022). 2021년 아동자립지원 통계현황보고서. 서울: 아동권리보장원. https://www.ncrc.or.kr/ncrc/na/ntt/selectNttInfo.do?mi=1469&bbsId=1127&nttSn=4135&cataGori=all&tabName=all
안성희, 성현모, 김보람, 이상민 (2022). 실무실습, 실무수련, 실무교육, 실무경력: 심리상담사의 실무능력배양의 방향성. 상담학연구, 23(3), 39-49. https://doi.org/10.15703/kjc.23.3.202206.39
이정우, 이소연 (2023). 자립준비청년 내담자의 심리상담 경험에 대한 내러티브 탐구. 한국청소년연구, 34(3), 129-158. https://doi.org/10.14816/sky.2023.34.3.129
양은별, 김태우, 박은혜, 이소연, 정익중 (2015). 청소년의 학교적응에 영향을 미치는 요인: 일반청소년 및 가정외 보호 청소년 비교를 중심으로. 학교사회복지, 31, 311-331. https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002020265
왕은자, 유정이, 김선경 (2018). 국내 개인상담 성과의 측정 및 평가에 대한 분석: 학술지 <상담학연구>와 <한국심리학회지: 상담 및 심리치료>를 중심으로. 상담학연구, 19(5), 1-23. https://doi.org/10.15703/kjc.19.5.201810.1
원성두, 장은진 (2022). 대한민국 심리서비스 관련 법령 및 적용 현황. 한국심리학회지: 일반, 41(3), 257-270. https://doi.org/10.22257/KJP.2022.8.41.3.257
원재순 (2018). 청소년 집단상담 프로그램 효과에 대한 메타분석: 국내 프로그램 개발 논문 중심으로. 경북대학교 대학원 박사학위논문. https://www.riss.kr/link?id=T14916043
유기웅, 정종원, 김영석, 김한별 (2012). 질적연구방법의 이해. 서울: 박영사. https://doi.org/10.979.11303/50899
윤희섭, 정현희 (2010). 상담일반: 상담자의 애착유형과 발달수준에 따른 역전이 관리

능력의 차이. 상담학연구, 11(2), 485-505. https://doi.org/10.15703/kjc.11.2.201006.485

이미혜 (2002). 시설보호형태에 따른 보호청소년의 자아개념 비교 연구. 서울대학교 대학원 석사학위논문. https://s-space.snu.ac.kr/handle/10371/38307

이상정, 김지민, 안은미, 김무현 (2020). 보호종료(예정)아동 심리정서 실태조사. https://www.ncrc.or.kr/ncrc/na/ntt/selectNttInfo.do?mi=1176&bbsId=1048&nttSn=2953&cataGori=all&tabName=

이상정, 류정희, 김지연, 김무현, 김지민 (2019). 가정 외 보호 아동의 자립 준비 실태와 자립 지원 체계 개선 방안 연구. https://repository.kihasa.re.kr/bitstream/201002/34541/1/%EC%97%B0%EA%B5%AC%EB%B3%B4%EA%B3%A0%EC%84%9C%202019-22.pdf

이상정, 김지민, 류정희, 조정우, 홍문기, 안은미 (2021). 자립준비청년 지원 강화를 위한 보호서비스 전달체계 개선 연구. repository.kihasa.re.kr/en/bitstream/201002/39742/2/연구보고서%202021-30.pdf

이정애 (2018). 가정외보호 퇴소청소년의 자립에 관한 혼합연구. 이화여자대학교 사회복지학과 박사학위논문. https://www.riss.kr/link?id=T14886248

이정애, 이화조, 정익중 (2017). 가정외보호 퇴소아동의 자존감과 사회적 지지가 자립생활기술에 미치는 영향: 공동체의식의 매개효과. 지역과세계(구사회과학연구), 41(1), 181-207. https://doi.org/10.33071/ssricb.41.1.201704.181

이종원, 황진구, 모상현, 정은주, 강현철, 한영근, 허효주, 문은옥, 이영화 (2014). 한국아동 . 청소년패널조사Ⅴ: 사업보고서. 한국청소년정책연구원연구보고서, 1-208. https://nypi.re.kr/brdrr/boardrrView.do?menu_nix=4o9771b7&brd_id=BDIDX_PJk7xvf7L096m1g7Phd3YC&cont_idx=483&seltab_idx=0&edomweivgp=R

이태연, 최은숙, 이세정 (2019). 아동양육시설 퇴소 후 청소년들의 생활경험과 자립간의 관계에 대한 사례 연구. 청소년학연구, 26(4), 293-322. https://doi.org/10.21509/KJYS.2019.04.26.4.293

이훈진, 원호택 (1995). 자기개념과 편집증적 경향. 심리과학, 4, 15-29. https://scholar.kyobobook.co.kr/article/detail/4010009105071

임민경, 이지혜, 이한나, 김태동, 최기홍 (2013). 근거기반 실천과 심리치료. 한국심리학회지: 일반, 32(1), 251-270. https://accesson.kr/kpageneral/assets/pdf/15871/journal-32-1-251.pdf

장윤정 (2013). 아동양육시설 청소년의 자아존중감, 사회적지지, 우울, 공격이 자립에 미치는 영향. 명지대학교 대학원 석사학위논문. https://www.riss.kr/link?id=T13093505

장정은, 전종설 (2018). 양육시설 퇴소청소년의 초기 자립 경험. **청소년복지연구**, 20(2), 95-125. https://doi.org/10.19034/KAYW.2018.20.2.05

장혜림, 정익중 (2017). 가정외보호 퇴소 대학생의 생활 경험. **청소년복지연구**, 19(2), 47-80. https://doi.org/10.19034/KAYW.2017.19.2.03

전용오 (2000). 대학상담에서 상담자-내담자 동맹관계와 상담성과 간의 연계적 관계.

서울대학교 박사학위논문.
https://www.riss.kr/link?id=T7863259

정선욱 (2015). 대학에 진학한 시설 퇴소 청년의 진로준비행동 영향요인 - 사회적 지지와 자아정체감을 중심으로. 사회복지연구, 46(1), 191-214.
https://doi.org/10.16999/kasws.2015.46.1.191

정익중 (2007). 미국 요보호아동의 퇴소 후 자립 관련 프로그램과 시사점. 사회과학연구, 13, 35-52.
https://kiss.kstudy.com/Detail/Ar?key=2672313

정익중, 김주현 (2019). 가정위탁종결청소년의 자립 경험. 한국가족복지학, 64, 131-163.
https://doi.org/10.16975/kjfsw.2019..64.005

제철웅, 장영인 (2019). 성인기 전이과정에 있는 보호 대상 청소년 지원 방안-특별법제정의 필요성을 중심으로. 서울법학, 27(1), 35-77.
https://doi.org/10.15821/slr.2019.27.1.002

조한익, 김영숙 (2016). 청소년의 미래지향목표와 자아정체감, 공동체의식 및 진로정체감의 종단적 구조관계. 교육심리연구, 30(4), 783-810.
https://doi.org/10.17286/KJEP.2016.30.4.06

최미혜 (2015). 부모의 헬리콥터형 양육태도가 대학생 자녀의 친구관계, 공동체의식에 미치는 영향 연구: 자아정체감의 매개효과를 중심으로: 자아정체감의 매개효과를 중심으로. GRI 연구논총, 17(2), 181-205.
https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002021471

최정호, 한영주 (2020). 십오통활집단상담에 참여한 중학생들의 체험 연구: 해석학적 현상학 방법을 적용하여. 질적탐구, 6(2), 131-166.
http://dx.doi.org/10.30940/JQI.2020.6.2.131

통계청. KOSIS 국가통계포털. 보호대상아동 현황보고.
https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=TX_117341138&conn_path=I3. 에서 2020.12.06.인출

홍예영, 김유숙 (2020). 아동양육시설 퇴소를 앞둔 청소년의 심리적 경험에 대한 질적 연구. 청소년학연구, 27(2), 275-304.
https://doi.org/10.21509/KJYS.2020.02.27.02.275

홍은택, 김현진, 박수현, 최기홍 (2023). 국내 심리서비스 관련 민간자격 현황 연구. 한국심리학회지: 임상심리 연구와 실제, 9(3), 481-505.
https://doi.org/10.15842/CPKJOURNAL.PUB.9.3.481

홍은택, 박수현, 최기홍 (2023). 국가 차원의 근거-기반 심리서비스 제도의 필요성. 한국심리학회지: 일반, 42(4), 287-307.
https://doi.org/10.22257/kjp.2023.12.42.4.287

황수연 (2018). 아동양육시설 퇴소성인들의 가족생활 어려움과 극복 경험 연구. 한국사회복지학, 70(1), 33-61.
https://doi.org/10.20970/kasw.2018.70.1.002

Arnett, J. J. (Ed.). (2015). *The Oxford handbook of emerging adulthood*. Oxford University Press.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *The American Psychologist, 61*(4), 271-285.
https://doi.org/10.1037/0003-066X.61.4.271

Barkham, M., Hardy, G. E., & Mellor-Clark, J. (Eds.). (2010). *Developing and delivering practice-based evidence:A guide for the psychological therapies*. John Wiley &Sons.

Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J (2001). Pushing the quality envelope: A new outcomes management system. *Psychiatric Services, 53*, 925-934. https://doi.org/10.1176/appi.ps.52.7.925

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice-oriented research: Approaches and application. In M. J. Lambert (Ed.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp.85-133). New York, NY: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd ed*. Hillsdale, NJ: Erlbaum. http://dx.doi.org/10.1080/00401706.1989.10488618

Coopersmith, S. (1981). *Self-esteem inventories*. Consulting Psychologists Press.

Cunningham, M. J., & Diversi, M. (2013). Aging out: Youths' perspectives on foster care and the transition to independence. *Qualitative Social Work*, *12*(5), 587-602. https://doi.org/10.1177/1473325012445833

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York: Marcel A.

Dean, S. I. (1958). Treatment of the reluctant client. *American Psychologist, 13*(11), 627-63.

Derogatis, L. R. (2001). *BSI 18, Brief Symptom Inventory 18: Administration, scoring and procedures manual*. NCS Pearson, Incorporated.

Erikson, E. H. (1968). *Identity: Youth and crisis (No. 7)*. New York, W. W. Norton & company.

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing* (3rd ed.). Baltimore: Johns Hopkins Press.

Herman, J. L (1992). Complex PTSD: A syndrome in survivors of prolonged and repeated trauma. *Journal of Traumatic Stress, 5*, 377-391. https://doi.org/10.1002/jts.2490050305

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research, 17*, 1-14. https://doi.org/10.1080/10503300601032506

Lambert, M. J., & Vermeersch, D. A. (2008). Measuring and improving psychotherapy outcome in routine practice. *Handbook of counseling psychology*, *4*, 233-248.

Marcia, J. E. (1980). Identity in adolescence. *Handbook of adolescent psychology*, *9*(11), 159-187. https://www.researchgate.net/publication/233896997_Identity_in_adolescence

OECD (2015). 정신보건의료의 중요성에 대한 인식증대: 정신보건의료 문제를 방치한 결과 발생하는 사회경제적 비용. OECD 대한민국 정책센터.

Patton, M. Q. (1990). Qualitative evaluation and research methods(2nd ed.). Newbury Park, CA: Sage.

Reilly, T. (2003). Transition from care: Status and outcomes of youth who age out of foster care. *Child welfare*, 727-746. https://thomreillypublications.com/docs/2003_Transition_From_Care.pdf

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). Acceptance and commitment therapy.

Measures package, 61(52), 18. https://integrativehealthpartners.org/downloads/ACTmeasures.pdf#page=61

Wampold, B. E. (2007). Psychotherapy: the humanistic (and effective) treatment. *American Psychologist*, *62*(8), 857. https://doi.org/10.1037/0003-066X.62.8.857

1차원고접수 : 2023. 11. 28
2차원고접수 : 2024. 05. 21
최종게재결정 : 2024. 07. 22

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 173-197
http://dx.doi.org/10.22257/kjp.2024.9.43.3.173

Mixed Methods Research on the Effectiveness of Counseling on Youth who are in Readiness for Self-Reliance*

HaeYoun Choi **JeeSung Baek**

Professor, Department of Psychology, Chungbuk National University,

Given the need for empirical evidence to inform evidence-based interventions and systematic policy development for the mental health of Youth who are in Readiness for Self-Reliance(YRSR), this study examined the need for and effectiveness of professional counseling for YRSR. Study 1 examined the changes in 31 young adults (mean age 21.81 years, 18 females) before and after receiving a mean of 9.90 (SD=2.13) counseling sessions from a counselor with a first-level license as a counseling psychologist. After counseling, there was a significant decrease in psychological symptoms and a significant increase in self-identity and self-esteem. They reported significantly lower levels of major complaints and higher levels of counseling satisfaction. Study 2 explored the process of counseling effectiveness through a qualitative study of the counseling experiences of YRSR. We analyzed the counseling experiences of five young adults who participated in Study 1 and organized the 20 concepts into four categories: empathic honesty, time to get to know me, freedom from irrational beliefs, and increased positivity. An empathic counseling relationship allows for emotional stability and self-expression as it is, which in turn promotes mindfulness, conscious reflection, acceptance, and integration, leading to the reduction of psychological symptoms and the strengthening of an empowered self-concept. The need to secure the professionalism of counseling personnel was discussed as a key factor in the effectiveness of counseling.

Key words : Youth in readiness for self-reliance, Counseling effectiveness, Counseling Professional Qualification Standard, Therapeutic factors.

---

* This work was supported by the research grant of the Chungbuk National University in 2022

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 199-230
http://dx.doi.org/10.22257/kjp.2024.9.43.3.199

구조방정식 모형의 전반적인 평가 및 효과크기와 연속성에 대한 숙고

유 소 현　　김 수 영†

이화여자대학교 심리학과

잠재변수 간 관련성을 설명하는데 활발히 사용되고 있는 구조방정식 모형은 적합도를 통하여 그 유용성을 판단할 수 있다. 모형 적합도의 통계적 유의성을 평가하는 χ^2검정과 실질적 유의성을 평가하는 적합도 효과크기 지수는 각각 이분법적 해석과 연속적 해석 방식을 이용하여 모형의 유용성을 평가한다. 실질적 유의성의 경우 적합도의 수준을 연속적으로 평가하는 것이 목적이지만, 이를 위해 사용되는 효과크기 지수는 현재 대다수의 연구에서 이분법적으로 해석되고 있다. 본 연구는 모형 적합도의 실질적 유의성을 평가하기 위하여 효과크기 지수의 관점에서 적합도의 수준을 연속적으로 해석하는 방법과 올바른 가이드라인의 사용에 대해 다룬다. 먼저 χ^2검정을 이용한 통계적 유의성 평가의 의의에 대해 간단히 설명한 뒤, 구조방정식 모형의 맥락에서 효과크기의 개념을 정의한다. 이후 다양한 종류의 적합도 효과크기 지수를 소개하며, 해당 지수들의 해석에 사용되는 가이드라인의 특징에 대해 설명한다. 마지막으로, 추정된 효과크기 지수 값을 연속적으로 해석하고자 할 때 참고하기에 적절한 가이드라인의 올바른 예시를 제공하며, 연속성이 반영되지 않았을 때의 잘못된 모형 평가 사례와 가이드라인을 근소하게 만족하지 못하는 모형에 대한 올바른 해석 방식에 대해 논의한다.

주요어 : 구조방정식 모형, 모형 적합도, 효과크기, 연속성, 가이드라인

† 교신저자: 김수영, 이화여자대학교 심리학과, 서울시 서대문구 이화여대길 52
Tel: 02-3277-3792, E-mail: suyoung.kim@ewha.ac.kr

심리학을 포함한 사회과학 영역에서 구조방정식 모형은 관찰변수 또는 잠재변수 간 관련성을 설명하기 위해 보편적으로 사용되는 모형 중 하나이다. 구조방정식 모형을 이용할 때 연구자는 모형이 자료를 설명하는 정도를 나타내는 적합도의 개념을 통해 설정한 모형이 유용한가에 대한 판단을 내리게 된다. 모형의 유용성은 일반적으로 통계적 유의성(statistical significance)과 실질적 유의성(practical significance)을 통해 판단할 수 있으며, 구조방정식 모형에 대한 적절한 판단을 내리기 위해서는 적합도의 통계적 유의성과 실질적 유의성을 평가하는 도구들의 서로 다른 목적과 역량을 이해하고 이를 올바르게 사용하는 것이 중요하다. 그럼에도 불구하고 지금까지 출판된 국내외 논문들에서는 모형적합도의 평가 과정을 정확하게 이해하지 못하고 추정의 결과를 단편적으로 서술하는 경우가 매우 빈번하였다. 이에 본 연구는 모형 적합도의 유의성을 판단하는 기존의 다양한 평가 도구들의 목적을 바탕으로 전반적인 적합도 평가 과정을 통합 정리하고, 특히 실질적 유의성을 평가하는 단계에 효과크기의 개념을 반영하여 보다 실용적이며 적절한 모형 평가 방식에 대하여 논의한다. 이는 다양한 종류의 구조방정식 모형이 광범위하게 이용되고 있는 현재, 내용 영역의 연구자(substantive researchers)들이 모형의 유용성을 올바르고 적절하게 판단하기 위해 꼭 짚고 넘어가야 하는 주제이다.

적합도의 통계적 유의성을 평가하는 대표적인 도구인 χ^2검정(Jöreskog, 1969)의 경우, 모형이 자료를 완벽하게 설명한다는 완전 적합(exact fit) 가설을 설정하고 이에 대한 통계적 검정을 진행한다. 유의수준 α와 p값을 비교하여 적합도를 완전 적합 혹은 불완전 적합으로 해석하는 이분법적 방식을 통해 연구자는 모형이 자료를 완벽하게 설명하는지 아닌지에 대한 통계적 유의성의 정보를 얻게 된다. 반면 적합도의 실질적 유의성의 경우 모형과 자료 간 차이가 완전 적합으로부터 얼마나 떨어져 있는가를 연속선 상에서 평가한다. 일반적으로 통계 모형이나 모수의 실질적 유의성을 평가하기 위해 Cohen의 d, η^2, R^2과 같은 효과크기 지수들이 사용되는데, 구조방정식의 경우 적합도 지수를 이용해 적합도의 효과크기를 확인할 수 있다. 적합도 지수는 χ^2검정이 기각되었을 때 적합도가 완전 적합으로부터 얼마나 떨어져 있는지에 대한 효과크기를 연속적인 관점에서 평가한다. Yuan과 Marshall (2004)은 구조방정식 모형의 맥락에서 일반적인 효과크기의 형태와 유사한 새로운 적합도 평가 지수를 제안하였으며, Maydeu-Olivares (2017)는 통계적으로 유의한 차이 검정 결과에 대하여 효과크기를 이용해 그 차이를 질적으로 확인하듯, 유의한 적합도 검정 결과에 대하여 적합도 지수를 이용해 모형과 자료 간 차이의 효과크기를 확인할 것을 제안하였다. 나아가, Gomer 등(2019) 역시 다양한 종류의 구조방정식 모형의 효과크기를 제시하고 이를 이용한 시뮬레이션 결과들을 제공하였다.

χ^2검정과 적합도 지수는 이분법적 평가와 연속선 상에서의 평가라는 각각의 방식을 통해 모형의 유용성을 평가한다. 이는 곧 χ^2검정 결과를 연속적으로 해석하거나, 반대로 추정된 적합도 지수 값을 이분법적으로 해석하는 것은 잘못된 평가 방식임을 의미한다. 특히 적합도 지수의 경우, 그 해석에 사용되는 가이드라인을 적합도 평가의 절대적 기준으로 사용해서는 안 됨을 여러 연구에서 지속적으로 경고하였으나(Barrett, 2007; Gomer et al.,

2019; Markland, 2007; Marsh et al., 2004), 그럼에도 이를 모형의 유용성에 대한 이분법적 통과 기준으로 사용하는 관행은 사라지지 않고 있다. 본 연구에서는 총 여섯 개의 프리미엄 저널[1]에서 2020년부터 2022년까지 출판된 연구 가운데 적합도 지수를 이용해 구조방정식 모형을 평가한 250개의 연구를 분석한 결과, 대부분의 연구에서 적합도 지수와 그 가이드라인은 마치 χ^2검정의 p값과 α의 관계처럼 사용되고 있었다. p값이 α보다 작으면 영가설을 기각하고 α보다 크면 영가설 기각에 실패하듯, 적합도 지수가 가이드라인에서 제시하는 기준을 만족하면 모형이 좋거나(good) 적절하고(acceptable) 기준을 만족하지 못하면 나쁘거나(bad) 적절하지 않다(unacceptable)고 해석하는 현상이 대부분의 연구에서 확인되었다.

나아가, 250개의 연구 가운데 설정된 모형에 대하여 적합도 지수 값이 가이드라인의 기준에 근접하나 만족은 하지 못한 적합도(marginal fit)를 나타낸 32건의 사례 가운데 절반 이상의 연구에서 해당 모형은 자료를 충분히 설명하고 있지 못한 것으로 보고되었다. 특히 Tyler 등(2020)의 경우 적합도 지수가 기준에 매우 근접함에도 불구하고(TLI=.89, CFI=.91, RMSEA=.06) 제시된 모형을 배제하는 등, 현재 적합도 지수는 많은 연구에서 단순히 특정한 값을 기준으로 적합도에 대한 이분법적 판단을 내리는 통계적 검정과 유사한 방식으로 사용되고 있었다.

적합도 지수가 본래의 목적에 맞게 연속적으로 사용되기 위해서는 적합도 지수를 통계적 검정이 아닌 효과크기 지수로서 인식하고 사용할 필요가 있다. 통계적 검정이 α를 영가설 기각여부의 절단점으로 사용하는 반면, 효과크기 지수의 가이드라인은 어디까지나 적합도의 수준이 얼마나 높아졌거나 낮아졌는가를 나타내는 일종의 알림판 역할에 불과하다. 가이드라인의 기준값은 적합도를 이분법적으로 평가하는 근거가 될 수 없으며, 단순히 연속선 상에서 모형 적합도의 위치를 해석하는 과정에서 참고하는 보조적인 지표에 불과하다.

적합도 지수를 효과크기 지수로 사용하는 과정에 이와 같은 가이드라인의 연속성(continuity)이 제대로 고려되지 않을 경우, 특히 효과크기 지수의 특성에 대해 완벽히 이해하지 못하고 있는 연구자는 적합도의 효과크기를 이분법적으로 해석하는 잘못을 저지를 수 있다. 실제로 적합도 지수를 해석할 때 인용되는 저명한 가이드라인들(Bentler & Bonett, 1980; Browne & Cudeck, 1993; Hu & Bentler, 1999)은 적합도 지수의 연속적인 성격을 제대로 반영하지 못하는 절단적인 기준과 해석을 제시하며, 이와 같은 문제는 비교적 최근 제안된 새로운 가이드라인(Asparouhov&Muthén, 2018; Shi et al., 2018)까지 이어지고 있다. 그러나 적합도를 평가하고 모형을 수정하는 과정에서 연속성이 반영되지 않은 가이드라인이 이용될 경우, 가이드라인의 기준값을 아주 조금이라도 만족하지 못한다는 이유 하나만으로 충분히 유용한 모형을 배제하는 비효율적이며 실용적이지 못한 모형의 평가를 하는 문제가 발생할 수 있다.

본 연구의 목적은 적합도 효과크기 지수의 종류 및 연속선 상의 해석 방식에 대한 논의

1) Journal of applied psychology, Journal of applied developmental psychology, Journal of educational psychology, Journal of counseling psychology, Journal of abnormal psychology, Journal of personality and social psychology

를 제시하는 것이다. 이를 통하여 모형 자체의 유용성에 대해 지금 당장 평가를 내려야 하는 내용 영역 연구자들이 직접적으로 참고할 수 있는 실용적인 모형 평가 및 해석 방법을 제안하고자 한다. 적합도 효과크기 지수의 경우 이미 알려진 여러 적합도 지수 이외에도 Maydeu-Olivares(2017) 및 Gomer 등(2019)의 최근 연구들을 통해 제안된 새로운 지수들이 존재하여 이를 소개하고자 한다. 구조방정식 모형의 맥락에서 사용할 수 있는 다양한 효과크기 지수들을 소개하는 것은 적합도의 실질적 유의성의 평가 방법에 대한 이해를 넓힐 수 있을 것이다. 또한, 본 연구는 적합도의 효과크기 지수가 본래의 쓰임에 맞게 사용될 수 있도록 가이드라인의 연속성과 임의성이라는 특징을 바탕으로 각 지수의 기준값과 해석 방식을 재정리한 가이드라인의 예시를 제공한다. 해당 가이드라인을 통해 효과크기 지수를 연속적으로 해석함에 따라 기존에는 배제되었던 모형이 유용한 모형으로 평가될 수 있는 상황에 대해 논의하고 적합도 평가도구의 올바른 해석 방식의 중요성을 재고한다.

이와 같은 목적을 달성하기 위하여 본 연구에서는 우선 적합도의 첫 번째 평가단계에 해당하는 χ^2검정 과정에 대해 간략히 리뷰한다. 이후, χ^2검정의 대안으로 제시된 적합도 지수를 적합도의 효과크기 지수로 적용하는 것에 대한 가능성과 함께 다양한 종류의 효과크기 지수들을 정리하여 제시한다. 마지막으로, 적합도의 효과크기를 해석하기 위한 기준값과 가이드라인의 올바른 예시 및 사용 방식을 논하고, 실제 연구에서 효과크기의 연속성을 반영하지 못함으로 인해 발생하는 잘못된 적합도 평가 사례를 확인한다. 특히 아주 근소한 차이로 기준값을 만족하지 못하는 모형(marginal fit model)에 대하여 연속성이 반영된 가이드라인을 이용해 해석할 경우 모형의 유용성과 무용성에 대한 결과가 다르게 나타날 수 있음을 실제 사례를 통해 제시한다.

적합도의 통계적 유의성

적합도의 효과크기 개념에 대해 본격적으로 논의하기 전, 구조방정식 발전 초기 대표적인 모형 평가 방법으로 사용되었던 χ^2검정을 이용해 적합도의 통계적 유의성을 확인하는 과정을 검토한다. 또한, 실제 연구에서 표본크기와 관련된 χ^2검정 결과의 유용성에 대해 논의하고 검정 결과가 적합도 평가 과정에서 갖는 의미에 대하여 재고한다.

완전 적합가설의 검정

공분산 구조 분석(covariance structure analysis)의 맥락에서 Jöreskog(1969)가 소개한 χ^2검정은 수집된 자료의 공분산 행렬과 추정된 공분산 행렬(모형 함의 공분산 행렬) 간의 차이, 즉 적합도의 통계적 유의성을 평가한다. χ^2검정의 영가설은 $\Sigma = \Sigma(\theta)$로, 모집단의 공분산 행렬(Σ)과 모수 기반의 모형 함의 공분산 행렬($\Sigma(\theta)$) 간 차이가 없음을 가정한다. 이때 모집단의 수준에서 두 행렬 간 차이는 직접적으로 계산할 수 없기 때문에, 자료와 모형의 차이는 표본의 수준에서 두 공분산 행렬 간 차이를 최소화하는 합치함수 $F(S,\Sigma(\hat{\theta}))$을 통해 추정된다. 다양한 종류의 합치함수 가운데 가장 대표적으로 사용되는 최대우도 (maximum likelihood, ML) 합치함수 F_{ML}은 아래와 같다.

$$F_{ML} = \log|\Sigma(\hat{\theta})| + tr(S\Sigma^{-1}(\hat{\theta})) - \log|S| - p \quad (1)$$

위에서 S와 $\Sigma(\hat{\theta})$은 각각 표본의 공분산 행렬과 추정된 공분산 행렬을, $\hat{\theta}$은 추정치 벡터를, p는 변수의 개수를 의미한다. 식 1을 통해 계산된 F_{ML}을 이용해 χ^2검정통계량 T_{ML}을 아래와 같이 계산할 수 있다.

$$T_{ML} = (n-1)F_{ML} \text{ 또는 } nF_{ML} \quad (2)$$

F_{ML}의 계산에 사용되는 표본의 공분산 행렬 S가 어떤 분포를 따르는가에 따라 F_{ML}에 n을 곱할지 $(n-1)$을 곱할지가 달라진다. 예를 들어, Mplus의 경우 S가 다변량 정규분포를 따른다고 가정함에 따라 F_{ML}에 n을, LISREL의 경우 S가 Wishart 분포를 따른다고 가정(Hayduk, 1987)함에 따라 F_{ML}에 $(n-1)$을 곱한다. 두 계산 방식은 표본크기가 증가함에 따라 점근적으로 유사한 결과를 제공한다. 식 1에서 표본크기가 충분히 크고, 내생변수들이 다변량 정규성을 만족하며, 모형이 올바르게 설정되었다는 가정 하에 T_{ML}은 점근적으로 χ^2분포를 따른다.

앞서 언급한 바와 같이, χ^2검정의 영가설 $(\Sigma = \Sigma(\theta))$은 모형이 자료를 완벽하게 반영하는 완전 적합을 가정한다. 이는 중요한 함의점을 가지고 있는데, χ^2검정의 기각이 연구자가 설정한 모형이 자료를 설명하는 데 무조건 실패했음을 가리키는 것이 아니라는 것이다. 다만 χ^2검정의 기각은 모형이 자료를 완벽하게 설명하고 있지는 않음을 의미한다. 즉, χ^2검정은 연속선 상에 놓인 적합도의 여러 수준 가운데 완전 적합이라는 하나의 기준에 대한 이분법적 판단을 통해 적합도를 평가하는 통계적 기법인 것이다. 모형 적합도 평가 도구로 소개된 이래 여러 연구에서 적합도 평가 시 반드시 χ^2검정 결과를 보고할 것을 지속적으로 제안하였으며(Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; Kline, 2016; Markland, 2007), 실제로 본 연구에서 검토한 250개의 연구 가운데 218개의 연구가 χ^2검정의 결과를 보고하였다.

χ^2검정과 표본크기

연구자가 적합도 평가를 위해 통계적 검정을 이용할 경우, 결과에 대한 해석은 검정의 영가설에 대하여 이루어져야 한다. 이는 곧 χ^2검정 결과의 해석이 완전 적합을 기준으로 이루어져야 함을 의미한다. 그러나 χ^2검정 결과를 보고한 218개의 연구 가운데 기각된 검정 결과에 대하여 모형이 자료에 완벽하게 적합하지는 않음을 설명하는 연구는 찾을 수 없었다. 나아가, χ^2검정의 p값조차 보고하지 않은 사례도 많아, 전반적으로 χ^2검정 결과에 대한 보고와 해석이 제대로 이루어지고 있지 않음을 확인할 수 있었다.

χ^2검정의 완전 적합 영가설 기각 결과를 무시하는 근거들을 제시한 연구들 가운데 대다수는 표본크기를 그 근거로 제시하였다. χ^2검정에 사용되는 검정 통계량 T_{ML}은 점근적으로 χ^2분포를 따르며, 이에 따라 표본크기가 충분히 크다면 검정 통계량이 χ^2분포를 따르지만 표본크기가 작을 경우 χ^2분포를 따르지 않을 수 있다. 또한, 표본크기가 클 경우 검정 통계량이 커짐에 따라 모형과 자료의 실제 차

이가 작더라도(즉, 합치함수 F_{ML}의 값이 작더라도) 모형은 기각될 수 있다. 실제로 표본크기가 200 혹은 400만 넘어가도 대부분의 모형에서 χ^2검정은 기각되는 것으로 알려져 있다(Barrett, 2007; Kenny, 2020).

표본크기가 검정의 결과에 영향을 미치는 문제는 여러 연구에서 꼽은 χ^2검정의 대표적인 한계에 해당하며(Fan, Thompson, & Wang, 1999; Hu & Bentler, 1995; Marsh & Balla, 1994), 그에 따라 χ^2검정 결과를 그대로 받아들여서는 안 된다거나(Goffin, 2007), 혹은 χ^2검정을 대체할 수 있는 다른 종류의 통계적 검정을 제안하는 연구(Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996)들이 발표되었다. 그러나 χ^2검정이 t검정, 혹은 F검정과 같은 일반적인 통계적 검정의 종류 중 하나임을 생각했을 때, 표본크기의 문제는 χ^2검정뿐 아니라 일반적인 통계적 검정 전반에서 나타나는 대표적인 한계점이다(Thompson, 1996). 통계적 검정이 갖는 한계들을 언급한 다양한 연구들은(Kirk, 1996; Meehl, 1967; Tukey, 1991; Wilkerson & Olson, 1997) 영가설의 비현실성 또는 유의수준을 이용한 이분법적 판단의 문제와 함께 표본크기가 검정 결과에 영향을 미치는 문제를 언급하였으며, 특히 Fan(2001)은 통계적 검정이 갖는 여러 문제 가운데 가장 대표적인 한계점으로서 표본크기를 꼽았다. 하지만 표본크기를 근거로 t검정이나 F검정, 혹은 통계적 검정 자체가 의미 없다고 주장하는 연구는 없으며, 일반적으로 표본크기의 영향을 받지 않고 검정 결과를 해석할 수 있는 대안적인 도구(예, 효과크기)를 개발하는 방향으로 발전되었다.

나아가 χ^2검정의 경우 현재 모형 적합도를 평가하는 여러 방법 중 거의 유일하게 적합도의 통계적 유의성을 평가한다(Barrett, 2007). 가장 일반적으로 사용되는 적합도 지수 중 하나인 RMSEA(root mean square error of approximation)를 이용한 근사 적합(close fit) 검정이 있기는 하지만, 이를 제외한 TLI (Tucker-Lewis index), CFI(comparative fit index), SRMR(standardized root mean square residual) 등의 경우 RMSEA와 동일하게 비중심 χ^2분포를 이용하거나 새로운 분포 기반의 불편향 추정치(Maydeu-Olivares, 2017)가 제안되었음에도 아직 이를 이용해 적합도에 대한 통계적 검정을 진행하는 과정은 대중화되지 못하였다. 즉, 현재 χ^2검정은 적합도의 통계적 유의성을 확인할 수 있는 대표적인 평가 도구에 해당하며, 표본크기를 비롯한 몇 가지 문제들이 χ^2검정의 결과를 무시하고 제대로 보고하지 않는 근거가 될 수는 없다. χ^2검정을 통해 얻게 되는 통계적 유의성 결과에 더하여 모형과 자료 간 차이의 수준이라는 실질적 유의성에 대한 해석이 보충된다면 χ^2검정은 그 자체로 충분히 의미 있는 적합도 평가에 해당한다(Maydeu-Olivares, 2017; Steiger, 1989).

적합도의 실질적 유의성

연구자는 χ^2검정을 통해 완전 적합에 대한 통계적 유의성을 검정하고, 검정이 기각되면 적합도의 효과크기 지수를 통해 모형이 완전 적합으로부터 얼마나 떨어져 있는지에 대한 실질적 유의성을 확인함으로써 적합도를 종합적으로 평가할 수 있다. 그림 1은 이와 같은 적합도의 전반적인 평가 과정을 도식화하여 제시한다.

구조방정식 모형의 효과크기

실질적 유의성의 확인

일반적인 통계적 검정이 끝나고 통계적 유의성을 확인한 뒤, 연구자는 효과크기 지수와 같은 도구를 이용해 검정의 실질적 유의성에 대하여 확인한다. 통계적 유의성이 검증되었음에도 불구하고 추가적으로 실질적 유의성을 확인해야 하는 이유는 가설검증 결과가 표본크기의 영향을 받기 때문이다. 두 집단의 평균 차이가 동일하더라도, 표본크기에 따라 통계적 유의성의 결과는 달라진다. 만일 두 집단의 평균 차이가 매우 작음에도 불구하고 표본크기가 크다면, 표집분포의 표준오차 값은 줄어들며 결과적으로 검정통계량은 매우 큰 값으로 계산되어 통계적으로 유의한 결과를 제시할 확률이 올라간다. 그러나 이는 실질적인 유의미성을 의미하는 것이 아닌, 표본크기에 의해 왜곡된 통계적 유의미성에 지나지 않는다. 표본크기의 영향을 배제하고 두 집단 간의 실질적인 차이를 확인하기 위해서는 검정의 실질적 유의성을 확인할 필요가 있다.

완전 적합 가설 검정에 대한 효과크기

모형의 실질적 유의성을 해석하는 대표적 평가도구인 효과크기는 '해당 현상이 모집단에 존재하는 정도(the degree to which the phenomenon is present in the population)' 또는 '영가설이 잘못된 정도(the degree to which the null hypothesis is false)'(Cohen, 1988)를 나타내는 지수이다. 이와 같은 맥락에서 구조방정식 모형 적합도의 효과크기는 χ^2검정의 영가설인 완전 적합 가설이 잘못된 정도, 즉 모형의 적합도가 연속선상에서 완전 적합으로부터 떨어진 정도를 나타내는 개념으로 정의할 수 있다. 이는 곧 적합도의 수준과 효과크기의 부적 관계를 의미하는데, 구체적으로 모형과 자료 간 차이가 증가할수록 효과크기의 값은 커지게 되며, 그에 따라 적합도의 수준은 낮아지고, 연구자의 모형은 점점 지지할 수 없게 됨을 의미한다.

구조방정식 모형 적합도의 효과크기는 일반

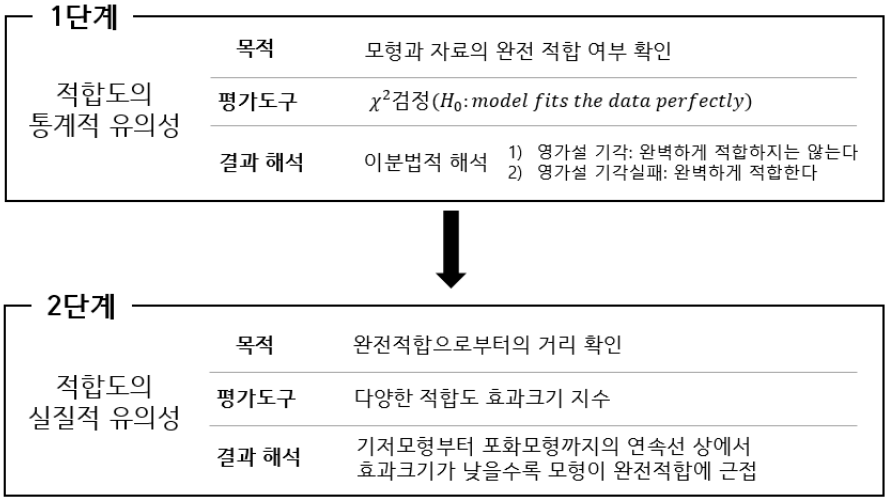


그림 1. 모형적합도의 전반적인 평가 과정

적으로 값이 커질수록 연구자가 주장하고자 하는 가설을 지지하는 전통적인 효과크기(예, Cohen의 d)와 달리 그 값이 작아질수록 모형을 지지한다는 점에서 구분되는 해석상의 차이를 갖는다. 이는 일반적인 차이 검정에서 사용하는 영가설과 달리 구조방정식 모형의 완전 적합 검정은 영가설을 기각하지 않아야 연구가설이 지지 되는 수용-지지 검정(accept-support test)[2]에 해당하기 때문이다(Kline, 2016). 완전 적합 가설을 기각하는 데 실패하는 것이 연구자의 모형을 지지하는 일이 되며, 완전 적합 가설에서 멀어짐에 따라 효과크기는 증가하고 모형의 설명력은 낮아진다.

적합도 지수를 이용한 효과크기의 평가

χ^2검정이 갖는 표본크기 등의 한계를 보완함과 동시에 이분법적 프레임을 벗어난 적합도의 평가를 위해 발전된 적합도 지수는 연속성이라는 특징을 바탕으로 적합도의 효과크기 지수로서 사용될 수 있다. 1980년대부터 최근까지 다양한 적합도 지수들이 제시되었으며, 어떠한 관점에서 적합도를 정의하는가에 따라 몇 가지 범주(예, 상대적 적합도 지수, 절대적 적합도 지수 등)로 분류될 수는 있으나, 모든 지수는 공통적으로 적합도의 수준을 연속적으로 확인한다는 점에서 효과크기와 동일한 목적을 갖는다.

적합도 지수와 일반적인 효과크기 지수 간의 유사성, 또는 적합도 지수에 효과크기의 개념을 적용할 수 있다는 주장은 이전부터 지속적으로 제기되었다(Hu & Bentler, 1999; Maydeu-Olivares, 2017; Yuan & Marshall, 2004). 적합도 지수와 효과크기는 자료와 모형 간의 관계, 또는 독립변수와 종속변수 간의 관계의 정도를 기술적으로 나타내며, 그와 동시에 모수를 기반으로 하는 분포(예, 비중심 χ^2분포) 하에서 점추정치 및 구간 추정치의 형태로 정의될 수 있다. 또한, Jöreskog와 Sörbom(1981)은 초기 적합도 지수인 GFI(goodness of fit index)를 일반 선형 모형의 효과크기 지수인 R^2과 유사한 역할을 하는 지수로 소개하였으며, 이후 GFI는 구조방정식 모형의 결정계수로 사용되기도 하였다(Tanaka & Huba, 1989). NFI(normed fit index), CFI(comparative fit index), TLI(Tucker-Lewis index) 역시 선형회귀분석에서의 R^2과 같은 역할을 하는 것으로 해석할 수 있다(Laar & Braeken, 2022). Hu와 Bentler(1999)도 적합도 지수는 R^2과 같이 자료의 분산 가운데 모형에 의해 설명된 분산의 양을 측정하는데 사용해야 하며, 적합도 지수를 통계적 검정의 도구처럼 사용하자는 의견(Maiti & Mukherjee, 1991)에 대하여 적합도 지수의 목적과 부합하지 않음을 주장했다.

반면, 적합도 지수를 이용해 모형의 효과크기를 확인하는 것이 절대적으로 불가능한 것은 아니지만 선호되지도 않는다는 주장 또한 제기되었다. 대표적으로 Gomer 등(2019)은 현재 적합도 지수 가이드라인의 기준값이 가설 검정의 맥락에서 2종 오류의 통제(Hu & Bentler, 1999)를 위해 설정된 값에 불과하며, 또한 자료가 정규성을 만족하지 못하거나 모형 조건이 달라짐에 따라 적합도 지수가 편향될 수 있다는 문제점 등을 이유로 적합도 지수를 효과크기와 같은 개념으로 보는 것이 적절하지 않음을 주장하였다. 그러나 자료의 비정규성 문제의 경우, 합치함수 F_{ML}을 추정하

2) accept-support test는 통계 철학적으로 옳지 않은 표현이지만, 실질적으로 구조방정식 모형 적합도 검정의 특성을 잘 표현하고 있다.

는데 사용되는 최대우도 추정법이 본래 정규성 가정의 위반에 상당히 강건한 것으로 알려져 있으며(Schermelleh-Engel et al., 2003), 또한 현재 Mplus와 EQS 등의 통계 프로그램이 다양한 적합도 지수의 추정에 이용되는 검정 통계량 T_{ML}을 $T_{SB}(=\frac{T_{ML}}{\hat{c}}$, $\hat{c}$는 자료의 비정규성 수준을 고려한 척도화 계수)로 대체함으로써 비정규성에 대한 교정을 적용할 수도 있어(Brosseau-Liard & Savalei, 2014), 자료의 비정규성 문제로 인해 적합도 지수를 효과크기 지수로 사용하지 못할 이유는 없다고 볼 수 있다. 비록 여러 시뮬레이션 연구에서 적합도 지수가 표본크기나 지표변수의 개수와 같은 모형 조건에 의해 영향을 받는다는 사실이 밝혀졌으나(Ding, Velicer, & Harlow, 1995; Fan, Thompson, & Wang, 1999; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998), 해당 연구들은 공통적으로 적합도 지수의 편향과 가이드라인을 맹신하는 관행에 대하여 경고했을 뿐 적합도 지수를 이용해 모형이 잘못 설정된 정도를 확인하는 행위 자체에 대해 의문을 제기하지는 않았다. 본래의 목적에 맞게 연속선 상에서 모형과 자료의 차이를 확인하는 도구로 사용한다면 적합도 지수는 충분히 구조방정식 모형 적합도의 효과크기 지수로서 활용될 수 있다.

적합도 효과크기 지수의 종류

전통적 효과크기 지수

RMSEA, SRMR, CFI와 같이 오래전부터 사용되었던 적합도 지수들의 경우, 적합도의 수준을 연속선 상에서 파악한다는 점에서 구조방정식 모형의 효과크기 역할을 담당할 수 있다. 자유도에 의해 조정된 모형과 자료 간의 거리를 나타내는 RMSEA는 완전 적합 가설이 옳지 않다는 가정 아래의 분포인 비중심 χ^2 분포의 비중심 모수(noncentrality parameter) λ를 이용해 추정된다. 표본크기가 지수에 미치는 영향에 대한 교정과 모형의 복잡성에 대한 페널티, 그리고 단위의 통일 등을 거쳐 최종적으로 RMSEA는 다음과 같이 정의 된다(Steiger, 1989).

$$RMSEA = \sqrt{\frac{\lambda}{df(n-1)}} \tag{3}$$

위에서 n은 표본크기, df는 모형의 자유도를 나타낸다.

식 3의 RMSEA는 모집단의 공분산 행렬 Σ를 이용하여 정의되는 모수에 해당하기 때문에 실제로 값을 구할 수는 없으며, 일반적으로 RMSEA를 이용한 적합도의 평가는 점 추정치인 $\widehat{RMSEA}$과 구간 추정치인 90% 신뢰구간을 통해 이루어진다.

$$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}} \tag{4}$$

RMSEA는 모형과 자료 간 차이의 크기를 나타내는 데 있어 추정치의 단위가 표준화 되어있지 않으며 원변수의 단위를 그대로 이용함에 따라(Maydeu-Olivares, Shi, Rosseel, 2018) 전통적 효과크기 지수 가운데에서도 대표적인 비표준화 지수에 해당한다. 비록 비표준화 지수의 경우 추정된 값의 해석이 모형의 구조와 크기에 따라 달라지는 한계를 갖지만(Chen, Curran, Bollen, Kirby, & Paxton, 2008; Kline,

2016), 그럼에도 다양한 통계 프로그램에서는 RMSEA의 추정치를 이용한 근사 적합(close fit) 검정 결과를 제공하고 있으며, 현재 가장 대표적인 적합도 효과크기 지수 중 하나로 사용되고 있다.

SRMR은 Σ와 $\Sigma(\theta)$ 간 차이인 잔차 행렬을 이용해 적합도의 수준을 확인하는 지수로, 아래와 같이 정의된다.

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(\sigma_{ij}-\sigma_{ij}^{0})/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}\right\}^{2}}{p(p+1)/2}} \quad (5)$$

위에서 σ_{ij}와 $\sigma^{0}{}_{ij}$는 각각 Σ와 $\Sigma(\theta)$의 요소를 의미하며, p는 변수의 개수를 의미한다. σ_{ij}와 $\sigma^{0}{}_{ij}$를 각각의 표준편차로 나눠주는 표준화 과정을 통해 SRMR은 자료와 모형 간의 차이를 표준화된 값으로 나타낸다.

적합도의 실제 평가에 사용되는 추정치 $\widehat{SRMR}$의 경우 S와 $\Sigma(\hat{\theta})$의 요소인 s_{ij}와 $\hat{\sigma}_{ij}$을 이용해 아래와 같이 정의된다.

$$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij}-\widehat{\sigma_{ij}})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^{2}}{p(p+1)/2}} \quad (6)$$

SRMR은 대표적인 표준화 효과크기 지수로, RMSEA와 비교하여 '잔차에 대한 표준화된 공분산 행렬의 평균', 혹은 '표준화된 효과크기의 평균' 등의 정의와 함께 추정치의 크기를 직접적으로 해석할 수 있다는 장점이 있다. 다만, 표준화된 효과크기 지수라 해서 SRMR이 반드시 0에서 1 사이의 값을 갖는 것은 아니다. 실제 연구에서는 대부분의 SRMR 값이 1 이하의 값을 나타내지만, 이론적인 SRMR의 범위는 1 이상의 값을 가질 수 있다(West et al., 2012). 또한 표준화 된 자료의 공분산 행렬과 모형 함의 공분산 행렬 간의 차이를 나타내는 SRMR의 정의를 고려했을 때, 일반적으로 추정된 값이 2를 넘는 것은 이론적으로 불가능할 것으로 예상할 수 있다.

CFI는 변수 간 관계를 설정한 연구모형이 변수 간 어떠한 관계도 존재하지 않는 영모형(null model)에 비하여 자료를 얼마나 더 잘 설명할 수 있게 되었는지를 나타내는 상대적 효과크기 지수 가운데 가장 범용적으로 사용되는 지수이다. 상대적 효과크기 지수는 증분적합도 지수라고도 하는데, 일반적인 증분 적합도 지수의 모수 Δ는 아래와 같이 정의된다(Bentler, 1990).

$$\Delta = 1 - \frac{\lambda_M}{\lambda_N} \quad (7)$$

위에서 λ_M은 연구모형의 비중심 모수, λ_N은 영모형의 비중심 모수를 가리킨다. λ의 값이 커진다는 것은 모형이 자료를 제대로 설명하지 못함을 의미하며, 이에 따라 λ_N에 비하여 λ_M이 작아질수록 Δ는 커지게 된다. CFI는 Δ의 다양한 추정치 가운데 값의 범위를 0에서 1 사이로 고정하여 구하는 추정치로, 아래와 같이 정의된다.

$$\begin{aligned} CFI &= 1 - \frac{\widehat{\lambda_M}}{\widehat{\lambda_N}} \\ &= 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)} \end{aligned} \quad (8)$$

CFI는 부스트래핑을 이용해 신뢰구간을 추정하는 것이 가능하며(Cheng & Wu, 2017; Lai, 2019; Zhang & Savalei, 2016), 현재 TLI와 함께 대표적인 상대적 효과크기 지수로 사용되고 있다. 다만, TLI와 CFI는 상관이 높기 때문에 두 지수 중 하나만 보고할 것이 제안된다(Kline, 2016). 또한, CFI는 프로그램에 따라 조금 다른 값이 제시되기도 하는데, 이는 각 프로그램이 정의하는 영모형이 다르기 때문이다(예, Mplus와 EQS).

새로운 효과크기 지수

최근 모형의 적합도를 효과크기의 관점에서 평가해야 한다는 주장들(Gomer et al., 2019; Maydeu-Olivares, 2017; Maydeu-Olivares & Shi, 2017)과 함께 새로운 종류의 적합도 효과크기 지수들이 제안되었다. 기존의 전통적 효과크기 지수들과 비교하여 새로운 지수들은 한층 더 발전된 형태의 추정치를 가지며, 효과크기의 성격을 잘 나타내고 있다. 대표적으로 Maydeu-Olivares(2017) 및 Maydeu-Olivares와 Shi (2017)는 전통적 효과크기 지수인 $\widehat{SRMR}$이 모수 SRMR을 특히 작은 표본에서 과대추정하고 있음을 밝히며, 이에 따라 모수에 대한 불편향 추정치인 $\widehat{SRMR}_u$(unbiased SRMR)을 새롭게 제안하였다. 대부분의 통계 소프트웨어에서 식 6을 통해 추정하는 $\widehat{SRMR}$의 경우 일반적으로 실제 모수를 과대추정하고 있으며, 이에 따라 소프트웨어를 통해 추정된 모형의 $\widehat{SRMR}$은 실제보다 낮은 수준의 적합도를 나타낸다(Shi, Maydeu-Olivares, & DiStefano, 2018). 이와 같은 문제를 해결하기 위하여 Maydeu-Olivares(2017)는 정규분포 하에서 정의되는 $\widehat{SRMR}_u$을 이용해 적합도의 효과크기를 확인할 것을 새롭게 제안하였다. 모수 SRMR에 대한 불편향 추정치 $\widehat{SRMR}_u$은 아래와 같이 정의된다.

$$\widehat{SRMR}_u = \widehat{k^{-1}}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}),0)}{t}} \quad (9)$$

위에서 $\sqrt{\frac{\max(e'e - tr(\hat{\Xi}),0)}{t}}$는 잔차의 제곱합을 의미하는 $e'e$의 기댓값을 이용하여 구한 SRMR의 추정치로, e는 자료와 모형 간의 표준화된 잔차를, $\hat{\Xi}$은 e의 공분산 행렬을 나타내며, t는 공분산 행렬의 독립적인 정보의 개수를 의미한다. $\hat{k}^{-1}$은 식 10과 같이 추정되며, 식 9를 통해 얻은 SRMR의 추정치가 편향되지 않도록 조정 해주는 역할을 한다.

$$\widehat{k^{-1}} = 1 - \frac{tr(\widehat{\Xi_s^2}) + 2e_s\hat{\Xi}e_s}{4(e_s'e_s)^2} \quad (10)$$

$\widehat{SRMR}_u$의 경우 정규분포를 바탕으로 신뢰구간을 추정할 수 있으며, 근사 적합에 대한 검정 역시 가능하다. 또한, $\widehat{SRMR}_u$이 새롭게 제안된 이후 시뮬레이션을 통해 $\widehat{SRMR}_u$을 이용해 구하는 구간 추정치와 근사 적합 검정 결과가 RMSEA를 통해 얻는 결과보다 더 정확하다는 연구(Maydeu-Olivares, Shi, & Rosseel, 2018)가 제시됨에 따라, $\widehat{SRMR}_u$이 적합도의 효과크기 수준을 나타내는 데 비교적 우월한 지수일 가능성이 확인되었다. 한편, Asparouhov와 Muthén(2018)은 $\widehat{SRMR}_u$이 Mplus에서 제공하는 $\widehat{SRMR}$과 다

른 방식으로 정의되어 있으며, $\widehat{SRMR}$은 큰 표본크기에서 사용될 수 있는 반면 $\widehat{SRMR}_u$의 경우 작은 표본크기에서의 SRMR의 추정에 핵심을 두고 있다도 주장하였다.

$\widehat{SRMR}_u$과 $\widehat{SRMR}$이 모두 SRMR에 대한 추정치로 사용되는 적합도 효과크기 지수라면, Gomer 등(2019)의 ϵ[3]는 두 집단 간의 평균 차이를 표준화한 값을 나타내는 Cohen의 d(Cohen, 1988)의 개념을 바탕으로 제안된 적합도 효과크기 지수이다. d는 검정의 종류에 따라 다양한 형태로 정의되는데, 이 가운데 독립표본 t검정의 맥락에서 d는 아래와 같다.

$$d = \frac{\overline{X_1} - \overline{X_2}}{s_p} \tag{11}$$

위에서 $\overline{X}_1$과 $\overline{X}_2$는 각 표본의 평균치를, s_p는 통합된 표준오차를 의미한다. 해당 수식을 모집단의 수준에서 정의할 경우 d의 모수 δ는 아래와 같이 정의된다.

$$\delta = \frac{\mu_1 - \mu_2 - 0}{\sigma_p} \tag{12}$$
$$= \frac{E[\overline{X_1} - \overline{X_2} \mid H_1] - E[\overline{X_1} - \overline{X_2} \mid H_0]}{SD}$$

위에서 $E[\overline{X}_1 - \overline{X}_2|H_1]$과 $E[\overline{X}_1 - \overline{X}_2|H_0]$는 각각 대립가설과 영가설 하에서의 표본 평균의 차이를 나타낸다. 구조방정식의 경우 기본적으로 다변량 구조를 따르고 있으며 공분산 행렬을 이용하기 때문에 두 표본 평균의 차이를 상수의 형태로 직접 변환할 수 있는 개념은 존재하지 않으나, 그 대신 자료와 모형의 차이를 나타내는 F_{ML} 혹은 T_{ML}을 이용하여 해당 개념을 대신할 수 있다. T_{ML}을 이용하여 식 12를 구조방정식의 맥락에 맞게 변형한 형태는 아래와 같다(Gomer et al., 2019).

$$\epsilon = \left(\frac{E[T_{ML}|H_1] - E[T_{ML}|H_0]}{N-1} \right)^{1/2} \tag{13}$$

ϵ는 Gomer 등(2019)이 d의 개념을 바탕으로 제안한 다양한 효과크기 지수 가운데 가장 표본크기의 영향을 적게 받음과 동시에 모형과 자료 간 차이를 제대로 탐지하는 것으로 밝혀졌다. ϵ의 추정치인 $\hat{\epsilon}$은 식 14와 같이 구할 수 있으며, 부스트래핑을 이용해 추정되는 신뢰구간과 함께 사용할 것이 추천된다.

$$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T^*_{ML}}}{N-1} \right)^{1/2} \tag{14}$$

위에서 $\overline{T}^*_{ML}$는 부스트래핑을 이용해 추정되는 T_{ML}의 값에 해당한다(Yuan & Marshall, 2004). 표 1은 전통적 적합도 효과크기 지수와 새로운 적합도 효과크기 지수의 모수와 추정치, 그리고 각 지수가 따르고 있는 분포를 종합적으로 제시하고 있다. 현재 대다수의 통계 프로그램에서는 적합도 효과크기 지수의 추정치로 $\widehat{RMSEA}$, $\widehat{SRMR}$, 그리고 CFI를 제공하며, 대부분의 연구자들은 해당 지수를 가이드라인과 비교하여 모형의 적합도를 평가한다.

3) Gomer 등(2019)은 해당 연구에서 ϵ를 '입실론'이 아닌 영어 알파벳 'E'로 명명하였다.

표 1. 전통적 적합도 효과크기 지수와 새로운 적합도 효과크기 지수

모수	추정치	분포
$RMSEA = \sqrt{\frac{\lambda}{df(n-1)}}$	$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}}$	비중심 χ^2분포
$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(\sigma_{ij} - \sigma_{ij}^0)/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$	$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$ $\widehat{SRMR}_u = \hat{k}^{-1}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}), 0)}{t}}$	- 정규분포
$\Delta = 1 - \frac{\lambda_M}{\lambda_N}$	$CFI = 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)}$	비중심 χ^2분포
$\epsilon = \left(\frac{E[T_{ML}\|H_1] - E[T_{ML}\|H_0]}{N-1}\right)^{1/2}$	$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T}^*_{ML}}{N-1}\right)^{1/2}$	-

주. 표에 제시된 분포는 각 적합도 효과크기 지수를 정의하는데 이용된 분포를 의미한다.

가이드라인을 이용한 적합도의 효과크기 해석

효과크기 지수를 바탕으로 적합도의 실질적 유의성을 평가하는 과정에서 연구자는 가이드라인을 통해 적합도의 수준을 연속적으로 해석해야 한다. 이는 곧, 가이드라인이 제시하는 기준값이 적합도의 좋고 나쁨에 대한 절대적인 절단점이 아닌, 연속선 상에서의 해석을 위한 지표로 사용되어야 함을 의미한다.

적합도 효과크기 지수 가이드라인의 특징

기준값의 연속성과 임의성

적합도의 효과크기는 모형과 자료 간 차이를 나타내는 모수로 정의되며, 모수에 대한 추정치를 구하여 그 수준을 평가한다. 연구자는 적합도 추정치와 관습적으로 제시하는 가이드라인의 기준값들을 비교하여 연속선 상에서 모형과 자료 간의 거리를 파악할 수 있다.

적합도 지수의 가이드라인을 포함하여 통계학에서 사용되는 대다수의 관습적인 가이드라인들의 기준값을 정의하는 과정에는 연속성과 임의성이 반영된다. 대표적인 효과크기 지수 d(Cohen, 1988)의 경우, 추정된 d값의 해석을 위한 가이드라인에 따라 d= .50이 중간 정도의 효과크기를 나타낸다고 해석한다. 이때 추정된 d값이 .49 혹은 .48이라고 해서 처치의 효과가 작다고 해석하는 경우는 없다. 독립표본 t검정에서의 d는 두 모집단 간 차이의 수준을 연속적으로 나타내는 지표이며, .50이라는 기준값은 임의적으로 결정된 값이기 때문이다. RMSEA를 이용한 근사 적합 검정의 경우 일반적으로 사용되는 기준은 .05이다. 그러

나 근사 적합의 개념 역시 연속적인 적합도 수준의 한 지점을 의미할 뿐이며, .05는 임의로 정해진 상수로 모형의 설정이나 표본크기에 따라 기준값은 변할 수 있다(Chen et al., 2008).

나아가 RMSEA, SRMR, CFI와 같은 대표적인 적합도 효과크기 지수들의 가이드라인이 제시하는 기준값의 경우, 대부분 엄격한 수리적 과정을 통해 도출된 것이 아닌 연구자의 오랜 경험을 기반으로 제시된 값에 해당한다. Browne과 Cudeck(1993)은 다양한 모형의 RMSEA를 추정한 결과 .05 이하의 값이 나오는 모형은 자료와 상당히 근접한 것으로 평가할 수 있음을 제안했다. Bentler와 Bonett(1980)은 TLI를 비롯한 여러 증분적합도 지수들의 초기 형태를 제안하며, 본인들의 경험을 바탕으로 .90이라는 기준값을 제시하였다. 참고로 TLI의 경우 본래 신뢰도 지수로서 소개되었는데, 신뢰도 지수의 가이드라인 역시 다양한 경험과 직관을 바탕으로 합의된 임의적인 기준에 해당한다(Helmstadter, 1964).

임의적으로 설정된 기준의 경우 통계적 기반이 부족하다는 한계를 갖고 있기 때문에(McDonald and Marsh, 1990; Marsh & Balla, 1994; Marsh, Hau, & Grayson, 2005), 1980~1990년대의 다양한 연구들은 시뮬레이션을 이용하여 가이드라인을 설정하고자 시도하였다. 현재 적합도 지수 평가 과정에서 대표적인 참고문헌으로 사용되고 있는 Hu와 Bentler(1999)는 잘못 설정된 모형(misspecified model)을 이용한 시뮬레이션을 통해 가이드라인을 제시하였다. 구체적으로 TLI, CFI는 .95 이상일 때, RMSEA는 06, SRMR은 .08 이하일 때 2종 오류, 즉 잘못 설정된 모형을 기각하지 못하는 확률이 낮아지는 시뮬레이션 결과를 보고하였다. 그러나 해당 가이드라인 역시 모든 모형에 적용 가능한 통일된 기준이라 할 수는 없다. 일반적으로 시뮬레이션 연구의 경우 특정 조건을 지닌 모형을 이용하기 때문에 모든 조건에 대하여 일반화된 규칙으로 사용하기 어렵기 때문이다(Kline, 2016; Sivo, Fan, Wittal, & Willse, 2006). 연구자의 모형이 Hu와 Bentler(1999)의 시뮬레이션에서 사용된 모형과 차이가 클수록 해당 가이드라인의 정확성과 유용성은 낮아질 수밖에 없다.

이와 같은 한계에도 불구하고 대부분의 연구자들이 몇몇 참고문헌(예, Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2016)에서 제공한 적합도 지수 가이드라인을 절대적인 규칙으로 사용하는 이유는 어떠한 평가 대상에 대하여 하나의 정해진 규칙이 있으면 판단하기에 더 용이하기 때문이다(Marsh, Hau, & Wen, 2004). 모든 조건에 대하여 동일하게 적용되는 기준이 있을 경우 연구자는 적합도를 평가할 때 자신의 모형이 어떠한 특징과 조건을 갖고 있는지 고민할 필요가 없는 것이다. 그러나 모형의 종류와 상관없이 모든 상황에 통용될 수 있는 단 하나의 규칙은 존재할 수 없으며(Fan et al., 1999), Hu와 Bentler(1999)를 포함한 여러 연구자는 가이드라인을 엄격하게 지키는 방향보다는 단순히 해석에 도움이 될 수 있는 보조적인 역할로 사용할 것을 강조하였다(Lai & Green, 2016).

모형 조건에 따른 기준값의 조정

적합도 효과크기 지수의 가이드라인을 모든 상황에 적용 가능한 절대적인 규칙으로 사용할 수 없는 또 다른 이유는, 모형의 조건에 따라 효과크기 지수의 추정치에 편향이 생길 수 있기 때문이다. CFI, RMSEA, SRMR을 포함

한 거의 모든 적합도 효과크기 지수들은 표본크기나 변수의 개수 등과 같은 모형 조건에 따라 모수에 대한 편향된 추정치를 제공할 수 있다. 예를 들어, 자료의 표본크기는 적합도 지수가 탄생한 배경과 직접적으로 연결되는 요인으로, 적합도 지수는 본래 χ^2검정과는 다르게 표본크기의 영향을 받지 않을 것이라는 믿음 아래 발전되었다(Bentler & Bonett, 1980; Jöreskog & Sörbom, 1981, 1984). 그러나 적합도 지수의 사용이 대중화되고 적합도 지수 추정치의 편향에 관한 연구들이 증가하면서 적합도 지수 역시 표본크기의 영향을 받는다는 결과들이 제시되었다. 대표적인 초기 적합도 지수 GFI(goodness of fit index)와 AGFI(adjusted goodness of fit index)의 경우 발전 당시 표본크기로부터 독립적이라고 가정되었으나(Jöreskog & Sörbom, 1984), 이후 진행된 시뮬레이션 연구(Anderson & Gerbing, 1984)는 두 지수의 수리적 계산과정 안에 표본크기가 반영되지 않을 뿐 분포 자체는 표본크기의 영향을 받으므로, 표본크기가 증가할수록 GFI와 AGFI도 함께 증가함을 밝혔다. 이는 곧 표본크기가 클 경우 좋은 수준의 적합도를 나타내기 위하여 GFI와 AGFI의 가이드라인이 기준값보다 더 높은 값으로 설정되어야 함을 의미한다.

나아가, Marsh 등(1988)은 χ^2/df, GFI, AGFI, NFI, TLI 등을 포함한 30개 가량의 초기 적합도 지수들의 표본크기에 대한 편향을 확인한 결과, TLI만이 상대적으로 표본크기에 독립적임을 확인하였다. Marsh와 Balla(1994)는 CFI와 동일한 모수를 추정(Goffin, 1993)하는 RNI (relative noncentrality index; McDonald, 1989)가 비교적 표본크기에 독립적임을 발견하였으며, Fan 등(1999)은 TLI, CFI, RMSEA가 상대적으로 표본크기로 인해 발생하는 편향이 작음을 확인하고 해당 지수들을 중점적으로 이용할 것을 추천하였다. 다만 Curran 등(2003)의 경우 200 이하의 표본크기에서는 RMSEA 점추정치가 과대 추정되는 경향이 있음을 밝혀, 작은 표본크기의 자료에 대하여 모형을 추정하는 연구자들은 RMSEA 값을 해석할 때 기존의 가이드라인이 상대적으로 엄격한 기준이 될 수 있다.

적합도 지수의 편향에 대한 초기 연구가 표본크기를 중심으로 진행되었다면, 모형의 크기(model size)가 미치는 영향에 관한 연구는 90년대 중반까지 상대적으로 적은 비중을 차지하였다(Ding, Velicer, & Harlow, 1995). 확인적 요인분석 모형에서 전체 지표변수의 개수, 요인 당 지표변수의 개수, 혹은 자유도 등으로 정의(Shi, Lee, & Terry, 2017)되는 모형 크기의 효과란, 특히 작거나 중간 정도의 표본크기에서 모형의 크기가 증가할수록 T_{ML}이 정적으로 편향되며 1종 오류가 증가하는 현상을 의미한다(Herzog, Boomsma, & Reinecke, 2007). T_{ML}에 편향이 발생함에 따라 T_{ML}을 이용해 추정되는 TLI, CFI, RMSEA에도 모형 크기가 영향을 미치는데, 구체적으로 작은 표본크기(예, 200 이하)의 조건 아래에서 지표변수의 개수가 증가할수록 올바르게 설정된 모형임에도 불구하고 TLI와 CFI는 좋지 않은 적합도를, RMSEA는 반대로 좋은 적합도를 보여준다(Kenny & McCoach, 2003). 이처럼 모형 크기 조건이 다르게 작동하는 이유 중 하나는 RMSEA가 TLI나 CFI와 다르게 추정과정에서 모형과 자료 간 차이를 나타내는 χ^2-df를 df로 나누어주기 때문이다. 이와 같은 과정은 모형의 복잡성에 대한 페널티를 부여하는데, 일반적으로 지표변수의 개수가 증가할수록 자유도도 함께 증가함에 따라(Shi, Lee, &

Terry, 2017) RMSEA는 작은 값을 나타내게 된다. 반대로 자유도가 감소할 경우 RMSEA는 정적으로 편향되어 좋지 않은 적합도를 나타내기 때문에 기존의 RMSEA 가이드라인(예, Browne & Cudeck, 1993)이 과도하게 엄격한 기준이 될 수 있다(Kenny, Kaniskan, & McCoach, 2015). 이와 같은 연구결과들은 모형 크기의 효과로 인하여 적합도 지수의 편향이 발생할 경우 연구자는 일반적인 가이드라인보다 다소 조정된 값을 기준으로 적합도의 효과크기를 해석할 필요가 있음(Moshagen, 2012)을 시사한다. 나아가 현재 범용적으로 사용되고 있는 가이드라인(예, Hu & Bentler, 1999)을 근거로 적합도의 좋고 나쁨을 절대적으로 평가하는 것은 적절한 평가 방식이 아닐 수도 있음을 함의한다.

앞에서 언급된 연구결과들의 경우 애초에 모형 크기의 영향을 받는 T_{ML}을 통해 추정된 적합도 지수 위주로 제시된 반면, SRMR은 잔차 행렬을 이용해 추정되기 때문에 T_{ML}의 편향이 SRMR의 편향으로 이어진다고 보기 어렵다. 그러나 실제로 모형의 df가 감소함에 따라 SRMR도 함께 감소하거나(Taasoobshirazi & Wang, 2016), 작은 표본크기에서 SRMR이 좋지 않은 적합도를 나타내는 경향이 지표변수의 개수가 증가함에 따라 더욱 강해지는(Ximénez, Maydeu-Olivares, Shi, & Revuelta, 2022) 등의 연구결과들을 확인하였을 때, SRMR 역시 모형 크기의 영향으로부터 자유롭지 않음을 알 수 있다.

적합도 효과크기 지수의 연속적 해석

효과크기 가이드라인의 올바른 예시

적합도 효과크기 지수들의 해석을 위한 가이드라인은 그 기준값 자체가 경험적인 배경을 바탕으로 설정됨에 따라 동일한 지수에 대하여 학자마다 조금씩 다른 값을 제안한다. 그럼에도 불구하고 대부분의 가이드라인은 각 효과크기 지수에 대하여 어느 정도 유사한 기준을 지니며, 표 2는 이 가운데 현재 모형 적합도의 보고에 가장 대중적으로 사용되고 있는 기준값들을 제시하고 있다.

적합도 효과크기 지수는 결과를 연속적으로 해석한다는 점에서 χ^2검정과 매우 큰 평가 방법의 차이를 보인다. 이분법적인 해석을 도출하는 통계적 검정과 달리 효과크기 지수는 결과의 해석에 연속성이 반영되어야 한다. 안타깝게도 대다수의 연구 상황에서 적합도의 효과크기는 이분법적으로 해석되고 있는데(Gomer et al., 2019; Lai & Green, 2016), 그 대표적인 원인 중 하나는 가이드라인을 처음 제시하는 과정 자체에서 기준값이 채택 가능한 모형 적합도의 최저 기준, 또는 절단점으로서 제안되기 때문이다. 예를 들어, 적합도 지수를 해석하는 가이드라인으로 가장 많이 사용되고 있는 Browne과 Cudeck(1993)의 연구(RMSEA ≤ .05) 및 Hu와 Bentler(1999)의 연구(CFI≥ .95, RMSEA≤ .06, SRMR≤ .08)의 경우, 모두 기준값을 제시하는 과정에서 적합도 지수가 특정 값 '이상' 또는 '이하' 등의 기준을 제시하였다. 이후, 모형 적합도에 대한 리뷰 연구를 진행한 Schermelleh-Engel 등(2003) 및 Hooper 등(2008) 역시 마찬가지로 RMSEA≤ .10이 적절한 적합도를 나타낸다는 이분법적 기준을 제시하였는데, 이와 같은 가이드라인을 이용해 모형을 평가할 경우 RMSEA=.11과 같이 .10의 기준값을 근소하게(marginally) 달성하지 못하는 모형을 단순히 나쁜 모형으로 해석하는 상황이 발생할 수 있다. 그러나 RMSEA의 가이드

표 2. 대표적으로 사용되고 있는 적합도 효과크기 지수의 가이드라인

적합도 효과크기 지수 추정치	가이드라인		
	기준값	해석	출처
$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}}$	.10	good fit	Steiger(1989)
	.05	very good fit	
	.01	outstanding fit	
	.06	good fit	Hu와 Bentler(1999)
	.08	reasonable model	Browne과 Cudeck(1993)
	.05	close fit	
$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$	.08	good fit	Hu와 Bentler(1999)
	.10	acceptable fit	Schermelleh-Engel 등 (2003)
	.05	good fit	
$\widehat{SRMR}_u = \hat{k}^{-1}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}), 0)}{t}}$	$.10 \times \overline{R^2}$ [4]	acceptable fit	Shi 등(2018)
	$.05 \times \overline{R^2}$	close fit	
$CFI = 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)}$	.90	.90 이하의 모형은 발전 필요	Bentler와 Bonett(1980)
	.95	good fit	Hu와 Bentler(1999)
	.95	acceptable fit	Schermelleh-Engel 등 (2003)
	.97	good fit	
$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T}^*_{ML}}{N-1}\right)^{1/2}$	.82	large effect size	Gomer 등(2019)
	.60	medium effect size	
	.42	small effect size	

라인에서 제시하는 .10, .08, .05와 같은 값들은 연속선 상의 한 지점에 해당할 뿐이며, 이 값을 가까스로 만족했다고 하여 연구모형이 적합하다고 평가하고 미세하게 만족하지 못했다고 해서 부적합하다고 평가하는 것은 적절하지 않다.

효과크기 지수의 연속성을 직관적으로 이해하기 위하여 그림 2와 같은 가상의 연속선 상에 표시된 각 지수의 가이드라인과 효과의 크기 예시를 제공하였다. 그림 2는 최악의 적합을 의미하는 기저모형부터 완전 적합을 의미하는 포화모형까지 가상의 연속선 상에서 근소하게 기준값을 만족하지 못하는 값(marginal value)의 상대적 위치를 나타낸다. 포화모형은 자료에 대한 모형의 완전 적합을 나타내는 모형으로, 기저모형은 자료를 거의 설명하지 못하는 수준의 적합을 나타내는 모형으로 해석할 수 있다. 적합도 효과크기의 정의에 따라

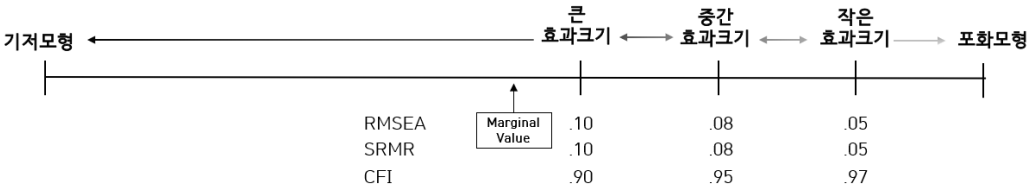


그림 2. 연속선 상에서 근소하게 기준을 만족하지 못하는 적합도 수준의 상대적 위치

효과크기 값이 작을수록 모형이 자료를 잘 설명하고 있음을 고려하였을 때, 적합도의 수준이 완전 적합에 가까워질수록 추정된 효과크기는 작아진다. 만일 연구모형의 RMSEA 값이 .11일 경우, 이는 큰 효과크기의 기준인 .10을 초과하게 된다. 그러나 그림 2를 바탕으로 해당 모형을 해석할 때 모형의 효과크기는 큰 효과크기에서 멀리 떨어져 있는 것도 아니며, .10을 초과했다고 해서 기저모형 쪽에 매우 가까이 위치한 것도 아니다.

이와 같은 연속적인 관점에서의 해석 방식을 가이드라인에 더욱 잘 반영하기 위해서는 가이드라인의 기준값을 적합도의 마지노선으로 인식하는 것이 아닌, 효과크기의 수준을 나타내는 연속선 상의 한 지점으로 이해하고 사용할 필요가 있다. 가이드라인 자체에서 'RMSEA=.10은 큰 효과크기를 나타낸다'고 말하는 해석방식을 제시할 경우 연구자는 .11이라는 RMSEA 값을 가진 모형이 .10의 값을 가진 모형과 설명력의 측면에서 멀리 떨어져 있지 않다고 해석할 수 있다. 반대로 특정 적합도 효과크기 지수가 기준값을 반드시 만족해야 하는 것처럼 가이드라인을 제시할 경우(예, χ^2검정이 기각되고 $SRMR > .08$일 경우 poor fit으로 정의[Asparouhov&Muthén, 2018]), 효과크기의 연속성에 대해 제대로 이해하고 있지 못한 연구자는 이를 절대적 기준으로 받아들이게 되며, 이는 가이드라인의 잘못된 사용방식으로 이어진다(Chen et al., 2008; Kenny, 2015; Kline, 2016; Markland, 2007; West et al., 2012).

적합도 효과크기 지수가 이분법적으로 해석되는 또 다른 원인으로는 기존의 가이드라인들이 기준값에 대한 해석을 하는 과정에서 '좋은 적합도(good fit)', '충분한 적합도(adequate fit)', 또는 '허용 가능한 적합도(acceptable fit)'등과 같은 표현을 이용하기 때문이다. 이와 같은 해석은 적합도의 좋고 나쁨, 혹은 적절성을 질적으로 판단하는 방식으로, 이는 곧 적합도에 대한 이분법적 평가로 이어지게 된다. 하지만 적합도 효과크기 지수들을 이용해 적합도를 평가한다는 것은 모형과 자료 간의 차이를 양적으로 해석하는 것을 의미한다. 일반적으로 대다수의 효과크기 지수들은 연속선 상에서 크기에 대한 정의를 내리기 위해 관습적으로 '작은', '중간 정도의', '큰' 등과 같은 표현들을 이용하였으며(Cohen, 1988), Gomer 등(2019)은 실제로 이와 같은 방식을 이용하여 적합도 효과크기 지수 ϵ를 해석하였다.

'좋은(Good)', 혹은 '적절한(adequate)'과 같은 질적 판단이 아닌 '작은(small)', '중간의(medium)', '큰(large)'과 같은 양적 평가를 기반으로 이루어지는 해석의 경우, 기존의 표현과

4) $\overline{R^2}$=내생변수들의 평균적인 R^2

표 3. 연속성을 반영한 효과크기 지수 가이드라인

적합도 효과크기 지수		효과크기 기준값		
		작은 효과크기	중간 정도의 효과크기	큰 효과크기
RMSEA		.05	.08	.10
SRMR	$\widehat{SRMR}$	.05	.08	.10
	$\widehat{SRMR}_u$	$.05\times\overline{R^2}$	$.10\times\overline{R^2}$	-
ϵ		.42	.60	.82
CFI		.97	.95	.90

주. RMSEA, SRMR, ϵ은 모수이지만 CFI는 추정치에 해당. CFI의 모수 Δ의 경우 다른 지수들과 달리 일반적으로 모수보다 추정치의 형태로 보고되기 때문에 이와 같이 표기함.

비교하여 모형과 자료 간 차이를 한층 더 연속적인 관점에서 나타낸다. 표 3은 각 효과크기 지수가 연속선 상에서 작은 효과크기, 중간 효과크기, 큰 효과크기를 나타내는 지점을 제시한 가이드라인으로, 기준값은 표 2에서 언급된 다수의 대표적인 가이드라인을 통합한 결과이다. 앞에서 언급한 바와 같이 완전 적합 검정의 경우 검정 결과에 대한 효과크기 값이 작을수록 높은 수준의 적합도를 의미하며, 각 지수의 적합도에 대한 정의에 따라 높은 수준의 적합도를 가진 모형에 대하여 RMSEA와 SRMR, ϵ는 작은 값을, CFI는 큰 값을 이용해 작은 효과크기를 나타내게 된다. 표 3을 이용하여 적합도의 효과크기 지수를 해석할 경우 연구자는 앞서 언급된 기준값의 절단적 해석 문제를 해결함과 더불어 적합도의 수준을 양적으로 표현하는 것이 가능하다.

근소하게 기준을 만족하지 못하는 적합도(marginal fit)에 대한 평가

효과크기 지수를 해석하는 과정에서 연속성의 중요성은 모형에 대한 실질적인 평가를 내릴 때 더욱 강조된다. 효과크기 지수를 연속적으로 해석하는지 아닌지에 따라 모형의 유용성에 대한 평가가 달라지기 때문이다.

표 4에 제시된 연구들(Heller et al., 2021; Thomsen & Lessing, 2020; Tyler et al., 2020; Yang & McGinley, 2021)은 적합도 평가 과정에서 추정된 효과크기 지수 값이 그 기준을 근소하게 만족하지 못할 경우(marginal fit) 적합도에 대해 질적으로 부정적인 평가를 제시하며 모형을 수정하거나 배제한 사례이다. 이 가운데 Thomsen과 Lessing(2020)의 경우 적합도 효과크기 지수 값을 추정하기 전 이미 가이드라인을 충족하는 모형은 배제할 것을 연구에 명시하였다. 이와 같은 결과들은 적합도 평가 과정에서 가이드라인을 절대적으로 충족해야 하는 이분법적 규칙처럼 사용하는 관행이 계속 이어지고 있음을 가리킨다.

앞서 언급한 바와 같이, 적합도 효과크기 지수가 연속적으로 해석되지 않는 주요 원인 가운데 하나는 기존의 가이드라인들이 제시되는 과정 자체에 연속성이 제대로 반영되지 않았기 때문이다. 만일 연구자가 표 3과 같이

표 4. 근소하게 기준을 충족하지 못한 모형에 대한 해석

출처	효과크기 지수	값	평가에 사용한 가이드라인	해석
Heller 등(2021)	CFI	.97	.95	mixed fit
	RMSEA	.07	.06	
	SRMR	.08	.08	
Thomsen과 Lessing(2020)	CFI	.91	.95	without acceptable fit
	RMSEA	.08	.08	
	SRMR	.09	.08	
Tyler 등(2020)	TLI	.89	.90, .95	mediocre fit
	CFI	.91	.90, .95	
	RMSEA	.06	.05, .08	
Yang과 McGinley(2021)	CFI	.88-.89	.90, .95	poor model fit
	RMSEA	.08	.06, .08	
	SRMR	.08-.11	.08, .10	

주. 회색 영역은 연구자가 이분법적 기준을 바탕으로 문제가 있다고 보고한 효과크기 지수와 값

연속선 상에서 효과크기의 해석을 제시하는 가이드라인을 사용할 경우, CFI=.89의 모형은 CFI=.90의 모형들과 동일하게 단순히 큰 효과크기를 나타내며 비슷한 수준의 유용성을 지닌 모형으로 평가될 수 있다. 기존의 이분법적 가이드라인들이 CFI=.90의 모형은 통과시키면서 CFI=.89의 모형은 적합도의 수준이 낮다는 이유로 배제하였다면, 표 3의 가이드라인은 CFI=.89의 모형과 CFI=.90의 모형을 동일선상에서 사용할 수 있는 근거를 제공한다.

나아가, Tyler 등(2020) 및 Thomsen과 Lessing (2020)의 경우 근소하게 가이드라인을 만족하지 못하는 모형에 대해 오차 간 상관을 포함하거나 변수를 통제하는 등의 기법을 사용하여 적합도를 임의적으로 올렸다. 그러나 효과크기 지수의 가이드라인을 특정 모형을 수정하거나 탈락시키는 유일한 근거로 사용하는 것은 적절하지 않다(Bagozzi & Yi, 1988; Hu & Bentler, 1998; Kenny et al., 2015; Marsh & Balla, 1994; McDonald & Ho, 2002). 반대로 Heller 등(2021) 및 Yang과 McGinley(2021)의 경우 앞의 연구들과 동일하게 가이드라인을 만족하지 못하는 모형을 수정하였으나, 그 과정에서 기준값 외에 모형 자체의 문제(예, Heywood case) 등을 함께 근거로 제시하였다. 이처럼 근소하게 기준값을 만족하지 못하는 효과크기 지수가 산출되었을 때, 해당 모형이 유용하지 않다고 주장하기 위해서는 가이드라인의 기준값 이외에 또 다른 확실한 근거가 필요하다(Marsh & Hau, 1996; Schreiber et al., 2006). 애초에 효과크기의 가이드라인은 모형의 유용성에 대한 절대적 기준이 아니기 때문

이다.

이와 반대로, 실제로 적합도 지수를 연속적으로 해석하여 모형을 평가하는 사례도 다수 존재하는데, 해당 연구들의 경우 기준값이 시뮬레이션으로부터 결정된 값이기 기준값에서 떨어졌다고 해서 무조건 제안된 모형을 기각해서는 안됨을 설명하였으며(Gerpott et al., 2021, Marsh et al., 2004), 이와 같은 기준들을 엄격한 기준이 아닌 가이드라인 정도로 사용할 것을 명시하였다(Williams et al., 2021). 또한, 다양한 연구들에서 적합도 지수를 연속적으로 사용함에 따라 기준값을 근소하게 만족하지 못하는 모형임에도 이를 채택하는 사례들이 제시되었으며(Johnson et al., 2020; Rau et al., 2021), 심지어 RMSEA= .114가 보고되었음에도 다른 요인들을 고려해 해당 모형을 채택하는 등의 (Jansen et al., 2021) 결과도 존재하였다. 특히 CFI =.88 ~ .89 사이의 값을 나타내는 모형임에도 이를 적절한 모형으로 보고하는 연구들도 다수 존재하였는데(Lin et al., 2020; Rojas et al., 2020; Rosen et al., 2020; Thompson & Bergeron, 2020), 특히 Rosen 등(2020)은 .89라는 CFI 값이 다른 지수들과 비교해 상대적으로 조금 작은 것일 뿐이며 West 등(2012) 또는 Williams, O'Boyle과 Yu(2020)이 제안한 바와 같이 적합도 지수를 자동적으로 모형을 기각하는데 적용하는 규칙으로 사용하지 않을 것을 권고하였다.

적합도 효과크기 지수를 연속적으로 해석하는 사례는 국내에서도 적지 않게 확인할 수 있는데, 이들은 CFI 혹은 TLI의 값이 .88에서 .89 사이로 추정됨에도 해당 모형을 적합한 모형으로 보고하였다(김진숙 & 권석만, 2010; 조은영 & 임성문, 2012; 연수진 & 서수균, 2013). 이와 같은 연구 결과들은 현재 경험연구를 수행하는 연구자들 가운데 적합도 지수를 연속적인 지수로 인식해 모형을 평가하는 연구자와 이를 이용해 모형을 이분법적으로 '선정'하는 연구자들이 섞여 있음을 나타낸다. 그러나 연구자가 적합도 효과크기 지수를 바탕으로 선정한 모형이 반드시 가장 좋은 모형인 것은 아니다. 해당 자료를 더 잘 설명하는 모형은 충분히 존재할 수 있으며, 이는 즉 어떠한 모형을 선택하는 과정에서 모형의 옳고 그름에 대한 이분법적 의미를 반영하는 것 자체가 위험한 일임을 의미한다.

적합도 효과크기 지수와 가이드라인의 목적 및 연속성의 특징을 고려하였을 때 연구자는 추정된 효과크기 지수가 가이드라인의 기준값에 근접하게 되면 적합도 자체에 큰 문제가 없음을 명시하고 모형을 사용할 수 있다. 물론, 이와 같은 주장은 heywood case와 같은 모형 자체의 문제가 존재하지 않는다는 조건 하에 성립할 수 있다. 또는, 여전히 적합도 효과크기 지수가 기준값을 만족하지 못하면 해당 모형이 설명력의 측면에서 불완전하다고 생각되어 모형을 수정하고 적합도를 끌어올릴 수도 있다. 하지만 수정 지수를 이용하여 모형에 자유모수를 추가하는 행위에는 확실한 이론적 근거가 바탕이 되어야 한다(Marsh & Hau, 1996; Schreiber et al., 2006). 또한, 일반적으로 적합도를 향상하기 위해 모형을 수정하는 관행이 이미 여러 연구에서 모형의 타당성 및 일반화의 문제 등을 바탕으로 비판(Boomsma, 2000; MacCallum, 1986; MacCallum et al., 1992)받아 왔음을 고려하였을 때, 적합도에 큰 문제가 없음에도 불구하고 오로지 가이드라인을 만족하기 위해 모형을 수정하는 것은 상당히 위험한 행위임을 알 수 있다.

적합도의 효과크기를 확인한다는 것은 적합

도의 실질적 유의성을 확인한다는 의미이며, α와 같은 이분법적 절단점을 이용해 평가하는 통계적 유의성과는 달리 연속성을 바탕으로 효과의 크기를 파악하는 것이다. 표 4에서 제시된 바와 같이, 적합도의 수준을 확인하기 위해 효과크기 지수를 사용함에도 불구하고 이를 이분법적으로 해석하며 가이드라인을 만족하기 위해 모형을 수정하는 것은 효과크기 지수의 본래의 사용 목적에 맞지 않으며, 자료에 대한 충분한 설명력을 지니고 있음에도 불구하고 모형을 배제해 버리는 비효율적인 평가에 해당한다.

결론 및 논의

사회과학 영역에서 구조방정식 모형의 사용이 활발해짐에 따라 대다수 연구자는 가장 대표적인 적합도 평가 도구인 적합도 지수를 중점적으로 활용하여 모형의 유용성을 판단한다. 현재 적합도의 평가 관행은 추정된 적합도 지수가 Hu와 Bentler(1999), Browne과 Cudeck (1993) 등의 가이드라인에서 제시하는 기준값을 만족하면 모형을 통과시키고, 그렇지 못하면 모형을 배제하는 방식이 만연하다(Heene et al., 2012). 심지어 연구모형이 기준값을 조금이라도 만족하지 못할 경우, 오차 간 상관을 임의로 포함하는 등의 기법을 이용하여 어떻게든 그 기준을 충족하고자 한다. 이와 같은 평가 방식은 적합도의 실질적 유의성을 확인하는 과정 자체에 대한 이해가 부족함에 따라 나타나는 문제이다. 본 연구는 이를 해결하기 위하여 효과크기의 관점에서 적합도를 평가하는 다양한 지수들을 소개하고, 연구자들이 적합도를 연속적으로 해석하는데 실질적인 도움이 될 수 있는 효과크기 가이드라인의 예시 및 사용 방법에 대해 논의하였다.

모형 평가의 전반적인 과정에 대한 이해를 돕기 위하여 본 연구에서는 우선 적합도 평가의 첫 번째 단계인 χ^2검정을 간략하게 소개하고, 실제 연구에서 빈번하게 기각되는 χ^2검정 결과가 어떠한 의미를 갖는지에 대해 논하였다. 표본크기를 비롯한 몇몇 한계점들로 인하여 χ^2검정 결과는 현재 형식적으로만 보고되고 있으나 그 형식 자체도 제대로 지켜지고 있지 않으며, 검정의 결과를 완전 적합 영가설에 대하여 해석하는 사례는 거의 찾을 수 없다. 본문에서도 강조했듯이 χ^2검정은 적합도의 통계적 유의성을 평가하는 거의 유일한 도구로서 매우 중요한 의의를 지닌다. 검정의 p값조차 제대로 보고하지 않고 표본크기를 근거로 χ^2검정 자체를 배제하기보다, 검정이 기각됨에 따라 모형이 자료에 완벽하게 적합하지는 않으며 완전 적합에서 얼마나 떨어져 있는지에 대해서는 실질적 유의성을 통해 파악한다고 해석하는 것이 적절한 χ^2검정 결과의 해석이라 볼 수 있다.

다음으로, 본 연구에서는 적합도의 실질적 유의성을 평가할 수 있는 다양한 지수에 대한 소개가 이루어졌다. 과거부터 오랜 기간 사용되고 있는 전통적 지수부터 최근 새롭게 발전한 지수까지 다양하고 핵심적인 종류의 평가지수를 이용해 적합도의 실질적 유의성을 확인하는 것이 가능하다. 이와 같은 지수들을 효과크기로 이용하는 과정에서 첫 번째로 주의를 요하는 개념은 일반적인 검정 결과에 대한 효과크기와 달리 적합도에 대한 효과크기의 경우 완전 적합검정 결과에 대한 것이기 때문에 효과크기가 작을수록 모형과 자료가

서로 합치함을 나타낸다는 것이다. 적합도의 수준은 효과크기와 부적 관계를 이루고 있으며, 효과크기 지수 값이 작을수록 연구자는 모형 적합도의 수준이 높다고 주장할 수 있다.

적합도의 실질적 유의성을 평가하는 과정에서 주의해야 하는 두 번째 요점이자 적합도 효과크기 지수의 목적을 달성하기 위하여 가장 중점적으로 고려해야 하는 요소는 추정된 지수의 해석과정에서 연속성을 반영하는 것이다. 실질적 유의성의 경우 연속선 상에서 모형과 자료 간의 차이를 확인하는 것이기 때문에 추정된 적합도 효과크기 지수 값은 연속적으로 해석되어야 하며(Hu & Bentler, 1998), 이는 곧 효과크기 지수의 가이드라인에서 제시되는 기준값들을 데드라인, 또는 절단 값이 아닌 말 그대로의 가이드라인 정도로 사용해야 함을 의미한다(Marsh, Hau & Wen, 2004). 본 연구는 기존의 효과크기 지수 가이드라인들을 정리하여 연속성이 반영된 새로운 가이드라인의 예시를 제공하였으며, 이를 바탕으로 가이드라인의 기준값을 근소하게 만족하지 못하는 모형을 사용하는 것에 논리적으로 문제가 없음을 설명하기 위해 노력하였다.

본 연구는 모형 적합도를 평가하고 해석하는 과정에서 어려움을 겪는 내용 영역 연구자들에게 실용적으로 도움이 되고자 하는 목적 아래 적합도를 평가할 수 있는 다양한 지수를 재소개하고 연속적인 해석을 위한 가이드라인의 사용 방식을 제안하였다. 그럼에도 불구하고 실제로 적합도 효과크기 지수를 사용하는 과정에는 여러 종류의 문제들이 복합적으로 존재한다. 본 연구에서도 설명하였듯이 적합도 효과크기 지수는 모형 조건의 영향을 받으며, 추정된 지수 값이 기준을 만족하지 못하는 것이 실제로 모형 설정 과정에서의 심각한 문제에 해당하는지 아닌지에 대한 명확한 이유를 알기 위해서는 잔차 행렬 등을 통해 모형을 복합적으로 진단하는 과정이 요구된다(McDonald & Ho, 2002). 나아가, 효과크기를 해석하는 과정에는 효과크기 모수에 대한 구간 추정치인 신뢰구간을 함께 확인하여 효과크기의 정확성을 평가하는 단계도 필요하다(Maydeu-Olivares, 2017; Maydeu-Olivares & Shi, 2017). RMSEA의 경우 현재 대부분의 통계 프로그램에서 신뢰구간을 함께 제공하고 있으나, 그 외의 지수들은 구간 추정치의 정보가 디폴트로 제공되지 않는다. 이에 따라, 본 연구에서 제시한 다양한 효과크기 지수를 신뢰구간의 관점에서 어떻게 추정하고 사용할 수 있는가에 대한 확장된 논의가 필요할 수 있다.

적합도 지수가 처음 발전된 이래 이를 이용하여 모형을 평가하는 행위 자체에 대한 근본적인 한계에 대한 연구들 역시 지속적으로 제기되어왔다. 기본적으로 적합도 지수는 점추정치 형식으로 제공되는데, 그에 따른 표집오차의 문제는 함께 수반될 수 밖에 없으며 그에 따라 동일한 모형에 대해 표본이 바뀔 때마다 추정치가 다른 값을 제공하게 된다(Kline, 2016). 또한, 적합도 지수는 모형을 평가하는 다양한 요소 중 하나일 뿐이며(Marsh & Balla, 1994), 이외에도 모형의 타당성을 주장하기 위해서는 추정치의 해석 가능성이나 모형의 복잡성 등 다양한 요인이 고려되어야 한다(Hu & Bentler, 1998).

본 연구의 주요 목적은 현재 구조방정식 모형의 평가 과정에 사용되는 주요 적합도 지수를 이용해 적합도의 실질적 유의성을 해석한다는 것이 모형 평가의 측면에서 어떠한 의미

를 지니는지에 대해 논의하는 것이었다. 이와 같은 논의를 바탕으로 실제로 모형 적합도를 평가하는 과정에 있는 연구자, 특히 적합도 지수 값이 Hu와 Bentler(1999)나 Browne과 Cudeck(1993) 등과 같은 가이드라인에서 제공하는 기준에 근접한 결과를 가진 연구자들이 해당 모형을 배제하지 않을 수 있는 근거를 제시할 수 있을 것으로 기대된다. 연구자는 적합도 평가의 가장 큰 목적이 가이드라인에서 제시하는 기준을 충족하는 것이 아님을 인지하고, 이를 바탕으로 모형의 유용성에 대해 넓은 관점에서 효율적인 판단을 내려야 한다.

참고문헌

김진숙. & 권석만. (2010). 인지행동적 요인과 부부 불만족도 사이의 관계. 한국심리학회지: 일반, 29(2), 265-288.

연수진. & 서수균. (2013). 이성관계에서 안정애착이 갈등해결전략과 관계만족도에 미치는 영향: 자기효과와 상대방효과. 한국심리학회지: 일반, 32(2), 411-428.

조은영. & 임성문. (2012). 자아해석과 주관적 안녕감 및 우울간의 관계: 인지적 유연성, 자기개념 명확성의 매개효과와 자기복잡성의 조절효과. 한국심리학회지: 일반, 31(2), 493-519.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2), 155-173. https://doi.org/10.1007/BF02294170

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus.

Bagozzi, R., & Yi, Y. (1988). On the Evaluation of Structural Equation Models. *Journal of the Academy of Marketing Sciences*, 16(1), 74-94. http://dx.doi.org/10.1007/BF02723327

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815-824. https://doi.org/10.1016/j.paid.2006.09.018

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons. https://doi.org/10.1002/9781118619179

Boomsma, A. (2000). Reporting Analyses of Covariance Structure. *Structural Equation Modeling: A Multidisciplinary Journal,* 7(3), 461-483. https://doi.org/10.1207/S15328007SEM0703_6

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting Incremental Fit Indices for Nonnormality. *Multivariate Behavioral Research,* 49(5), 460-470. https://doi.org/10.1080/00273171.2014.933697

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park,

CA: Sage.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cut-Off Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods and Research,* 36(4), 462-494. https://doi.org/10.1177/0049124108314720

Cheng, C., & Wu, H. (2017). Confidence Intervals of Fit Indexes by Inverting a Bootstrap Test. *Structural Equation Modeling: A Multidisciplinary Journal,* 24(6), 870-880. https://doi.org/10.1080/10705511.2017.1333432

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite Sampling Properties of the Point Estimates and Confidence Intervals of the RMSEA. *Sociological Methods & Research*, 32(2), 208-252. https://doi.org/10.1177/0049124103256130

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of Estimation Methods, Number of Indicators per Factor, and Improper Solutions on Structural Equation Modeling Fit Indices. *Structural Equation Modeling: A Multidisciplinary Journal,* 2(2), 119-143. https://doi.org/10.1080/10705519509540000

Fan, X. (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin. *The Journal of Educational Research,* 94(5), 275-282. http://www.jstor.org/stable/27542335

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal,* 6(1), 56-83. https://doi.org/10.1080/10705519909540119

Gerpott, F. H., Rivkin, W., & Unger, D. (2021). Stop and Go, Where is My Flow? How and When Daily Aversive Morning Commutes are Negatively Related to Employees' Motivational States and Behavior at Work. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000899

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. *Multivariate Behavioral Research,* 28(2), 205-214. https://doi.org/10.1207/s15327906mbr2802_3

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences,* 42(5), 831-839. https://doi.org/10.1016/j.paid.2006.09.019

Gomer, B., Jiang, G., & Yuan, K.-H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(3), 371-389. https://doi.org/10.1080/10705511.2018.1545231

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Johns Hopkins University Press.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S.

(2007). Testing! Testing! One, two, three-Testing the theory in structural equation models! *Personality and Individual Differences,* 42(5), 841-850. https://doi.org/10.1016/j.paid.2006.10.001

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling: A Multidisciplinary Journal,* 19(1), 36-50. https://doi.org/10.1080/10705511.2012.634710

Heller, A. S., Stamatis, C. A., Puccetti, N. A., & Timpano, K. R. (2021). The distribution of daily affect distinguishes internalizing and externalizing spectra and subfactors. *Journal of Abnormal Psychology,* 130(4), 319-332. https://doi.org/10.1037/abn0000670

Helmstadter, G. C. (1964). Principles of psychological measurement. Appleton-Century-Crofts.

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal,* 14(3), 361-390. https://doi.org/10.1080/10705510701301602

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods,* 6(1), 53-60. https://doi.org/10.21427/D7CF7R

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Sage Publications, Inc.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal,* 6(1), 1-55. https://doi.org/10.1080/10705519909540118

Jansen, M., Becker, M., & Neumann, M. (2021). Dimensional comparison effects on (gendered) educational choices. *Journal of Educational Psychology,* 113(2), 330-350. http://dx.doi.org/10.1037/edu0000524

Johnson, A., Nelson, J. M., Tomaso, C. C., James, T., Espy, K. A., & Nelson, T. D. (2020). Preschool executive control predicts social information processing in early elementary school. *Journal of Applied Developmental Psychology.* 71, Article 101195. https://doi.org/10.1016/j.appdev.2020.101195

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods. Chicago: International Educational Services

Jöreskog, K. G., & Sörbom, D. (1984). *Advances in factor analysis and structural equation models.* Lanham: Rowman & Littlefield Publishers.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* 34(2, Pt.1), 183-202.

https://doi.org/10.1007/BF02289343

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal,* 10(3), 333-351. https://doi.org/10.1207/S15328007SEM1003_1

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research,* 44(3), 486-507. https://doi.org/10.1177/0049124114543236

Kenny, D. A. (2020). Measuring model fit. https://davidakenny.net/cm/fit.htm

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement,* 56(5), 746-759. https://doi.org/10.1177/0013164496056005002

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York, NY: The Guilford Press.

van Laar, S., & Braeken, J. (2022). Caught off Base: A Note on the Interpretation of Incremental Fit Indices. *Structural Equation Modeling: A Multidisciplinary Journal,* 29(6), 935-943. https://doi.org/10.1080/10705511.2022.2050730"

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal,* 26(5), 757-777. https://doi.org/10.1080/10705511.2018.1562351

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research,* 51(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

Lin, S.-H. (J.), Chang, C.-H. (D.), Lee, H. W., & Johnson, R. E. (2021). Positive family events facilitate effective leader behaviors at work: A within-individual investigation of family-work enrichment. *Journal of Applied Psychology,* 106(9), 1412-1434. https://doi.org/10.1037/apl0000827

MacCallum, R. C., Roznowski. M., Necowitz. L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin,* 111(3), 490-504. https://doi.org/10.1037/0033-2909.111.3.490.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin,* 100(1), 107-120. https://doi.org/10.1037/0033-2909.100.1.107

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods,* 1(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Maiti, S. S., & Mukherjee, B. N. (1991). Two new goodness-of-fit indices for covariance matrices with linear structures. *British Journal of Mathematical and Statistical Psychology,* 44(1), 153-180. https://doi.org/10.1111/j.2044-8317.1991.tb00953.x

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting

model fit in structural equation modelling. *Personality and Individual Differences,* 42(5), 851-858. https://doi.org/10.1016/j.paid.2006.09.023"

Marsh, H. W., & Balla, J. R. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity,* 28(2), 185-217. https://doi.org/10.1007/BF01102761

Marsh, H. W., & Hau, K.-T. (1996). Assessing Goodness of Fit: Is Parsimony Always Desirable? *The Journal of Experimental Education,* 64(4), 364-390. https://doi.org/10.1080/00220973.1996.10806604

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin,* 103(3), 391-410. https://doi.org/10.1037/0033-2909.103.3.391

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research,* 33(2), 181-220. https://doi.org/10.1207/s15327906mbr3302_1

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal,* 11(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275-340). Lawrence Erlbaum Associates Publishers.

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models: Standardized residual covariances and residual correlations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* 13(1), 23-30. https://doi.org/10.1027/1614-2241/a000129

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika,* 82(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7"

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: a Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling: A Multidisciplinary Journal,* 25(3), 389-402. https://doi.org/10.1080/10705511.2017.1389611

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods,* 7(1), 64-82. https://doi.org/10.1037/1082-989X.7.1.64

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin,* 107(2), 247-255. https://doi.org/10.1037/0033-2909.107.2.247

McDonald, R. P. (1989). An index of

goodness-of-fit based on noncentrality. *Journal of Classification*, 6(1), 97-103. https://doi.org/10.1007/BF01908590

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: *A Methodological Paradox. Philosophy of Science,* 34(2), 103-115. https://doi.org/10.1086/288135

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal,* 19(1), 86-98. https://doi.org/10.1080/10705511.2012.634724

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models. *Educational and Psychological Measurement,* 81(1), 110-130. https://doi.org/10.1177/0013164420926231

Rau, R., Carlson, E. N., Back, M. D., Barranti, M., Gebauer, J. E., Human, L. J., ... & Nestler, S. (2021). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology,* 120(3), 745-764. http://dx.doi.org/10.1037/pspp0000278

Rojas, N. M., Yoshikawa, H., & Melzi, G. (2020). Preschool teachers' use of discourse practices with Spanish-speaking dual language learners. *Journal of Applied Developmental Psychology.* 69, Article 101158. https://doi.org/10.1016/j.appdev.2020.101158

Rosen C.C., Dimotakis N., Cole M.S., Taylor S.G., Simon L.S., Smith T.A., Reina C.S. (2020). When challenges hinder: An investigation of when and how challenge stressors impact employee outcomes. *Journal of Applied Psychology.* 105(10), 1181-1206. https://doi: 10.1037/apl0000483.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research*, 8(2), 23-74.

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research,* 99(6), 323-337. https://doi.org/10.3200/JOER.99.6.323-338

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The Relationship Between the Standardized Root Mean Square Residual and Model Misspecification in Factor Analysis Models. *Multivariate Behavior Research,* 53(5), 676-694. https://doi.org/10.1080/00273171.2018.1476221.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement,* 79(2), 310-334. https://doi.org/10.1177/0013164418783530

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research,* 53(5),

676-694.
https://doi.org/10.1080/00273171.2018.1476221

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education,* 74(3), 267-288.
https://doi.org/10.3200/JEXE.74.3.267-288

Steiger, J. H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Systat, Inc.

Taasoobshirazi, G., & Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods,* 11(3), 31-40.

Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology,* 42(2), 233-239.
https://doi.org/10.1111/j.2044-8317.1989.tb00912.x

Tanaka, J. S. (1993). Multifaceted Conceptions of Fit in Structural Equation Models. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 10-39). Newbury Park, CA: Sage.

Thomsen, T., & Lessing, N. (2020). Children's emotion regulation repertoire and problem behavior: A latent cross-lagged panel study. *Journal of Applied Developmental Psychology,* 71, 101198.
https://doi.org/10.1016/j.appdev.2020.101198

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher,* 25(2), 26-30.
https://doi.org/10.2307/1176337

Thompson P. S., Bergeron D.M., Bolino M.C. (2020) No obligation? How gender influences the relationship between perceived organizational support and organizational citizenship behavior. *Journal of Applied Psychology*. 105(11):1338-1350.
https://doi: 10.1037/apl0000481

Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science*, 6(1), 100-116.

Tyler, C. P., Olsen, S. G., Geldhof, G. J., & Bowers, E. P. (2020). Critical consciousness in late adolescence: Understanding if, how, and why youth act. *Journal of Applied Developmental Psychology,* 70, 101165.
https://doi.org/10.1016/j.appdev.2020.101165

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). The Guilford Press.

Williams, A. L., Craske, M.G., Mineka, S., Zinbarg, R.E. (2021). Neuroticism and the longitudinal trajectories of anxiety and depressive symptoms in older adolescents. *Journal of Abnormal Psychology*. 130(2), 126-140.
https://doi.org/10.1037/abn0000638.

Williams, L. J., O'Boyle, E., & Yu, J. (2020). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives.

Organizational Research Methods. 23, 6-29. http://dx.doi.org/10.1177/1094428117736137

Wilkerson, M., & Olson, M. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology: Interdisciplinary and Applied,* 131(6), 627-631. https://doi.org/10.1080/00223989709603844

Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022). Assessing Cutoff Values of SEM Fit Indices: Advantages of the Unbiased SRMR Index and Its Cutoff Criterion Based on Communality. *Structural Equation Modeling: A Multidisciplinary Journal,* 29(3), 368-380. https://doi.org/10.1080/10705511.2021.1992596

Yang, P. J., & McGinley, M. (2021). Commonalities and specificities of positive youth development in the US and Taiwan. *Journal of Applied Developmental Psychology,* 73(1), 101251. https://doi.org/10.1016/j.appdev.2021.101251

Yuan, K. H., & Marshall, L. L. (2004). A new measure of misfit forcovariance structure models. *Behaviormetrika,* 31(1), 7-90. https://doi.org/10.2333/bhmk.31.6

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal,* 23(3), 392-408. https://doi.org/10.1080/10705511.2015.1118692

1차원고접수 : 2024. 03. 20
2차원고접수 : 2024. 07. 01
최종게재결정 : 2024. 07. 28

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 199-230
http://dx.doi.org/10.22257/kjp.2024.9.43.3.199

Overall evaluation of structural equation models and reflection on effect size and continuity

So-Hyun Yoo **Su-Young Kim**

Department of Psychology, Ewha Womans University

Structural equation model, which is widely used to describe the relationship between latent variables, can be judged by its goodness of fit. The χ^2 test for statistical significance and the effect size index for practical significance of model fit use dichotomous and continuous interpretation approach to evaluate the usefulness of the model, respectively. However, despite the fact that the level of fit is represented on a continuum for practical significance, the calculated effect size index is interpreted dichotomously by using the guideline as an absolute standard. The present study discusses the process of assessing the practical significance of fit in terms of the effect size index and the correct use of guidelines so that researchers evaluating the fit of a model can interpret the level of fit on a continuum. We begin with a brief discussion of the importance of assessing statistical significance using χ^2 test, and then define the concept of effect size in the context of structural equation models. We then introduce the different types of goodness of fit effect size indices and describe the characteristics of the guidelines used to interpret them. Finally, we provide examples of appropriate guidelines for interpreting calculated effect size index values on a continuum and discuss examples of incorrect model evaluation when continuity is not reflected, as well as the correct interpretation of models with marginal fit.

Key words : Structural equation models, interpretation of goodness of fit, effect sizes, continuity, use of guidelines.

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 231-260
http://dx.doi.org/10.22257/kjp.2024.9.43.3.231

반복측정 자료에 기반한 변화의 집단차 분석 방법: 차이점수 모형과 공분산분석 모형 비교*

이 영 수 서강대학교 심리학과 박사과정

석 혜 원† 서강대학교 심리학과 부교수

심리학 여러 분야에서 사전, 사후 시점에 반복측정한 자료에 기반하여 처치집단과 통제집단 간 변화의 차이를 살펴보는 연구를 자주 볼 수 있다. 이때 연구자들이 가장 널리 사용하는 분석 모형은 차이점수 모형과 공분산분석 모형이다. 하지만, 이 두 모형은 때로 상이한 결과를 산출하기 때문에, 많은 연구자들은 언제 어떠한 방법을 사용해야 하는지 혼란을 겪고 있다. 이에, 본 연구는 두 모형을 이론적, 경험적으로 비교한 연구를 개관하고, 이에 기반하여 언제 어느 모형을 사용하는 것이 적절한지 가이드라인을 제시하고자 하였다. 이를 위해, 우선 두 모형을 각각 소개하고, 예시 자료를 통해 두 모형이 서로 다른 분석 결과를 산출할 수 있음을 보였다. 다음으로, 차이점수 사용과 관련된 논쟁을 살펴보고, 차이점수에 대한 전통적인 비판이 지나치게 단순화된 가정과 잘못된 믿음에 근거한 것임을 확인하였다. 이어서, 인과추론의 맥락에서 두 방법이 어떤 숨겨진 가정을 내포하고 있는지 이론적으로 살펴보고, 이러한 가정 및 시뮬레이션 연구 결과들에 기반하여, 실험집단에 참여자를 할당하는 방법과 분석 목적에 따라 어떤 방법을 사용하는 것이 적절한지 가이드라인을 제시하였다. 본 연구를 통해 연구자들이 보다 적절한 분석 방법을 선택하고, 엄밀하고 효과적으로 분석을 수행하는 데 도움을 제공할 수 있을 것으로 기대된다.

주요어 : 차이점수, 공분산분석, 인과추론, 처치효과, Lord의 역설

---

* 이 연구는 재단법인 플라톤 아카데미의 지원(2023년도 서강대학교 특별연구비, 202315003.01) 및 2023년도 서강대학교 교내연구비(202312024.01) 지원을 받아 수행되었음.

† 교신저자: 석혜원, 서강대학교 심리학과, 서울특별시 마포구 백범로 35 (신수동) 서강대학교 다산관 334호, Tel: 02-705-8328, E-mail: hsuk2@sogang.ac.kr

심리학 여러 분야에서 변화의 집단차를 살펴보는 연구를 흔히 볼 수 있다. 예를 들어, Kim과 Park(2019)는 대학생 교육기부 프로그램이 중학생의 역량 발전과 정서 변화에 효과적인지 살펴보기 위해, 프로그램에 참여한 중학생 집단과 참여하지 않은 중학생 집단을 참여 전, 후 두 시점에 걸쳐 측정하고, 역량과 정서 변화에 집단 차이가 있는지 검증하였다. Roh와 Chang(2006)은 대졸취업자의 지각된 과잉자격과 자존감 및 정신건강 간 관련성을 살펴보기 위해, 대학 4학년생을 대학 재학 당시와 졸업 후, 두 시점에 걸쳐 측정하고, 이들을 미취업 집단, 취업자 중 낮은 과잉자격 지각 집단, 취업자 중 높은 과잉자격 지각 집단으로 분류하여, 세 집단 간에 자존감과 정신건강 변화에 차이가 있는지 검증하였다.

이처럼 동일한 대상을 두 시점에 걸쳐 반복측정하고 이에 기반하여 변화에서의 집단차를 분석하고자 할 때 연구자들이 가장 널리 사용하는 분석 모형이 차이점수(difference score) 모형과 공분산분석(Analysis of Covariance; ANCOVA) 모형이다(Castro-Schilo & Grimm, 2018; Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013; Van Breukelen, 2006, 2013). 차이점수 모형은 이름 그대로 처치 전, 처치 후 두 시점 간 점수 차이를 계산하여 이 차이점수에서의 집단차를 검증하는 방법이고, 공분산분석 모형은 처치 전 시점의 점수를 통계적으로 통제하고 처치 후 시점 점수에서의 집단차를 검증하는 방법이다.

그런데, 이 두 분석 방법은 동일한 자료에 대해 때로 상이한 결론을 산출한다. 예를 들어, 자료를 차이점수 모형으로 분석하면 변화에서의 집단차가 유의하지 않지만, 공분산분석 모형으로 분석하면 유의한 집단차가 나타날 수 있다. 이처럼 두 모형의 결과가 상이하게 나타나는 현상을 Lord의 역설(Lord's paradox)이라고 한다(Lord, 1967). 문제는, Lord의 역설이 상당히 빈번히 발생함에도 불구하고, 둘 중 어느 결과가 타당한 것인지 판단하기가 쉽지 않다는 것이다.

Petscher와 Schatschneider(2011)는 2002년부터 2007년까지 Journal of Education Psychology와 Developmental Psychology 두 학술지에 출판된 무선할당 실험 연구 중 집단간 차이 검증을 수행한 연구들이 어떤 분석 방법을 사용했는지 살펴보았는데, 전체 27편의 연구 중 약 44%(12편)가 공분산분석 모형을, 약 37%(10편)가 차이점수 모형을 사용하여, 두 모형이 유사한 정도로 자주 사용되고 있음을 관찰하였다. 동일한 분석 목적과 동일한 실험 설계에 기반한 연구들에서 두 방법이 비슷한 정도로 사용되고 있었던 것이다. 그러나, Jamieson (1994)이 지적했듯, 해당 분석 방법을 왜 선택했는지 근거를 제시하고 있는 연구는 찾아보기 어렵다.

차이점수 모형과 공분산분석 모형을 이론적, 경험적으로 비교하는 연구도 상당수 진행되었으나, 연구 결과들이 산발적으로 제시되어 있고, 이를 통합적으로 잘 정리하여 제시한 연구는 매우 부족한 실정이다. 때문에, 변화의 집단차를 검증하고자 하는 연구자들은 언제 어느 방법을 사용하는 것이 적절한가에 대해 여전히 혼란을 경험하고 있다.

이에, 본 연구는 차이점수 모형과 공분산분석 모형의 차이를 이해하기 위해 두 모형에 대한 이론적, 경험적 선행 연구들을 체계적으로 정리하고, 이를 기반으로 모형 선택에 대한 통합적인 가이드라인을 도출하고자 한다.

차이점수 모형과 공분산분석 모형

논의를 간단히 하기 위해, 두 집단을 두 시점에 결쳐 측정한 자료에 기반하여 변화의 집단차를 검증하는 상황을 가정하도록 하겠다. 첫 번째 시점과 두 번째 시점은 처치, 사건 등과 같은 특정 경험에 의해 구분되며, 이 두 시점에 측정된 점수를 각각 사전점수, 사후점수라고 명명하도록 하겠다. 비교하고자 하는 두 집단은 각각 통제집단, 처치집단이라 명명하되, 참여자가 두 집단에 무선할당(random assignment)되는 경우뿐만 아니라, 비무선할당(nonrandom assignment)된 경우를 모두 통칭하는 의미로 사용하도록 하겠다.

차이점수 모형

차이점수 모형은 동일한 측정 단위를 사용하여 얻은 사전점수와 사후점수 간 차이값을 종속변수로 하고, 집단을 독립변수로 하는 단순회귀모형이라고 할 수 있다. 이 모형을 수식으로 표현하면 식 (1)과 같다.

$$Y_i - X_i = \gamma_0 + \gamma_1 Z_i + e_{1i} \quad (1)$$

이때 X_i와 Y_i는 각각 참여자 i의 사전점수와 사후점수를 나타낸다[1]. 따라서, 식 (1)의 좌변에 위치한 종속변수는 사후점수에서 사전점수를 뺀 차이점수로, 양의 값은 이득이나 성장을, 음의 값은 손실이나 감소를 나타낸다. 식 (1)의 우변에 위치한 독립변수 Z_i는 집단을 나타내는 더미변수로, 0은 통제집단, 1은 처치집단을 가리킨다. 모형의 절편 γ_0는 독립변수 Z_i가 0일 때 기대되는 차이점수 즉, 통제집단의 평균 차이점수를 나타내고, 기울기 γ_1은 Z_i가 1단위 증가할 때 기대되는 차이점수의 변화량 즉, 통제집단에 비해 처치집단의 평균 차이점수가 얼마나 큰지(혹은 작은지)를 나타낸다. 따라서, 차이점수 모형에서는 기울기 γ_1이 바로 변화의 집단차를 나타내며, γ_1의 유의성을 검증하면 변화의 집단차가 유의한지 검증할 수 있다. 마지막으로, e_{1i}는 잔차 즉, 집단으로는 예측할 수 없는 차이점수에서의 개인차를 나타내고, 일반적으로 평균이 0이고 분산이 σ_1^2인 정규분포를 이룬다고 가정한다.

공분산분석 모형

다음으로, 공분산분석 모형은 사후점수 Y_i를 종속변수로 하고, 집단 Z_i을 독립변수, 사전점수 X_i를 공변인으로 하는 다중회귀모형으로, 식 (2)와 같이 나타낼 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_{2i} \quad (2)$$

이 식의 절편 β_0는 X_i와 Z_i가 모두 0일 때 즉, 통제집단에 속하고 사전검사 점수가 0인 사람에게 기대되는 사후검사 점수를 나타낸다. 사전점수 기울기 β_1은 집단 Z_i를 통제하고 사전점수 X_i로 사후점수 Y_i를 예측할 때

1) 흔히, 사전, 사후점수가 동일한 변수를 서로 다른 시점에 측정한 것임을 강조하기 위해 Y_{1i}, Y_{2i}과 같은 기호를 사용해서 사전, 사후점수를 나타내곤 한다. 그러나, 본 논문에서는 이후 인과추론 모형을 설명하는 부분에서, 수식 기호에 지나치게 많은 인덱스가 사용되어 복잡해지는 것을 방지하고자, Y_{1i}, Y_{2i} 대신 X_i, Y_i와 같은 기호를 사용하여 사전, 사후점수를 나타내었다.

의 기울기로, 각 집단 내에서 사전, 사후점수 간 관련성을 나타내는 자기회귀계수라 할 수 있다. β_2는 사전점수 X_i를 통제한 후, 집단변수 Z_i로 사후점수 Y_i를 예측할 때의 기울기로, 사전점수가 동일할 때 기대되는 사후점수에서의 집단차를 나타낸다. 마지막으로, e_{2i}는 잔차 즉, 집단과 사전점수로는 설명되지 않는 사후점수에서의 개인차를 나타내며, 일반적으로 평균이 0이고 분산이 σ_2^2인 정규분포를 이룬다고 가정한다. 참고로, 앞의 식 (1)에서의 잔차와 식 (2)에서의 잔차는 동일하지 않기 때문에, 각 식에서의 잔차를 e_{1i}, e_{2i}와 같이 서로 다른 기호를 사용하여 구분하였다.

식 (2)의 양변에서 $\beta_1 X_i$를 빼면 식 (3)을 얻을 수 있다. 이때 식 (3)의 좌변은 각 집단 내에서 사후점수 Y_i를 사전점수 X_i로 예측하고 난 후의 잔차를 나타낸다. 때문에, 공분산분석 모형을 잔차화된 차이점수(residualized change score) 모형이라고 부르기도 한다 (Castro-Schilo & Grimm, 2018).

$$Y_i - \beta_1 X_i = \beta_0 + \beta_2 Z_i + e_{2i} \quad (3)$$

즉, 공분산분석 모형에서 집단 간에 비교하는 변화란, 사전점수로 예측되는 것 이상으로 사후점수가 변화한 정도를 의미한다. 따라서, 공분산분석 모형에서는 집단변수 기울기 β_2가 바로 변화의 집단차를 나타내며, β_2의 유의성을 검증하면 변화의 집단차가 유의한지 검증할 수 있다.

Lord의 역설

이제 앞서 살펴본 차이점수 모형과 공분산분석 모형을 실제 자료에 적용했을 때 분석 결과에 어떠한 차이가 나타날 수 있는지 살펴보도록 하겠다.

분석에 사용된 예시 자료는 서강대학교 희망연구소에서 2022년 국내 3개 대학에 소속된 학생들을 대상으로 수집한 단기 종단자료의 일부로, 본 논문에서는 관련 문항에 응답한 186명의 자료를 분석에 사용하였다. 분석의 목적은 목표달성 정도에 따라 기본심리욕구 충족 수준 변화에 차이가 있는지 살펴보는 것이었다.

자료 수집 과정은 다음과 같다. 연구에 참여한 학생들에게 학기 초(2022년 3월)에 자신이 이번 학기에 추구하고자 하는 목표 세 가지를 적도록 하고, 학기 말(2022년 7~8월)에 각 목표의 달성 정도를 측정하였다. 각 목표에 대한 달성 점수는 '나는 이 목표를 향해 많은 진전을 이루었다', '나는 이 목표 계획을 순조롭게 진행하고 있는 것 같다', '나는 이 목표를 이룬 것 같다'의 세 문항에 대한 응답 평균으로 구하였고, 응답은 7점 척도(1=전혀 동의하지 않는다, 7=매우 동의한다)로 측정하였다. 그리고, 해당 측정치에 기반하여 참여자들을 목표달성 고집단과 저집단으로 구분하였는데, 학기 초에 제출한 세 가지 목표에 대한 달성 점수 평균이 전체 평균보다 높은 경우 목표달성 고집단으로, 낮은 경우 목표달성 저집단으로 구분하였다.

기본심리욕구 충족 수준은 Lee와 Kim(2008)이 개발한 한국형 기본심리욕구 척도를 사용하여 학기 초와 학기 말 두 시점에 측정하였다. 각 문항에 대한 응답은 6점 척도(1=전혀 아니다, 6=매우 그렇다)를 사용하여 측정하였고, 전체 18문항에 대한 총점을 계산하여 분석에 사용하였다. 척도의 신뢰도(Cronbach's α

는 학기 초와 학기 말에 각각 0.89, 0.90으로 나타났다.

표본 평균을 살펴본 결과, 그림 1에서와 같이 학기 초에 측정한 기본심리욕구 수준은 목표달성 고집단이 저집단에 비해 더 높은 것으로 나타났다. 두 집단 모두 학기 초에 비해 학기 말 기본심리욕구 충족 수준이 다소 낮아진 것으로 나타났으며, 감소 폭은 목표달성 고집단에 비해 저집단이 약간 크게 나타났다.

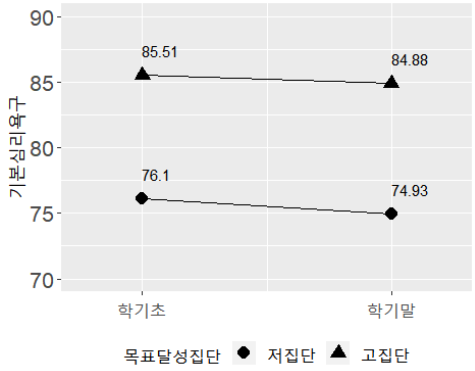


그림 1. 목표달성 집단별 기본심리욕구 평균 변화

목표달성 고집단과 저집단 간에 기본심리욕구 충족 수준 변화에 차이가 있는지 검증하기 위해, 차이점수 모형과 공분산분석 모형을 사용하여 해당 자료를 각각 분석하였다. 그 결과, 차이점수 모형에서는 목표달성 고집단과 저집단 간에 기본심리욕구 충족 수준의 변화에 유의한 차이가 나타나지 않은 반면(b=-0.54, t=-0.55, p=0.58), 공분산분석 모형에서는 학기 초 기본심리 욕구 충족 수준을 통제한 후 목표달성 고집단과 저집단 간 차이를 분석한 결과, 학기 말 기본심리욕구 충족 수준에서 유의한 집단 차이가 나타났다(b=2.66, t=1.21, p=0.029).

동일한 자료를 분석했음에도 불구하고, 차이점수 모형으로 분석하면 목표달성 수준이 기본심리욕구 충족 변화에 영향을 미치지 않는다는 결론이 도출되는 반면, 공분산분석 모형으로 분석하면 목표달성의 수준이 높은 집단에서 기본심리욕구 충족 수준이 덜 감소한다는 결론이 도출된다. 즉, 차이점수 모형과 공분산분석 모형이 변화의 집단차에 대한 서로 다른 결론을 도출하는 Lord의 역설이 발생했음을 알 수 있다.

Lord의 역설이 발생하는 조건

그렇다면, Lord의 역설은 언제 발생하는 것일까? 차이점수 모형과 공분산분석 모형을 보다 쉽게 비교하기 위해, 식 (2)의 좌변을 식 (1)과 동일하게 나타내보자. 공분산분석 모형을 나타내는 식 (2)의 양변에서 X_i를 빼면 식 (4)를 얻을 수 있다. 즉, 공분산분석 모형은 차이점수 $Y_i - X_i$를 종속변수로 하고, 사전점수 X_i와 집단변수 Z_i를 독립변수로 하는 다중회귀모형임을 알 수 있다(Werts & Linn, 1970). 따라서, 공분산분석 모형과 차이점수 모형은 모두 차이점수를 종속변수로 하지만, 사전점수 X_i를 공변인으로 포함하느냐 하지 않느냐의 측면에서 핵심적인 차이를 보인다.

$$Y_i - X_i = \beta_0 + (\beta_1 - 1)X_i + \beta_2 Z_i + e_{2i} \quad (4)$$

식 (4)와 식 (1)을 비교하면, 두 모형이 언제 동일한 결과를 산출하고, 언제 서로 다른 결과를 산출하는지 알 수 있다. 우선, 식 (4)에서 $\beta_1 = 1$일 때는 식 (4)와 식 (1)이 동일해짐을 알 수 있다. 따라서, $\beta_1 = 1$이 성립하면

$\gamma_1 = \beta_2$가 성립하고, 두 모형은 변화의 집단차에 대해 동일한 결론을 산출한다. 만약, 식 (4)에서 $\beta_1 \neq 1$일 때 $\gamma_1 = \beta_2$가 성립하려면, 사전점수 X_i와 집단변수 Z_i가 서로 독립적이어야 한다. 이 경우, Z_i의 기울기는 X_i가 모형에 포함되었는지의 여부에 영향을 받지 않는다. 만약 $\beta_1 = 1$혹은 X_i와 Z_i의 독립성, 이 두 조건 중 어느 하나도 충족되지 않으면 $\gamma_1 = \beta_2$가 성립하지 않고, 이 경우 두 모형이 변화의 집단차에 대한 서로 다른 결론을 산출하는 Lord의 역설이 발생한다(Castro-Schilo & Grimm, 2018).

Gollwitzer와 동료들(2014)은 Lord의 역설이 발생하는 보다 구체적인 상황을 제시하였다. 만약, 사전점수에 집단 차이가 없고, 사전, 사후점수 간 완벽한 안정성이 존재하여 모든 개인들의 변화 정도가 동일하고, 점수가 완벽하게 신뢰롭다면, 식 (2)에서 자기회귀계수 β_1은 1의 값을 갖게 되어 Lord의 역설이 발행하지 않는다. 그러나, 사전, 사후점수 간에 완벽한 안정성이 존재한다고 하더라도, 사전점수에 측정오차가 존재하여 신뢰도가 낮아지면, 이로 인해 자기회귀계수 β_1이 과소추정되어 1보다 작은 값을 갖게 된다. 이때 사전점수 X_i와 집단변수 Z_i가 독립적이지 않다면(즉, 사전점수에 집단차가 존재한다면) X_i의 기울기 β_1은 Z_i의 기울기 β_2에 영향을 미쳐 추정의 편향을 가져온다.

처치집단과 통제집단에 참여자를 무선할당하는 통제된 실험의 경우에는 사전점수에 집단차가 존재하지 않을 것으로 기대되지만, 비무선할당 연구에서는 집단 간 사전점수에 체계적인 차이가 존재할 수 있다. 따라서, 무선할당이 불가능하거나 실패하여 사전점수에 집단 차이가 존재하고, 사전점수에 측정오차가 개입되는 경우, $\gamma_1 = \beta_2$이 성립하지 않는 Lord의 역설이 발생하게 된다.

앞의 예시에서도, 비무선할당으로 인해 사전점수에 체계적인 집단 차이(즉, 목표달성 고집단의 평균이 저집단에 비해 높음)가 존재했고, 사전점수의 신뢰도가 1보다 낮았기 때문에 Lord의 역설이 발생한 것이라고 할 수 있다. 심리학 연구에서 대부분의 경우 사전점수의 측정오차를 완전히 제거하는 것은 불가능하고, 비무선할당 설계 또한 널리 사용된다는 점을 감안하면, Lord의 역설이 아주 드물게 예외적인 상황에서만 발생하는 것은 아님을 알 수 있다.

그렇다면, 두 모형이 산출한 서로 다른 결과 중 어느 결과가 정확한 것인가? 언제 어느 모형을 사용하는 것이 적절한가? 본 논문에서는 이러한 질문에 답하기 위해 지난 수십 년에 걸쳐 이루어진 관련논쟁과 연구를 다음과 같이 세 부분으로 제시해보고자 한다. 우선, 차이점수 사용에 대한 비판과 이에 대한 반박을 정리하고, 다음으로 처치효과 추정을 위한 인과추론(causal inference) 맥락에서 두 방법을 비교한 이론적 연구들을 제시하도록 하겠다. 마지막으로, 이론적 연구 및 시뮬레이션 연구에 기반하여, 언제 어느 방법을 사용하는 것이 적절한지 가이드라인을 제시하도록 하겠다.

차이점수에 대한 논쟁

차이점수는 신뢰도가 낮고, 사전점수와 부적 상관을 보인다는 두 가지 이유로 오랫동안 비판을 받아왔다. 이러한 비판이 널리 받아들여지면서, 분석 목적과는 상관없이 차이점수

를 분석에 사용하는 것 자체에 대해 현재까지도 많은 연구자들이 부정적 인식을 가지고 있다. 그러나, 차이점수에 대한 비판이 타당하지 않다는 반박 또한 꾸준히 제기되고 있다. 이에, 본 논문에서는 차이점수 비판의 근거를 짚어보고, 그 타당성에 대한 논쟁을 정리해보도록 하겠다.

차이점수의 신뢰도

차이점수가 비판을 받은 주된 이유는 바로 차이점수의 신뢰도가 낮다는 것이다. 특히, Cronbach와 Furby(1970)는 차이점수의 측정오차가 매우 크기 때문에 차이점수를 사용하면 왜곡된 결론을 도출하게 된다고 강하게 비판하였다. 또한, 이들은 변화를 보다 정확하게 측정하기 위해 여러 학자들이 제안한 차이점수 보정도 사용할 이유가 없다고 주장하였다. 이들의 논문이 발표된 이후로 많은 연구자들이 차이점수를 사용하는 것에 부정적 인식을 갖게 되었고, 차이점수를 사용하여 분석을 수행한 연구는 리뷰어들의 강한 비판에 직면하곤 하였다(Gollwitzer et al., 2014).

차이점수의 신뢰도가 낮다는 주장(e.g., Cronbach & Furby, 1970; Linn & Slinde, 1977; Lord, 1963)은 식 (5)에 제시된 차이점수 신뢰도 공식(Gulliksen, 1950)에 기반한다.

$$Rel(Y-X)=\frac{Rel(X)-\rho_{XY}}{1-\rho_{XY}} \qquad (5)$$

이 식에서 $Rel(\cdot)$는 신뢰도를 의미하고, ρ_{XY}는 사전, 사후점수 간 상관을 나타낸다. 참고로, 식 (5)는 사전점수와 사후점수가 동일한 신뢰도를 가질 때에만 성립한다. 즉, 식 우변에 등장하는 사전점수 신뢰도 $Rel(X)$는 사후점수 신뢰도인 $Rel(Y)$와 동일하다고 가정된다.

식 (5)에 따르면, 차이점수의 신뢰도는 각 시점에서 측정한 점수의 신뢰도 $Rel(X)$와 사전, 사후점수 간 상관 ρ_{XY} 이 두 값에 따라 달라진다. 우선, 검사점수의 신뢰도가 높아지면 차이점수의 신뢰도 또한 높아진다. 예를 들어, 사전, 사후점수 간 상관이 0.5일 때, 검사점수의 신뢰도가 0.8에서 0.9로 높아지면, 차이점수의 신뢰도는 (0.8-0.5)/(1-0.5)=0.6에서 (0.9-0.5)/(1-0.5)=0.8로 높아진다. 다음으로, 사전, 사후점수 간 상관이 높아지면 차이점수의 신뢰도는 낮아진다. 예를 들어, 사전, 사후점수의 신뢰도가 모두 0.8이고, 사전, 사후점수 간 상관이 0.2로 낮다면, 차이점수의 신뢰도는 (0.8-0.2)/(1-0.2)=0.75가 된다. 그러나, 사전, 사후점수 간 상관이 0.75와 같이 높다면, 차이점수의 신뢰도는 (0.8-0.75)/(1-0.75)=0.2로 매우 낮아진다.

즉, 차이점수가 신뢰롭기 위해서는 사전, 사후점수가 신뢰로워야 할 뿐만 아니라, 사전, 사후점수 간 상관이 낮아야 한다. 그런데, 사전, 사후점수 간에 어느 정도의 정적 상관이 존재한다면, 식 (5)에서 볼 수 있듯이 차이점수의 신뢰도는 사전 혹은 사후점수의 신뢰도보다 결코 높을 수 없다. 이것이 바로 차이점수 사용을 비판하는 주된 논거이다(Chiou & Spreng, 1996).

이러한 주장은 일견 매우 타당해 보이지만, 식 (5)가 성립하려면 다음 두 가지 가정이 만족되어야 한다. 첫째, 이미 언급한 것처럼 점수의 신뢰도가 사전, 사후에 모두 같아야 하고, 둘째, 두 시점의 점수가 동일한 표준편차를 가져야 한다(Chiou & Spreng, 1996;

Gollwitzer et al., 2014; Lord, 1956). 이 두 가지 가정이 만족되지 않는 보다 일반적인 상황에서 차이점수의 신뢰도는 식 (6)과 같이 나타낼 수 있다(Gollwitzer et al., 2014; Lord, 1963; Zimmerman & Williams, 1982, 1998).

$$Rel(Y-X)=\frac{Rel(X)+\lambda^2 Rel(Y)-2\lambda\rho_{XY}}{1+\lambda^2-2\lambda\rho_{XY}} \quad (6)$$

식 (6)에서 λ는 사전점수 표준편차(σ_X)와 사후점수 표준편차(σ_Y) 간 비율 즉, $\lambda=\sigma_Y/\sigma_X$을 나타낸다. 식 (6)에서 사전, 사후점수의 신뢰도가 동일하고, 사전, 사후점수의 표준편차가 동일하면($\lambda=1$), 식 (6)과 식 (5)가 같아진다는 것을 확인할 수 있다.

차이점수 모형의 낮은 신뢰도를 비판하는 연구자들은 이러한 가정에 대해 자세히 논의하지 않았다. 그러나, 현실에서는 이러한 가정이 만족되지 않는 경우가 종종 발생한다. 특히, 실제 연구에서는 사전, 사후점수의 표준편차가 서로 다르게 나타나곤 한다(Chiou & Spreng, 1996; Gollwitzer et al., 2014). 예를 들어, 사람에 따라 처치에 반응하는 정도가 다르다면, 사전점수에 비해 사후점수의 이질성이 증가하여 사후에 더 큰 폭의 개인차를 나타내는 확산효과(spreading effect; Gollwitzer et al., 2014)가 발생할 수 있다. 반대로, 처치효과가 매우 강력하다면, 처치를 경험한 사람들 모두가 유사한 수준의 높은 점수를 나타내게 되고, 사전점수에 존재하던 개인차가 처치 후에는 크게 줄어드는 축소효과(narrowing effect)가 발생하게 된다(Gollwitzer et al., 2014).

이처럼 확산 혹은 축소효과가 발생하여 사전, 사후점수의 표준편차가 서로 다를 경우, 식 (5)는 더 이상 성립하지 않고, 사전, 사후점수 간 상관이 높더라도 차이점수의 신뢰도는 식 (5)에서 산출하는 것만큼 낮아지지 않는다(Chiou & Spreng, 1996; Gollwitzer et al., 2014). Zimmerman과 Williams(1998)는 식 (6)에 포함된 항들 즉, 사전, 사후점수의 신뢰도, 사전, 사후점수 표준편차 비율, 사전, 사후점수 간 상관은 각각 따로 변하는 것이 아니라, 하나가 달라지면 다른 값들이 이에 따라 변한다는 것을 지적하였다. 그러면서, 만약 처치로 인해 사후점수(보다 정확히는 사후 진점수)의 표준편차가 변하면, 이에 따라 사후점수의 신뢰도 및 사전, 사후점수 간 상관도 모두 변하고, 이 경우 차이점수는 적어도 사전 혹은 사후점수 만큼의 신뢰도를 나타낸다고 하였다. 이들은 차이점수의 신뢰도가 낮다는 주장은 매우 예외적이고 비현실적인 가정(처치로 인해 진점수의 표준편차가 변화하지 않고, 사전점수의 신뢰도가 낮으며, 사전, 사후 진점수 간 상관이 높다는 가정)에 기반한 것이며, 사전점수가 충분히 신뢰롭기만 하다면, 대부분의 상황에서 차이점수는 연구에 사용할 수 있을만큼 충분히 신뢰롭다고 하였다.

Rogosa와 동료들 또한 차이점수의 신뢰도가 항상 낮은 것은 아닐 뿐만 아니라(Rogosa & Willett, 1983), 설령 차이점수의 신뢰도가 낮다고 하더라도 이것이 차이점수를 사용하지 말아야 할 타당한 이유는 아니라고 지적하였다(Rogosa et al., 1982). 고전검사이론(classical test theory; Lord et al., 1968)에 기반하면, 차이점수의 신뢰도는 관찰된 차이점수의 분산 중 측정오차로 인한 변화가 아닌 실제 점수 변화로 인해 발생한 분산의 비율을 나타낸다. 따라서, 측정오차가 거의 개입되지 않는다고 하더라도 실제 점수 변화로 인한 분산이 작다면 신뢰도

는 낮게 나타날 수밖에 없다. 극단적으로, 모든 사람들이 동일한 정도로 변화하여 모든 사람들의 실제 점수에 변화가 없다면, 아무리 측정을 정확하게 하더라도 차이점수의 신뢰도는 0이 된다. 이에, Rogosa와 동료들(1982)은 차이점수가 신뢰로우려면 변화에 개인차가 존재해야 하는 것은 맞지만, 변화의 개인차가 작다는 것(그래서 신뢰도가 낮다는 것)이 곧 차이점수가 무용하다는 의미는 아니라고 지적하였다.

Overall과 Woodward(1975) 그리고 Thomas와 Zumbo(2012)는 집단차를 검증하는 맥락에서 차이점수의 낮은 신뢰도는 문제가 되지 않는다고 주장하였다. 이들은 차이점수의 집단차를 검증할 때 차이점수의 신뢰도가 0인 경우(즉, 같은 집단에 속한 사람들이 모두 동일한 정도로 변화하는 경우) 역설적이게도 통계적 검증력은 최대가 된다는 것을 보이면서, 사전, 사후점수 자체의 신뢰도가 낮은 것은 우려할 만한 일이지만, 차이점수 자체의 신뢰도가 낮은 것은 문제가 되지 않는다고 하였다.

차이점수와 사전점수 간 부적 상관

낮은 신뢰도와 함께 차이점수를 비판하는 또 다른 주요한 근거는 바로 차이점수가 일반적으로 사전점수와 부적 상관을 나타낸다는 점이다. 사전점수와 차이점수 간 상관 $\rho_{X,Y-X}$는 식 (7)과 같이 나타낼 수 있다(Linn & Slinde, 1977). 이 식에서 ρ_{XY}는 사전, 사후점수 간 상관을, σ_X와 σ_Y는 각각 사전, 사후점수의 표준편차를 나타낸다.

$$\rho_{X,Y-X}=\frac{\rho_{XY}\sigma_Y-\sigma_X}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho_{XY}\sigma_X\sigma_Y}} \quad (7)$$

식 (7)에서 사전, 사후점수의 표준편차가 동일하도록 점수가 표준화되었다고 가정하면($\sigma_X=\sigma_Y$), 일반적으로 사전, 사후 점수 간 상관 ρ_{XY}은 1보다 작기 때문에 식 (7)의 분자가 0보다 작아지고, 사전점수와 차이점수 간 상관이 음의 값을 갖게 됨을 알 수 있다(Linn & Slinde, 1977).

사전점수와 차이점수 간 부적 상관을 때로 평균으로의 회귀(regression toward the mean) 현상으로 설명하기도 한다. Galton(1886)에 의해 처음 소개된 평균으로의 회귀란, 키와 같은 부모의 속성이 평균으로부터 더 극단적으로 멀어질수록 자식은 부모보다는 평균에 더 가까운 속성값을 보이는 현상을 가리킨다. 한 개인을 두 번 반복측정한 맥락에서 평균으로의 회귀란, 사전점수가 평균 이하인 개인의 사후점수는 증가하고, 사전점수가 평균 이상인 개인의 사후점수는 감소하여, 각 개인의 사후점수가 사전점수에 비해 평균에 더 가까워지는 경향을 의미한다(Furby, 1973; Nesselroade et al., 1980). 이 경우, 사전점수가 낮을수록 사후점수는 증가하므로 사후점수에서 사전점수를 뺀 차이점수는 커지고, 반대로 사전점수가 높을수록 사후점수는 감소하여 차이점수는 작아지기 때문에, 사전점수와 차이점수 간 부적 상관이 나타나게 된다.

평균으로의 회귀는 사전, 사후점수 간에 완벽한 상관이 존재하지 않을 경우 언제나 발생한다. 사전, 사후점수가 동일한 표준편차를 갖도록 표준화된 경우, 사전점수로 사후점수를 예측하는 회귀식은 식 (8)과 같다(Nesselroade et al., 1980).

$$Y_i-\mu_Y=\rho_{XY}(X_i-\mu_X) \quad (8)$$

이때 μ_X, μ_Y는 각각 사전, 사후점수의 평균을 나타내고, ρ_{XY}는 사전, 사후점수 간 상관을 나타낸다. 이 식에 따르면, 사전, 사후점수 간 상관이 완벽하지 않을 경우, 사전점수 X가 평균 μ_X에서 떨어진 거리보다 사후점수 Y가 평균 μ_Y로부터 떨어진 거리가 더 작아지는 평균으로의 회귀 현상이 발생할 수밖에 없다. 이에, Nesselroade와 동료들(1980)은 평균으로의 회귀와 사전, 사후점수 간 불완전한 상관은 결국 동일한 현상을 가리키는 것이라고 하였다.

사전, 사후점수가 완벽한 상관을 보이기 어렵다는 점을 고려하면, 평균으로의 회귀 현상은 보편적으로 발생한다고 할 수 있다(Furby, 1973). 즉, 사전점수와 차이점수 간에는 전형적으로 부적 상관이 존재한다고 할 수 있으며, 이는 차이점수에 대한 중요한 비판의 근거가 되었다. 예를 들어, '어떤 사람들의 자존감이 더 많이 향상되는가?', '어떤 환자들의 우울이 더 많이 감소하는가?' 등과 같이 개입이나 치료가 누구에게 더 효과적인지 평가하는 상황을 가정해보자. 이때 차이점수를 사용하면, 차이점수는 항상 사전점수와 부적 상관을 나타내므로, 사전점수가 더 낮은(혹은 높은) 사람들에게 유리한 결과가 나타날 수밖에 없고, 따라서 차이점수를 사용하는 것은 타당하지 않다는 것이다(Linn & Slinde, 1977). 또한, 사전점수가 집단과 연관되어 있으면, 차이점수와 집단 간에 의미있는 관련성이 존재하지 않더라도, 차이점수와 집단이 공통적으로 사전점수와 연관되어 있기 때문에 둘 간에 허위상관(spurious correlation)이 발생할 우려도 있다.

그러나, Rogosa(1995)는 사전점수와 차이점수 간에 항상 부적인 상관이 존재한다는 것은 잘못된 믿음이라고 지적하였다. 사전점수와 차이점수 간에는 상관이 존재하지 않을 수도 있으며, 확산(fan spread)이 발생하면 오히려 정적 상관이 존재하기도 한다. 식 (7)에서 사전, 사후점수 간에 정적 상관이 존재하고, 확산이 발생하여 사후점수의 표준편차가 사전점수의 표준편차보다 충분히 커지면, 사전점수와 차이점수 간에는 정적 상관이 존재하게 됨을 알 수 있다. 또한, Rogosa와 Willett(1985)는 사전점수가 언제 측정되었는가에 따라 사전점수와 차이점수 간 상관의 크기와 방향이 완전히 달라질 수 있음을 보였다. 이들은 개인들의 변화 궤적이 선형적인 경우를 가정하고 초기값과 변화량 간의 상관을 살펴보았을 때, 언제 측정한 값을 초기값이라고 정의하느냐에 따라 이 상관 값이 음수부터 양수까지 극적으로 달라짐을 보였다.

이와 더불어, Rogosa(1995)는 평균으로의 회귀 현상이 언제나 발생한다는 믿음은 사전, 사후점수의 표준편차가 동일하다는 가정하에서만 성립함을 지적하였다. Furby(1973)는 평균으로의 회귀를 "for a given score on x (e.g., x'), the corresponding mean score on y (e.g., y') is closer to $\overline{Y}$ in standard deviation units than x' is to $\overline{X}$ in standard deviation unit[사전점수가 표준편차 단위로 평균에서 떨어진 거리보다 사후점수가 표준편차 단위로 평균에서 떨어진 거리가 더 가까운 것]"이라고 정의했는데, 이는 평균으로의 회귀가 원점수가 아닌 표준화된 점수에서 발생함을 의미한다. Rosoga(1995)는 평균으로의 회귀를 표준화된 점수가 아닌 원점수에 대해 정의하는 것이 보다 현실적이라고 주장하였다. 그리고, 사전, 사후점수의 분산이 동일하게 표준화되지 않는다면, 사전점수와 차이점수 간 부적상관은 항상 발생하

는 것이 아니며, 이러한 부적상관이 존재할 경우에만 평균으로의 회귀가 발생하는 것이라고 하였다.

지금까지 차이점수에 대한 비판과 그에 대한 반박을 살펴보았다. 차이점수를 비판하는 첫 번째 근거는 낮은 신뢰도이다. 그러나, 차이점수 신뢰도가 낮게 나타나는 상황은 매우 제한적이며, 만약 변화에 개인차가 크지 않다면 아무리 정밀하게 이를 측정하더라도 신뢰도는 낮을 수밖에 없고, 차이점수에 대한 집단차 검증은 신뢰도가 낮을 때 오히려 높은 검증력을 나타낸다. 차이점수를 비판하는 두 번째 근거는 차이점수가 사전점수와 부적 상관을 갖는다는 것이다. 그러나, 이러한 부적 상관은 사전, 사후점수가 동일한 표준편차를 갖도록 표준화된 경우에 한해서만 항상 발생하며, 일반적으로 차이점수와 사전점수는 사전, 사후점수의 표준편차 비율에 따라 상관이 없거나 정적 상관을 보이기도 한다. 따라서, 차이점수를 비판하는 두 가지 근거는 모두 변화의 집단차 검증에서 차이점수를 사용하지 말아야 할 타당한 근거라고 하기 어렵다.

인과추론 맥락에서 차이점수 모형과 공분산분석 모형의 비교

차이점수 자체에 대한 비판과는 별개로, 차이점수 모형과 공분산분석 모형 중 어느 모형을 언제 사용하는 것이 적절한가에 대한 논의는 처치효과 추론 혹은 인과추론(causal inference; Holland & Rubin, 1983; Maris, 1998)의 맥락에서 빈번하게 이루어졌다. 본 논문에서는 Holland와 Rubin(1983)의 인과추론 모형(model for causal inference)에 기반하여, 차이점수 모형과 공분산분석 모형의 차이를 이론적으로 살펴보도록 하겠다.

인과추론 모형

어떤 처치(treatment)가 통제(control) 조건과 비교하여 종속변수에 대해 나타내는 처치효과(treatment effect; Maris, 1998) 혹은 인과효과(causal effect; Holland & Rubin, 1983)란, 한 개인이 통제조건에 노출되었을 때에 비해 처치조건에 노출되었을 때 발생하는 점수의 변화 혹은 차이를 나타낸다. 다른 모든 요인은 완전히 동일하게 유지하고, 한 개인을 통제 혹은 처치조건 중 어느 조건에 노출했는지만 바꾸었을 때 점수가 달라진다면, 이 점수 변화는 처치의 인과적 효과를 반영하는 것이라 할 수 있다.

식 (9)는 이러한 처치효과를 수식으로 나타낸 것이다(Holland, 1986; Maris, 1998). 여기서 Δ_i는 i번째 개인이 나타내는 처치효과를, Y_{ti}는 해당 개인이 처치조건에 노출되었을 때의 점수를, Y_{ci}는 동일한 개인이 통제조건에 노출되었을 때의 점수를 가리킨다.

$$\Delta_i = Y_{ti} - Y_{ci} \tag{9}$$

일반적으로, 한 개인에 대한 처치효과를 추정하는 것은 불가능하다. 실제 연구에서 한 개인은 처치 혹은 통제조건 중 한 조건에만 할당되므로, 할당된 조건(예를 들어, 처치조건)에서의 점수만 관찰할 수 있고, 할당되지 않은 조건(예를 들어, 통제조건)에서의 점수는 관찰할 수 없기 때문이다[2]. 이것을 인과추

2) 반복측정 설계의 경우, 한 개인을 처치조건과 통

론의 근본 문제(fundamental problem of causal inference)라 한다(Holland, 1986).

개개인에 대한 처치효과 추정은 불가능하지만, 모집단 수준에서의 처치효과 추정은 가능하다. 식 (9)에 기반하여 모집단에서의 평균 처치효과(average treatment effect; Maris, 1998) 혹은 평균 인과효과(average causal effect; Holland, 1986)를 정의하면 식 (10)과 같다. 여기서 $E(\cdot)$는 모집단에 속한 모든 개인 i들에 대해 구한 평균 즉, 기댓값을 나타낸다.

$$E(\Delta) = E(Y_t - Y_c) = E(Y_t) - E(Y_c) \quad (10)$$

식 (10)에서 볼 수 있듯이, 평균 처치효과 $E(\Delta)$는 모집단에 속한 모든 개인들이 처치조건에 노출되었을 때의 점수 평균 $E(Y_t)$에서 모든 개인들이 통제조건에 노출되었을 때의 점수 평균 $E(Y_c)$을 뺀 차이로 정의할 수 있다. 그런데, 여기서도 두 점수 Y_{ti}와 Y_{ci}를 동일한 개인에게서 모두 측정할 수 없다는 문제가 여전히 존재한다. 즉, 처치집단에 속한 개인들에게는 처치조건에 노출되었을 때의 점수만을, 통제집단에 속한 개인들에게는 통제조건에 노출되었을 때의 점수만을 측정할 수 있기 때문에, $E(Y_t)$, $E(Y_c)$와 같이 모집단 전체에 대해 정의되는 기댓값을 추정하는 것은 불가능하다.

대신, 처치집단에 속한 사람들의 Y_{ti}에 대한 기댓값, 통제집단에 속한 사람들의 Y_{ci}에 대한 기댓값과 같이 모집단 중 일부에 대한 기댓값은 추정 가능하다. 이렇게 모집단 전체가 아닌 모집단 중 일부 집단에 대해 정의되는 평균을 조건부 기댓값이라고 하며, 처치집단에 대한 Y_{ti}의 조건부 기댓값은 $E(Y_t \mid \text{처치})$, 통제집단에 대한 Y_{ci}의 조건부 기댓값은 $E(Y_c \mid \text{통제})$와 같이 나타낸다 (Holland & Rubin, 1983; Maris, 1998).

만약, 식 (11), (12)와 같이 모집단 전체에 대한 기댓값과 일부 집단에 대한 조건부 기댓값이 동일하다면, 평균 처치효과를 식 (13)과 같이 나타낼 수 있고, 이 경우 처치집단과 통제집단 각각의 표본 평균 점수를 사용하여 평균 처치효과를 편향없이(unbiasedly) 추정할 수 있다(Holland, 1986).

$$E(Y_t) = E(Y_t \mid \text{처치}) \quad (11)$$

$$E(Y_c) = E(Y_c \mid \text{통제}) \quad (12)$$

$$E(\Delta) = E(Y_t \mid \text{처치}) - E(Y_c \mid \text{통제}) \quad (13)$$

처치집단과 통제집단에 참여자를 무선적으로 할당하는 실험에서는 식 (11), (12)가 성립한다. 무선할당 절차는 실험집단과 점수 간의 독립성을 보장하며, 이로 인해 전체 모집단에 대한 기댓값과 실험집단에 따른 조건부 기댓값에 차이가 발생하지 않기 때문이다.

그러나, 질병 유무에 따른 처치효과 검증의 경우와 같이 무선할당이 윤리적으로 불가능하거나, 현실적 제약으로 인해 참여자를 실험집단에 비무선적으로 할당할 수밖에 없는 경우에는 식 (11), (12)가 성립하지 않고, 보다 일

제조건에 모두 노출시킬 수 있고, 따라서 Y_{ti}와 Y_{ci}를 동일한 개인에게서 모두 얻을 수 있다고 생각될 수도 있다. 그러나, 한 개인이 처치조건과 통제조건을 동시에 경험하는 것은 불가능하고, 하나씩 순차적으로 경험하는 것만 가능하다. 때문에, 이미 한 조건을 경험한 개인이 그 이전의 개인과 완전히 동일하다고 보기 어려우며, 이월효과(carryover effect)로 인해 점수가 달라질 가능성도 있다.

반적으로 식 (14), (15)가 성립한다(Holland & Rubin, 1983). 식 (14), (15)에서 P(처치) 및 P(통제)는 각각 처치집단과 통제집단에 속할 확률을 나타내며, 한 개인은 처치 혹은 통제집단 중 하나에 반드시 속하므로, 이 두 확률을 합치면 항상 1이 된다.

$$E(Y_t) = E(Y_t \mid \text{처치})P(\text{처치}) + E(Y_t \mid \text{통제})P(\text{통제}) \quad (14)$$

$$E(Y_c) = E(Y_c \mid \text{처치})P(\text{처치}) + E(Y_c \mid \text{통제})P(\text{통제}) \quad (15)$$

식 (14)와 (15)를 사용하면 평균 처치효과를 식 (16)과 같이 나타낼 수 있다(Maris, 1998). 식 (16)은 무선할당 실험이 아닌 경우에도 일반적으로 성립하지만, 이 식에 기반하여 평균 처치효과를 추정하는 것은 여전히 불가능하다. 왜냐하면, $E(Y_t \mid \text{통제})$를, 통제집단에 속한 개인들이 처치조건에 노출되었다면 얻어졌을 점수의 기댓값과 $E(Y_c \mid \text{처치})$를, 처치집단에 속한 개인들이 통제조건에 노출되었다면 얻어졌을 점수의 기댓값은 추정이 불가능하기 때문이다.

$$E(\Delta) = E(Y_t \mid \text{처치})P(\text{처치}) + E(Y_t \mid \text{통제})P(\text{통제}) - E(Y_c \mid \text{처치})P(\text{처치}) - E(Y_c \mid \text{통제})P(\text{통제}) \quad (16)$$

대신, 만약 이 값들을 다른 관찰 가능한 점수를 사용해서 예측하는 것이 가능하다면, 예측된 $E(Y_t \mid \text{통제})$와 $E(Y_c \mid \text{처치})$값을 사용하여 처치효과를 추정할 수 있을 것이다(Maris, 1998). 차이점수 모형과 공분산분석 모형은 $E(Y_t \mid \text{통제})$와 $E(Y_c \mid \text{처치})$를 예측하여 처치효과를 추정하며, 두 모형 모두 $E(Y_t \mid \text{통제})$와 $E(Y_c \mid \text{처치})$를 예측하기 위해 사전점수를 사용한다. 그러나, 두 모형은 사전점수를 사용하여 $E(Y_t \mid \text{통제})$와 $E(Y_c \mid \text{처치})$를 예측하는 방법에 있어 차이를 보이는데, 이러한 차이는 두 모형이 가진 서로 다른 가정에서 비롯된다.

사전, 사후 시점 간 점수 변화에 대한 가정

차이점수 모형은 처치 혹은 통제조건에 노출되었을 때 발생하는 사전, 사후시점 간 변화가 처치집단과 통제집단에 속한 사람들에게 평균적으로 동일하게 발생할 것이라고 가정한다(Maris, 1998). 예를 들어, 자존감 향상 프로그램의 효과성 검증 연구에서, 프로그램을 경험한 처치집단은 자존감 사후점수가 사전점수에 비해 평균적으로 3점 향상되었고, 프로그램을 경험하지 않은 통제집단은 자존감 사후점수가 사전점수와 평균적으로 동일했다고 해보자. 이때 차이점수 모형은, 만약 통제집단에 속한 참여자들이 이 프로그램을 경험했다면 처치집단에 속한 참여자들과 마찬가지로 평균 3점의 자존감 점수 향상을 나타냈을 것이고, 만약 처치집단에 속한 참여자들이 통제조건을 경험했다면 통제집단에 속한 참여자들과 동일하게 자존감 점수에 평균적으로 변화를 보이지 않았을 것이라고 가정한다.

이러한 가정을 수식으로 나타내면 식 (17), (18)과 같다(Maris, 1998). 식 (17)은 처치조건에 노출되었을 때의 사후점수 Y_t와 사전점수 X 간 차이가 처치집단과 통제집단에서 평균적으로 동일하다는 가정을 나타낸다. 마찬가지로, 식 (18)은 통제조건에 노출되었을 때의 사후점수 Y_c와 사전점수 X 간 차이가 두 집단에

서 평균적으로 동일하다는 가정을 나타낸다[3].

$$E(Y_t - X \mid \text{처치}) = E(Y_t - X \mid \text{통제}) \quad (17)$$
$$E(Y_c - X \mid \text{처치}) = E(Y_c - X \mid \text{통제}) \quad (18)$$

식 (17)과 (18)을 풀어서 정리하면, 각각 식 (19), (20)를 얻을 수 있고, 식 (19)와 (20)을 식 (16)에 대입하면, 식 (21)(Maris, 1998)을 얻을 수 있다.

$$E(Y_t \mid \text{통제}) = E(Y_t \mid \text{처치}) - E(X \mid \text{처치}) + E(X \mid \text{통제}) \quad (19)$$
$$E(Y_c \mid \text{처치}) = E(Y_c \mid \text{통제}) - E(X \mid \text{통제}) + E(X \mid \text{처치}) \quad (20)$$

$$E(\Delta) = [E(Y_t \mid \text{처치}) - E(X \mid \text{처치})] - [E(Y_c \mid \text{통제}) - E(X \mid \text{통제})] \quad (21)$$

즉, 차이점수 모형의 가정에 기반하면, 평균 처치효과는 식 (21)에서와 같이 처치집단에서의 사전, 사후점수 변화와 통제집단에서의 사전, 사후점수 변화 간 평균적인 차이와 같다. 그리고, 식 (21)의 평균 처치효과를 추정하는 것은 식 (1)의 차이점수 모형에서 γ_1(즉, 차이점수에서의 집단차)을 추정하는 것과 동일하다.

반면, 공분산분석 모형은 처치 혹은 통제조건에 노출되었을 때 나타나는 사전, 사후 점수 간 '관련성'이 처치집단과 통제집단에 속한 사람들에게 동일하게 나타난다고 가정한다(Maris, 1998). 예를 들어, 자존감 향상 프로그램 효과성 검증 연구에서 처치집단에 속한 참여자들의 경우 사전점수가 3점인 참여자들은 사후점수가 평균적으로 5점으로 향상되었고, 사전점수가 4점인 참여자들은 사후점수가 평균적으로 5.5점으로 향상되었다고 하자. 공분산분석 모형은 만약 통제집단에 속한 참여자들이 이 프로그램을 경험했다면 처치집단에서와 동일하게 사전점수가 3점인 참여들은 사후점수가 5점으로 향상될 것으로 기대되고, 사전점수가 4점인 참여자들은 사후점수가 5.5점이 될 것으로 기대된다고 가정한다. 즉, 사전점수가 동일하다면, 어느 집단에 할당되건 관계없이 처치조건에 노출되었을 때 기대되는 사후점수가 동일하다고 가정하는 것이다. 마찬가지로, 공분산분산 모형은 사전점수가 동일하다면 어느 집단에 할당되건 관계없이 통제조건에 노출되었을 때 기대되는 사후점수 또한 동일하다고 가정한다.

이러한 가정을 수식으로 나타내면 식 (22),

3) 참고로, 식 (17) 혹은 그 이후에 제시되는 수식에서 사전점수 X는 사후점수 Y와 달리 t 혹은 c와 같은 인덱스를 가지고 있지 않다. 본 논문에서 사용한 t 혹은 c 인덱스는 '실제로 할당된 집단'을 나타내는 기호가 아니라(이는 조건부 확률로 표시된다), '노출될 수 있는 조건'을 가리킨다. 즉, Y_{ti}는 i번째 개인이 '처치(t) 조건에 노출되었다면 얻어졌을 점수'이고, Y_{ci}는 i번째 개인이 '통제(c) 조건에 노출되었다면 얻어졌을 점수'를 나타낸다. 즉, 이 값들은 i번째 개인이 어느 조건에 실제로 할당되었는가에 따라 관찰될 수도 있고, 관찰하지 못할 수도 있는 잠재적 결과(potential outcomes)이다. 만약 i번째 개인이 처치 집단에 할당된다면, Y_{ti}는 관찰되지만 Y_{ci}는 관찰되지 않는다. 반대로, i번째 개인이 통제집단에 할당된다면 Y_{ci}는 관찰되지만 Y_{ti}는 관찰되지 않는다. 이와 달리, 사전점수 X_i의 경우 i번째 개인이 어느 조건에 노출되느냐에 따라 그 값이 달라지지 않으므로(조건에 노출되기 이전에 관찰되기 때문에), t 혹은 c 인덱스를 필요로 하지 않는다. 달리 말하면, X_i는 i번째 개인이 실제로 어느 집단에 할당되는가에 관계없이 항상 관찰 가능하다.

(23)과 같다. 여기서 $E(B \mid A, 집단)$은 해당 집단에서 변수 A의 값이 주어졌을 때 변수 B에 대해 기대되는 값을 나타낸다. 즉, 식 (22)는 사전점수 X가 주어졌을 때 기대되는 처치조건에서의 사후점수 Y_t가 처치집단과 통제집단에서 동일함을 나타내며, 식 (23)은 사전점수 X가 주어졌을 때 기대되는 통제조건에서의 사후점수 Y_c가 처치집단과 통제집단에서 동일함을 나타낸다.

$$E(Y_t \mid X, 처치) = E(Y_t \mid X, 통제) \quad (22)$$

$$E(Y_c \mid X, 처치) = E(Y_c \mid X, 통제) \quad (23)$$

일반적으로 공분산분석 모형에서는 사전, 사후 점수 간 관련성이 선형적이라고 가정한다. 이 경우, 식 (22)의 좌변과 우변을 각각 식 (24), (25)와 같은 선형 회귀식으로 나타낼 수 있다(Maris, 1998). 식 (24)와 (25)는 모두 사전점수 X로 처치조건에 노출되었을 때의 사후점수 Y_t를 예측하는 선형 회귀식으로, 식 (24)는 처치집단에서 이를 정의한 것이고 식 (25)는 통제집단에서 이를 정의한 것이다. 이 때 β_t는 X로 Y_t를 예측할 때의 기울기를 나타낸다. 식 (24)와 (25)를 식 (22)에 대입하면 식 (26)(Maris, 1998)을 얻을 수 있다.

$$E(Y_t \mid X, 처치) = E(Y_t \mid 처치) + \beta_t[X - E(X \mid 처치)] \quad (24)$$

$$E(Y_t \mid X, 통제) = E(Y_t \mid 통제) + \beta_t[X - E(X \mid 통제)] \quad (25)$$

$$E(Y_t \mid 통제) = E(Y_t \mid 처치) - \beta_t[E(X \mid 처치) - E(X \mid 통제)] \quad (26)$$

같은 방법으로, 식 (23)의 좌변과 우변을 각각 선형 회귀식으로 나타낸 후 이를 식 (23)에 대입하면 식 (27)(Maris, 1998)을 얻을 수 있다. 이때 β_c는 통제조건에 노출되었을 때의 사후점수 Y_c를 사전점수 X로 예측할 때의 기울기를 나타낸다.

$$E(Y_c \mid 처치) = E(Y_c \mid 통제) - \beta_c[(E(X \mid 통제) - E(X \mid 처치)] \quad (27)$$

마지막으로, 식 (26)과 (27)을 식 (16)에 대입하고, $\beta_t = \beta_c = \beta$라고 가정하면[4], 식 (28)(Maris, 1998)을 얻을 수 있다.

$$E(\Delta) = [E(Y_t \mid 처치) - E(Y_c \mid 통제)] - \beta[E(X \mid 처치) - E(X \mid 통제)] \quad (28)$$

즉, 공분산분석 모형의 가정에 기반하면, 평균 처치효과는 식 (28)에서와 같이 처치집단과 통제집단 간 사후점수의 평균적인 차이에서 두 집단 간 사전점수의 평균 차이로 예측되는 부분을 빼준 값과 동일하다. 그리고, 식 (28)에 기반하여 평균 처치효과를 추정하는 것은 곧 식 (2)의 공분산분석 모형에서 사전점수로는 설명되지 않는 사후점수에서의 집단차 β_2를 추정하는 것과 동일하다.

이와 같이, 차이점수 모형과 공분산분석 모형은 서로 다른 가정에 기반하며, 이로 인해 서로 다른 방식으로 처치효과를 추정한다. 차이점수 모형은 처치 혹은 통제조건에 노출되었을 때 발생하는 사전, 사후점수 간 '차이'가 두 집단에서 동일하게 나타날 것이라고 가정하는 반면, 공분산분석 모형은 처치 혹은 통

4) 이러한 가정이 반드시 필요한 것은 아니며, 여기서는 Maris(1998, p.316)에서와 마찬가지로 논의를 좀 더 단순화하기 위해 이러한 가정을 도입하였다.

제조건에 노출되었을 때 발생하는 사전, 사후 점수 간 '관련성' 즉, 사전점수가 주어졌을 때 기대되는 사후점수가 두 집단에서 동일하게 나타날 것이라고 가정한다. 따라서, 둘 중 어느 모형의 가정이 성립하느냐에 따라 처치효과 추론에 사용해야 할 적절한 모형이 달라진다. 문제는, 어느 모형의 가정이 맞는지 직접적으로 검증하는 것이 불가능하다는 것이다 (Gollwitzer et al., 2014; Holland & Rubin, 1983; Wainer, 1991). 통제집단에 속한 사람들이 처치를 받았다면, 혹은 처치집단에 속한 사람들이 처치를 받지 않았다면 어떤 점수를 얻었을지 관찰하는 것은 불가능하기 때문이다.

영가설 하에서 예측되는 결과 패턴

엄밀한 검증은 아니지만, 어느 모형의 가정이 타당한지 판단하기 위해, 영가설이 참일 때(즉, 평균 처치효과가 존재하지 않을 때) 두 모형의 가정이 각각 어떤 결과 패턴을 예측하는지 살펴보는 것이 도움이 된다.

앞서 설명했듯, 차이점수 모형의 가정이 성립한다면 평균 처치효과를 식 (21)과 같이 나타낼 수 있다. 이때 처치효과가 존재하지 않는다면 식 (21)은 0의 값을 갖게 되고, 이 경우 식 (21)을 식 (29)와 같이 나타낼 수 있다.

$$E(Y_t \mid \text{처치}) - E(Y_c \mid \text{통제}) \\ = E(X \mid \text{처치}) - E(X \mid \text{통제}) \qquad (29)$$

식 (29)가 보여주는 것은, 차이점수 모형의 가정이 성립한다면, 처치효과가 없을 때 처치집단과 통제집단에서 사후점수의 평균 차이는 두 집단 간 사전점수의 평균 차이와 같다는 것이다. 달리 말하면, 차이점수 모형은 평균 처치효과가 없을 때 처치집단과 통제집단 간 사전점수에서의 평균적 차이가 사후점수에서도 그대로 유지될 것으로 예상한다 (Castro-Schilo & Grimm, 2018; Maris, 1998). 그림 2의 패널 (a)는 바로 이러한 패턴을 보여준다. 즉, 두 집단에서 평균 점수가 사전, 사후 시점 간 감소한 폭이 동일하여, 사전점수에서의 집단 차이가 사후점수에서도 동일하게 유지됨을 알 수 있다.

반면, 공분산분석 모형의 가정이 성립한다면, 평균 처치효과는 식 (28)과 같이 나타낼 수 있다. 평균 처치효과가 0일 때식 (28)은 식

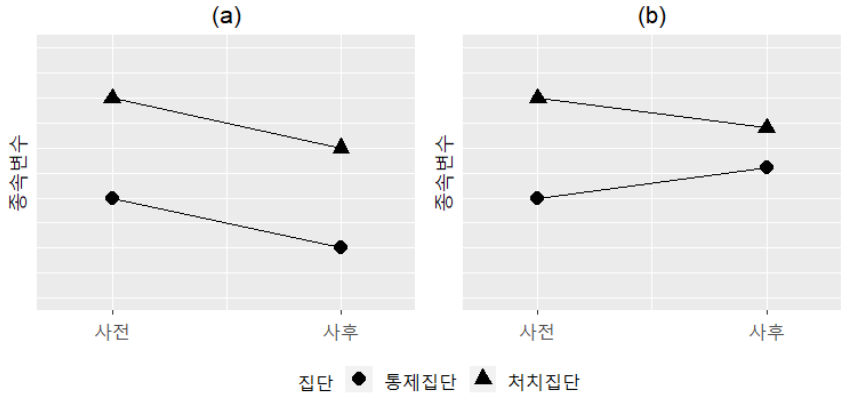


그림 2. 평균 처치효과가 존재하지 않을 때 기대되는 집단별 평균 점수 변화 양상

(30)과 같다.

$$E(Y_t \mid \text{처치}) - E(Y_c \mid \text{통제}) = \beta[E(X \mid \text{처치}) - E(X \mid \text{통제})] \quad (30)$$

식 (30)에서 볼 수 있듯이, 공분산분석 모형의 가정이 성립한다면, 처치집단과 통제집단 간 사후점수에서의 평균 차이는 두 집단 간 사전점수 평균 차이에 β를 곱한 값과 같아진다. 이때 기울기 β는 각 집단 내에서 사전점수에 기반하여 사후점수를 예측할 때의 기울기로, 처치효과가 존재하지 않고 점수가 완벽하게 신뢰롭지 않다면 보통 1보다 작은 양의 값을 나타낸다(Rausch et al., 2003, p.469). β가 1보다 작다면, 두 집단 간 사후점수에서의 평균적인 차이는 사전점수에서의 평균적인 차이보다 더 작아진다. 달리 말하면, 공분산분석 모형은 평균 처치효과가 없을 때 처치집단과 통제집단 간 사전점수에서의 평균적인 차이가 사후점수에서 그대로 유지되는 것이 아니라 줄어들 것으로 예상한다(Castro-Schilo & Grimm, 2018; Maris, 1998). 그림 2의 패널 (b)는 바로 이러한 패턴을 보여준다. 즉, 두 집단에서 평균 점수의 변화 폭이 동일하지 않고, 사전점수에서의 집단 차에 비해 사후점수에서의 집단 차가 더 작은 것을 볼 수 있다.

이와 관련하여, Van Breukelen(2013)은 차이점수 모형과 공분산분석 모형을 식 (31)과 같은 반복측정 모형으로 나타낼 수 있음을 보인 바 있다.

$$Y_{ijt} = \theta_0 + \theta_1 G_{ij} + \theta_2 T_{it} + \theta_3 G_{ij} T_{it} + e_{ijt} \quad (31)$$

식 (31)에서 Y_{ijt}는 j번째 집단(통제집단의 경우 j=1, 처치집단의 경우 j=2)에 속한 i번째 참여자를 t번째 시점(사전시점의 경우 t=1, 사후시점의 경우 t=2)에 측정한 값을 나타낸다[5]. G_{ij}는 집단을 나타내는 더미변수로, i번째 참여자가 통제집단에 속한 경우 0의 값을, 처치집단에 속한 경우 1의 값을 갖는다. T_{it}는 시점을 나타내는 더미변수로, Y_{ijt}가 사전점수를 나타낼 경우 0의 값을, 사후점수를 나타낼 경우 1의 값을 갖는다. e_{ijt}는 잔차로, 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정한다.

식 (31)에서 절편인 θ_0은 $G_{ij}=0$이고 $T_{it}=0$일 때 기대되는 Y_{ijt}값 즉, 통제집단의 사전점수 평균을 나타낸다. G_{ij}의 기울기인 θ_1은 $T_{it}=0$일 때 G_{ij}가 0에서 1로 변화하면 Y_{ijt}가 평균적으로 얼마나 달라지는지 즉, 사전점수에서 통제집단과 처치집단 간 평균 차이를 나타낸다. T_{it}의 기울기인 θ_2는 $G_{ij}=0$일 때 T_{it}가 0에서 1로 변화하면 Y_{ijt}가 평균적으로 얼마나 변화하는지 즉, 통제집단에서의 평균 사전-사후 변화를 나타낸다. G_{ij}와 T_{it}의 상호작용 기울기인 θ_3는 G_{ij}가 0에서 1로 변화할 때 T_{it}의 기울기가 얼마나 변화하는지 즉, 평균 사전-사후 변화에서의 집단차를 나타낸다. 따라서, 식 (31)에서 θ_3와 식 (1)의 γ_1은 동일하게 사전-사후 변화의 집단차를 나타내며, 식 (31)에서 θ_3를 추정하는 것은 곧 차이점수 모형에 기반하여 처치효과를 추정하는 것과 같다.

Van Breukelen(2013)은 수식과 예시를 통해

5) 지금까지 사전점수를 X_i와 같이 표기하였으나, 차이점수 모형과 공분산분석 모형을 반복측정 모형으로 나타낼 경우에는 사전점수가 Y_{ij1}와 같이 표기된다.

식 (31)에서 θ_1이 0의 값을 가질 때 이 모형이 공분산분석 모형과 동일해짐을 증명하였다. 이때 θ_1이 0이라는 것은 사전점수에서 통제집단과 처치집단 간 평균 차이가 존재하지 않음을 의미한다. 즉, 공분산분석 모형은 모집단 수준에서 사전점수에 집단차가 존재하지 않는다는 가정을 내포하고 있음을 알 수 있다. 이러한 공분산분석의 가정은 참여자를 통제집단과 처치집단으로 할당하는 절차가 사전점수 측정 이후에 시행되어, 사전점수를 측정하는 시점에는 오직 하나의 집단만 존재하는 경우 성립한다(Van Breukelen, 2013). 또한, 이러한 가정은 영가설이 참일 때 공분산분석이 예측하는 결과 패턴과도 일맥상통한다. 사전점수 측정 시점에 하나의 모집단에 속했던 참여자들이 이후 통제집단과 처치집단으로 구분된 것이라면, 사전점수에 나타난 집단간 표본평균차는 집단의 이질적 특성을 반영하는 것이 아니라, 측정오차로 인한 우연한 차이를 반영하는 것이라 할 수 있다. 따라서, 평균 처치효과가 존재하지 않는다면, 사전 시점에 나타난 우연에 의한 집단차는 사후 시점에 그대로 유지되기보다 줄어들 것으로 기대할 수 있다(Van Breukelen, 2013).

두 모형이 영가설 하에서 예측하는 결과 패턴이 다르다는 것에 기반하여 Lord의 역설을 설명할 수도 있다(Castro-Schilo & Grimm, 2018). 앞서 그림 1에 제시된 예시에서, 차이점수 모형은 변화의 집단차가 유의하지 않다는 결과를 산출하였으나, 공분산분석 모형은 유의한 결과를 산출하였다. 그림 1에서 보여지는 패턴은 사전점수에서의 집단 차가 사후점수에서 거의 그대로 유지되거나 혹은 아주 약간 증가하는 모습이라고 할 수 있다. 이러한 패턴은 차이점수 모형이 영가설 하에서 예측하는 패턴 즉, 그림 2의 패널 (a)와 유사하며, 따라서 차이점수 모형을 사용하면 영가설을 기각하지 못하는 결과를 얻게 된다. 반면, 공분산분석 모형의 경우에는 영가설이 참일 때 그림 2의 패널 (b)와 같이 사후점수에서 집단 차가 감소할 것으로 예측하는데, 그림 1의 패턴은 이러한 예측에 비해 사후점수에서의 집단 차가 더 크게 벌어진 것이라 할 수 있으며, 따라서 영가설을 기각하는 결과를 산출한 것이라고 이해할 수 있다.

적절한 분석 모형 선택을 위한 가이드라인

지금까지 인과추론 맥락에서 차이점수 모형과 공분산분석 모형이 가지고 있는 서로 다른 가정에 대해 자세히 살펴보았다. 어느 모형의 가정이 성립하는가는 실험 설계 방법 즉, 연구 참여자를 처치집단과 통제집단에 어떻게 할당하는가에 따라 달라진다(Kenny, 1975; Van Breukelen, 2013). 따라서, 앞서 살펴본 두 모형의 가정과 함께, 여러 시뮬레이션 연구 결과를 종합적으로 고려하여, 실험 설계 방법에 따라 어느 모형을 사용하는 것이 적절한지 가이드라인을 제시해보도록 하겠다.

무선할당

참여자를 통제집단과 처치집단에 무선적으로 할당하는 실험 연구에서는 차이점수 모형과 공분산분석 모형 모두 편향되지 않은 평균 처치효과 추정치를 제공하는 것으로 알려져 있다(Van Breukelen, 2006, 2013). 무선할당 실험의 경우, 두 집단에 속한 참여자들 간에 체

계적 차이가 존재하지 않고, 따라서 사전점수에 집단차가 없을 것으로 기대할 수 있다. 이는 곧 $E(X \mid \text{처치}) = E(X \mid \text{통제})$성립함을 의미하며, 이 경우 차이점수 모형의 가정에 기반하여 평균 처치효과를 정의한 식 (21)과 공분산분석 모형의 가정에 기반하여 평균 처치효과를 정의한 식 (28)이 동일해짐을 알 수 있다.

그러나, 차이점수 모형과 공분산분석 모형은 서로 다른 검증력을 나타낸다. 무선할당 실험에서 처치효과를 검증할 때, 차이점수 모형에 비해 공분산분석 모형의 검증력이 더 높은 것으로 알려져 있다(Huck & McLean, 1975). Petscher와 Schatschneider(2011)는 시뮬레이션 연구를 통해 무선할당 실험에서 두 모형의 수행을 비교하였는데, 표본크기, 점수의 정규성, 사전, 사후점수 간 상관, 사후점수 분산 등에 따라 달라지는 모든 조건 하에서 공분산분석 모형의 검증력이 차이점수 모형에 비해 일관되게 높은 것으로 나타났다. 다른 시뮬레이션 연구들에서도 무선할당 실험에서와 같이 사전점수의 집단차가 존재하지 않는 조건에서는 점수가 완벽하게 신뢰로운 경우를 제외하면 공분산분석 모형이 차이점수 모형에 비해 더 높은 검증력을 나타냈다(Jennings & Cribbie, 2016; Kisbu- Sakarya et al., 2013).

따라서, 무선할당 실험에서 처치효과 추정을 위해서는 차이점수 모형과 공분산분석 모형 모두 사용해도 무방하나, 처치효과에 대한 검증력을 최대한 확보하기 위해서는 공분산분석 모형을 사용할 것이 권장된다.

비무선할당 1: 사전 점수에 기반한 집단 할당

무선할당이 아닌 모든 집단 할당 방법은 비무선할당(nonrandom assignment)으로 구분된다. 비무선할당 중 처치효과 추론을 위한 실험 연구에서 종종 사용되는 방법의 하나가 바로 사전점수에 기반하여 참여자를 통제집단과 처치집단에 할당하는 방법이다. 예를 들어, 자존감 향상 프로그램의 효과성을 연구할 때, 이 프로그램은 자존감 수준이 높은 사람들보다는 낮은 사람들에게 더 필요할 것이므로, 사전 자존감 점수가 기준 점수보다 높은 사람들을 통제집단에, 기준 점수보다 낮은 사람들을 처치집단에 할당할 수 있다. 이렇게 사전점수의 특정 값을 기준으로 실험집단에 참여자를 할당하는 방법을 회귀 불연속 설계(regression discontinuity design; Shadish et al., 2002)라 한다.

이 경우, 사전점수에 기반하여 구분된 처치집단과 통제집단은 원래 서로 다른 이질적 모집단에 속한 사람들이 아니라, 하나의 모집단에 속한 사람들을 사전점수에 따라 구분하여 생성된 것이다. 즉, 집단 구분이 사전점수 측정 이후에 실시되었으므로, 모집단 수준에서 사전점수에 집단차가 존재하지 않는다는 공분산분석의 가정이 성립한다고 할 수 있다.

회귀 불연속 설계의 경우, 기준점보다 높은 사전점수를 가진 사람들을 하나의 집단으로, 기준점보다 낮은 사전점수를 가진 사람들을 다른 집단으로 구분하기 때문에, 이 결과로 얻어진 표본에서는 당연히 두 집단 간 사전점수 평균에 차이가 존재할 수밖에 없다. 때문에, 회귀 불연속 설계에서 어떻게 사전점수에 집단 차가 존재하지 않는다는 가정이 성립한다는 것인지 다소 의아하게 생각할 수도 있을 것이다.

그러나, 회귀 불연속 설계에서 관찰되는 표본 사전점수에서의 집단 차는 하나의 동질적 모집단으로부터 얻어진 표본 점수들을 기준점

을 중심으로 인위적으로 두 집단으로 구분함에 따라 발생할 수밖에 없는 집단 차이로, 이는 평균이 서로 다른 두 이질적 모집단으로부터 점수들을 표집했기 때문에 발생하는 체계적 집단 차이와는 구분되어야 한다. 즉, 표본에서 집단 간 평균 차이가 관찰된다고 하더라도, 이것이 항상 모집단에서의 평균 차이를 반영하는 것은 아니다. 모집단 사전점수에 집단 차가 존재하는가 그렇지 않은가(즉, 두 집단의 점수가 이질적 모집단으로부터 나왔는가 그렇지 않은가)에 대한 가정은 표본에서의 사전점수 평균 차이 유무로 결정되는 것이 아니라 집단 할당 메커니즘에 따라 결정된다.

Maris(1998)는 회귀 불연속 설계를 포함하여, 참여자를 실험집단에 할당할 때 사전점수에 기반하여 확률적으로 할당하는 경우(예를 들어, 사전점수가 높을수록 처치집단에 할당될 확률이 증가하도록 설계한 경우) 공분산분석 모형의 가정이 성립한다는 것을 이론적으로 보인 바 있다. Jennings와 Cribbie(2016)는 시뮬레이션 연구를 통해, 참여자를 사전점수에 기반하여 실험집단에 할당할 경우, 공분산분석 모형은 처치효과를 추정함에 있어 거의 편향을 나타내지 않은 반면, 차이점수 모형은 상대적으로 높은 편향과 제1종 오류율을 나타냄을 보였다. Wright(2006)의 시뮬레이션 연구에서도 사전점수에 기반해 실험조건에 참여자를 할당한 경우, 공분산분석은 점수의 신뢰도와 관계없이 항상 편향되지 않은 결과를 산출한 반면, 차이점수 모형은 점수의 신뢰도가 1인 경우에만 편향되지 않은 결과를 산출하고, 신뢰도가 낮아질수록 점점 더 편향된 결과를 산출하는 것으로 나타났다.

이러한 결과를 종합하면, 사전점수에 따라 참여자를 실험집단에 할당하는 실험 연구에서 처치효과를 추정하고자 할 때에는 공분산분석 모형을 사용하는 것이 적절하며, 차이점수 모형을 사용하는 것은 권장되지 않는다.

비무선할당 2: 비동질적 집단 설계

윤리적, 현실적 제약으로 인해 연구 참여자들을 실험집단에 무선적으로 할당할 수 없는 경우, 연구자들은 종종 비동질적 집단 설계(nonequivalent group design; Shadish et al., 2002)를 사용한다. 비동질적 집단 설계는 비무선할당 설계의 하나로, 서로 다른 이질적 집단에 속하는 참여자들을 처치집단과 통제집단에 할당하는 것을 뜻한다. 예를 들어, 참여자가 자신의 선호에 따라 처치집단과 통제집단 중 하나를 선택하거나, 참여자가 특정 속성[6]을 가지고 있는가를 관찰하여 이에 따라 참여자를 처치집단 혹은 통제집단에 할당하는 경우가 이에 해당된다. 앞서 목표달성이 기본심리욕구 충족 수준에 미치는 영향을 분석한 예시의 경우, 목표달성 정도를 측정하여 이에 따라 목표달성 고집단과 저집단을 구분했는데, 이 경우도 비동질적 집단 설계에 해당된다.

비동질적 집단 설계의 핵심적인 특징은 처치집단과 통제집단이 비동질적인 개인들로 이루어져 있어, 무선할당 실험에서와 달리 사전점수에 체계적인 집단 차이 즉, 모집단 수준에서의 집단 차이가 존재할 수 있다는 것이다. 바로 이 점에서, 비동질적 집단 설계는 회귀 불연속 설계와도 구분된다. 비동질적 집단

6) 여기서 말하는 특정 속성은 사전점수가 아닌 다른 속성을 의미한다. 만약 사전점수에 기반하여 참여자를 처치집단과 통제집단으로 할당한다면, 이는 비동질적 집단 설계가 아니라 앞서 언급한 회귀 불연속 설계에 해당된다.

설계와 회귀 불연속 설계는 모두 비무선할당 설계에 해당되지만, 회귀 불연속 설계에서는 동질적 모집단에 속한 개인들을 사전점수에 따라 인위적으로 두 집단으로 구분하여 할당하는 반면, 비동질적 집단 설계에서는 원래 서로 다른 이질적인 모집단에 속한 개인들을 처치집단과 통제집단에 할당한다.

공분산분석은 사전점수에 집단차가 존재하지 않으며, 사전점수 측정 시점에 처치집단과 통제집단은 동질적인 하나의 집단이었다는 가정을 내포하는데, 비동질적 집단 설계는 이러한 가정과 완전히 배치된다. 따라서, 비동질적 집단을 비교하여 처치효과를 추정할 때 공분산분석 모형을 사용하게 되면 부정확하고 편향된 결과를 얻을 수 있다(Maris, 1998; Van Breukelen, 2006).

Casto-Schilo와 Grimm(2018)은 간단한 시뮬레이션 연구를 통해 공분산분석 모형의 결과가 어떻게 편향될 수 있는지 보여주었다. 이들은 처치효과가 없는 상황을 가정하고, 처치집단과 통제집단 모두 사전, 사후점수에 변화가 없도록 자료를 생성하였다. 이때 사전점수에 집단 간 차이가 있는 첫 번째 조건과 사전점수에 집단 간 차이가 없는 두 번째 조건을 구분하였다. 첫 번째 조건은 차이점수 모형의 가정(사전점수의 집단차가 사후점수에 그대로 유지된다)에 부합하며, 두 번째 조건은 공분산분석 모형의 가정(사전점수에 집단차가 존재하지 않는다)에 부합한다. 이렇게 생성된 자료를 차이점수 모형과 공분산분석 모형으로 분석한 결과, 차이점수 모형은 두 조건 모두에서 처치효과가 유의하지 않다는 올바른 결론을 도출한 반면, 공분산분석 모형은 조건에 따라 다른 결과를 보였다. 사전점수에 집단차가 없는 두 번째 조건의 자료를 공분산분석 모형으로 분석했을 때는 처치효과가 유의하지 않았으나, 사전점수에 집단차가 존재하는 첫 번째 조건의 자료를 분석했을 때는 처치효과가 유의하다는 부정확한 결과를 도출하였다.

차이점수 모형과 공분산분석 모형의 수행을 비교한 다른 시뮬레이션 연구들에서도 사전점수에 집단차가 존재하고 사전점수의 신뢰도가 1보다 작을 때는 공분산분석 모형이 차이점수 모형에 비해 더 편향된 결과를 산출하는 것으로 나타났다(Jennings & Cribbie, 2016, Table 5). 차이점수 모형의 경우 사전점수의 신뢰도에 따라 제1종 오류율이 달라지지 않고 연구자가 설정한 유의수준과 유사한 정도의 오류율을 일관되게 유지했으나, 공분산분석 모형은 사전점수의 신뢰도가 낮아질수록 제1종 오류율이 증가하였다(Jennings & Cribbie, 2016, Table 6; Kisbu-Sakarya et al., 2013).

이러한 연구 결과를 종합하면, 비동질적 집단을 비교하여 처치효과를 추정하는 상황에서는 공분산분석 모형보다 차이점수 모형을 사용하는 것이 좀 더 적절하다고 할 수 있다(Casto-Schilo & Grimm, 2018). 그러나, 비동질적 집단 설계에서 차이점수 모형이 반드시 편향되지 않은 정확한 처치효과 추정치를 제공하리라는 보장은 없다. 차이점수 모형에서 가정하는 것처럼 처치효과가 없을 때 사전점수에 존재하는 집단차가 사후점수에서 그대로 유지되어야 할 필연적인 이유는 존재하지 않으며(Maris, 1998), 집단간 점수차에 이러한 안정성이 존재하리라는 것은 사실상 매우 강한 가정이기 때문이다(Van Breukelen, 2006, 2013). 차이점수 모형과 공분산분석 모형을 모두 적용해서 결과를 비교했을 때, 두 모형이 크기만 다소 다르고 동일한 방향의 처치효과를 추정한다면, 결과에 대한 확신이 증가할 수는 있

겠지만, 그렇다고 해서 이것이 곧 정확한 결과임을 보장해주는 것은 아니다(Van Breukelen, 2006, 2013).

결국, 차이점수 모형 혹은 공분산분석 모형을 사용해서 처치효과를 정확히 추정하기 위해서는 무선할당 혹은 사전점수에 기반한 실험집단 할당 방법을 사용해야 한다. 만약 비동질적 집단 설계를 사용할 수밖에 없는 상황이라면, 두 개 이상의 통제집단을 확보하거나 두 번 이상의 사전 점수를 측정하여 차이점수 모형의 가정이 성립하는지 살펴보는 것이 도움이 된다(Van Breukelen, 2006). 만약 두 통제집단에서 사전, 사후시점 간 동일한 변화를 보이거나, 통제집단과 처치집단이 두 사전시점 간에 동일한 정도의 차이를 보인다면, 이는 처치효과가 없을 때 사전점수의 집단차가 사후점수에서도 그대로 유지된다는 차이점수 모형의 가정에 부합하는 것이라 할 수 있다. 물론, 이것이 곧 차이점수 모형의 가정이 성립함을 입증한 것은 아니지만, 결과의 정확성에 대해 좀 더 확신을 가질 수는 있다.

비동질적 집단 설계를 사용해야 하는 경우, 사전시점에 참여자들로부터 사전점수 뿐만 아니라 다수의 다른 속성들도 함께 측정할 수 있다면, 성향 점수(propensity score; Rosenbaum & Rubin, 1983)를 사용하여 처치효과를 추론하는 것도 가능하다. 성향 점수 분석에 대한 자세한 논의는 본 논문의 범위를 벗어나므로, 관심있는 독자들은 West와 동료들(2014), 그리고 Kim(2019)을 참고할 것을 권한다.

지금까지 인과추론 맥락에서 정확한 처치효과 추론을 위한 분석 방법 가이드라인을 제시하였다. 이를 간단히 요약하면 표 1과 같다.

비동질적 집단에 기반한 관련성 분석

마지막으로, 인과추론이 아닌, 단순히 변화와의 관련성 분석을 위해 비동질적 집단을 비교하는 경우에 대해 추가적으로 고려해 보고자 한다. 변화의 집단차를 살펴보는 연구들 중에는 처치효과 입증을 목적으로 하는 것도 있지만, 단순히 변화와 집단 간 관련성을 살펴보는 것을 목적으로 하는 경우도 있기 때문이다.

처치효과 추론이 '처치에 의해 처치집단과 통제집단 간 변화에 차이가 나타나는가?'와

표 1. 인과추론 분석 방법에 대한 가이드라인

집단 할당 방법		분석 방법	
		차이점수 모형	공분산분석 모형
무선 할당		• 적절함	• 적절함 • 검증력 측면에서 차이점수 모형보다 권장됨
비무선 할당	사전 점수에 기반한 집단 할당	• 부적절함	• 적절함
	비동질적 집단 설계	• 경우에 따라 적절할 수 있음 • 그러나, 정확한 처치효과 추론을 보장하지는 않음	• 부적절함

같은 질문에 답하기 위한 것이라면, 변화와의 관련성(correlate of change) 분석은 '누가 더 많이 변하는가?'와 같은 질문에 답하기 위한 것이라고 할 수 있다. 이러한 관련성 분석을 위해서는 차이점수 모형과 공분산분석 모형에서 다루는 '변화'가 서로 다른 의미를 갖는다는 점에 주목할 필요가 있다. 구체적으로, 차이점수 모형은 '어느 집단이 더 많은 사전, 사후 시점 간 점수 차이를 나타냈는가?'와 같은 질문에 답을 제공한다면, 공분산분석 모형은 '만약 두 집단이 동일한 사전점수를 가지고 있었다면, 어느 집단이 더 많은 사전, 사후시점 간 점수 차이를 나타냈겠는가?'와 같은 질문에 답을 제공한다(Kisbu-Sakarya et al., 2013). 따라서, 둘 중 어느 질문이 연구 맥락에 보다 적절한가, 혹은 '두 집단이 동일한 사전점수를 가지고 있었다면'이라는 가정이 합당한가에 따라 분석 방법을 선택할 수 있다.

다만, 공분산분석 모형을 사용할 때, 사전점수에 측정오차가 개입되어 있으면 집단과 변화와의 관련성이 편향되어 추정되므로 주의가 필요하다. Culpepper와 Aguinis(2011)는 공분산분석 모형에서 집단변수의 기울기(이때 집단변수는 더미변수이고, 처치집단은 1, 통제집단은 0의 값을 가진다)를 추정할 때 발생하는 편향의 정도를 수식으로 제시하였고, Miyazaki와 동료들(2022)은 편향의 정도 와 방향을 함께 살펴보기 위해 근사 편향(asymptotic bias) 값을 수식으로 제시하였다[7]. 이들이 제시한 수식을 살펴보면, 사전점수를 측정오차 없이 완벽하게 신뢰롭게 측정할 수 있거나, 사전점수에 집단차가 존재하지 않으면, 공분산분석에서 집단변수 기울기를 편향없이 추정할 수 있다. 그러나, 사전점수의 신뢰도가 1보다 작고, 사전점수에 집단차가 존재하는 경우에는 집단변수 기울기가 편향되어 추정된다. Miyazaki와 동료들(2022), 그리고 Jamieson(1994, 1999)은 시뮬레이션 연구를 통해 이러한 편향이 실제로 발생한다는 것을 경험적으로 보여주었다.

집단변수의 기울기가 편향되어 추정될 때 편향의 방향은 사전, 사후점수 간 상관의 부호와 사전점수(보다 정확히는 사전 진점수)에서의 집단차 방향에 따라 결정된다(Miyazaki et al., 2022). 사전, 사후점수가 종종 그러하듯 정적 상관을 나타낸다고 가정했을 때, 만약 처치집단의 사전점수가 통제집단보다 높으면 집단변수의 기울기는 정적으로 편향되고, 처치집단의 사전점수가 통제집단보다 낮으면 집단변수의 기울기는 부적으로 편향되어 추정된다[8]. 따라서, 집단과 변화 간 관련성이 사전

7) Culpepper와 Aguinis(2011), 그리고 Miyazaki와 동료들(2022)은 공분산분석 모형을 참모형(true model)이라고 가정하고 모든 논의를 전개하였고, 공분산분석 모형에서 집단변수의 기울기를 처치효과라고 명명하였다. 그런데, 이들은 이와 동시에 사전점수에 집단차가 존재하는 조건을 연구에 포함시켰다. 앞서 설명했듯 사전점수에 집단차가 존재하는 것은 인과추론의 맥락에서 공분산분석 모형의 가정과 배치된다. 때문에, 이 경우 공분산분석 모형에서 집단변수의 기울기가 곧 인과추론에서의 '평균 처치효과'를 가리킨다고 볼 수 없다. 따라서, 본 논문에서는 이들이 사용한 처치효과라는 표현이 인과추론에서의 평균 처치효과가 아니라 집단과 변화와의 관련성을 가리킨다고 보았다.

8) 만약 사전, 사후점수가 부적 상관을 나타내면 반대 방향의 편향이 발생한다. 이 경우, 처치집단의 사전점수가 통제집단보다 높으면 집단변수 기울기가 부적으로 편향되고, 처치집단의 사전점수가 통제집단보다 낮으면 집단변수 기울기는 정적으로 편향되어 추정된다.

점수의 집단차 방향과 일치할 경우(예를 들어, 처치집단이 더 큰 폭의 점수 향상을 보이고, 사전점수도 처치집단에서 더 높을 때)에는 이러한 관련성이 과대추정되고, 이로 인해 검증력이 높아지면서 유의한 결과를 더 쉽게 얻게 된다. 반대로, 집단과 변화 간 관련성이 사전점수의 집단차 방향과 일치하지 않을 경우(예를 들어, 처치집단이 더 큰 폭의 점수 향상을 보이지만, 사전점수는 통제집단에서 더 높을 때)에는 이러한 관련성이 과소추정되고, 이로 인해 검증력이 낮아지면서 유의한 결과를 얻기가 더 어려워진다(Jamieson, 1994, 1999).

앞서 언급한 것과 같이, 공분산분석 모형은 사전점수가 완벽하게 신뢰롭다면 사전점수에 집단차가 존재하더라도 집단과 변화 간 관련성을 편향되지 않게 추정할 수 있다(Culpepper & Aguinis, 2011; Miyazaki et al., 2022). 따라서, 구조방정식을 사용하여 측정오차가 제거된 사전점수를 분석에 사용할 경우, 편향없이 관련성을 추정하는 것이 가능하다(Miyazaki et al., 2002).

결론 및 논의

지금까지 차이점수 모형과 공분산분석 모형 중 어느 모형을 언제 사용하는 것이 적절한지 알아보기 위해 다양한 선행 연구들을 개관하였다. 우선, 차이점수 사용 자체를 비판하는 주장의 근거를 살펴보고, 이러한 근거가 매우 제한된 가정 하에서만 성립함을 확인하였다. 다음으로, 차이점수 모형과 공분산분석 모형을 이론적, 경험적으로 비교한 연구를 개관하고, 이에 기반하여 인과추론의 맥락에서 적절한 모형 선택을 위한 가이드라인을 도출하였다. 이러한 가이드라인에 기반하여 앞으로 보다 많은 연구자들이 적절한 분석 방법을 선택하고, 동시에 분석 방법 선택의 근거를 논문에 제시한다면, 심리학 연구의 타당성과 투명성이 더욱 제고될 수 있을 것으로 기대된다.

인과추론 맥락에서 차이점수 모형과 공분산분석 모형은 무선할당 실험이 아닌 경우 서로 다른 결과를 산출할 수 있다. 비무선할당 연구에서 두 모형의 가장 큰 차이는 처치효과가 없을 때 점수 변화가 어떤 패턴을 보일 것이라고 예측하는가이다. 차이점수 모형은 사전점수의 집단차가 사후점수에서 그대로 유지될 것이라고 예측한다. 이는 처치집단과 통제집단이 서로 이질적인 모집단에 속한 개인들로 구성되었다는 가정을 내포한다. 반면, 공분산분석 모형은 사전점수의 집단차가 사후점수에서는 줄어들 것이라고 예측한다. 이는 처치집단과 통제집단이 동질적인 하나의 모집단에 속한 개인들로 구성되었다는 가정을 내포한다. 서로 다른 두 가정 중 어떤 가정이 성립하느냐에 따라 적절한 분석 방법이 달라지고, 두 가정이 모두 성립하지 않을 경우 두 모형 모두 편향된 결과를 산출할 수 있다.

흔히 연구자들은 공분산분석 모형을 사용하면 사전점수에 존재하는 집단차를 통제하고 변화의 집단차를 살펴볼 수 있기 때문에, 사전점수에 집단차가 존재하는 상황에서는 공분산분석 모형을 사용하는 것이 적절하다고 생각하곤 한다. 그러나, 인과추론 맥락에서 공분산분석은 모집단 수준에서 사전점수에 집단차가 존재하지 않는다는 가정을 내포하고 있다. 때문에, 회귀 불연속 설계에서와 같이 하나의 모집단에 속한 점수들을 기준점을 중심으로 인위적으로 구분하여 집단 할당이 이루어진 경우가 아니라면, 서로 다른 이질적 모집단으

로부터 나온 개인들로 구성된 집단 간에 사전 점수에 체계적 차이가 존재할 때 평균 처치효과 추론을 위해 공분산분석을 사용하는 것은 오히려 부적절하다. 만약, 집단과 변화 간 관련성을 분석하는 것이 연구의 목적이고, '두 집단이 사전시점에서 동일한 점수를 가졌다면'이라는 가정이 합당하다면, 공분산분석 모형을 사용할 수 있다.

본 연구의 제한점은 다음과 같다. 우선, 본 연구에서는 실험에 사용된 집단이 두 개인 경우만을 고려하였다. 그러나, 둘 이상의 통제(혹은 처치) 조건을 고려하는 실험에서와 같이 세 개 이상의 집단을 비교하는 경우도 존재한다. 이때 차이점수 모형과 공분산분석 모형을 사용해서 분석을 수행하려면, 식 (1)과 (2)에 추가적으로 더미변수를 투입해야 한다. 일반적으로, 집단이 G개일 때는 $G-1$개의 더미변수를 모형에 독립변수로 투입해야 하며, 각 더미변수의 기울기는 특정한 두 집단(보다 구체적으로, 연구자가 정한 기준 집단과 해당 더미변수가 나타내는 특정 집단) 간 평균 차이를 나타낸다. 더미변수를 사용하여 셋 이상의 집단을 비교하는 방법에 대한 보다 자세한 내용은 Cohen과 동료들(2003)의 8장을 참고할 것을 권한다.

다음으로, 본 연구는 두 모형을 비교함에 있어 사전, 사후 두 번의 측정 시점만을 고려하였다. 그러나, Rogosa와 Willett(1985)가 지적했듯 측정 시점이 오직 두 번 뿐인 경우 변화의 양상을 매우 제한적으로 분석할 수밖에 없고, 이러한 분석 결과는 측정 시점의 수가 증가하거나 초기값을 어느 시점으로 두느냐에 따라 완전히 달라질 수 있다. 많은 연구자들이 제한된 자원과 편의를 고려하여 사전-사후 시점 설계를 사용하고 있으나, 이에 기반하여 도출된 분석 결과는 사전, 사후시점 간 간격이 달라지거나 사전시점을 언제로 설정하느냐에 따라 일반화되지 않을 수 있음을 인지할 필요가 있다.

마지막으로, 본 연구는 관찰된 사전, 사후점수에 기반한 가장 단순한 형태의 차이점수 모형과 공분산분석 모형만을 비교했지만, 최근 이 모형들은 잠재변수를 사용한 구조방정식 모형과 결합되면서 보다 복잡하고 발전된 형태로 사용되고 있다. 잠재 변화점수 모형(latent change score model; Casto-Schilo & Grimm, 2018; McArdle, 2009)은 사전, 사후점수 간 차이를 잠재변수로 설정하여, 차이점수에서 측정 오차를 제거하고 분석하는 것을 가능하게 한다. 마찬가지로, 잠재변수를 사용한 자기회귀 모형(McArdle, 2009)은 사전, 사후점수에서 측정 오차를 제거하고 공분산분석을 수행하는 것을 가능하게 한다. 최근에는 잠재변수를 사용한 차이점수 모형과 공분산분석 모형의 수행을 비교한 연구도 수행되었다(Köhler et al., 2021). 평균 처치효과의 존재 유무 뿐만 아니라 처치효과의 메커니즘을 밝히기 위한 매개모형 또한 널리 사용되고 있는데, 이때 매개변수와 종속변수를 차이점수 혹은 잔차점수(사전점수로 예측되지 않는 사후점수) 등과 같이 다양한 형태로 사용할 수 있으며, 관찰변수가 아닌 잠재변수로 이 변수들을 정의하고 있다(Valente & MacKinnon, 2017).

이렇듯, 차이점수와 공분산분석에 기반한 모형들이 보다 복잡하고 다양한 형태로 발전하면서, 언제 어느 모형을 사용하는 것이 적절한지 판단하는 것은 점점 더 중요해지고 있다. 따라서, 이러한 분석 방법들의 목적과 가정을 명확히 이해하고, 다양한 조건에서의 수행 차이를 비교하는 체계적인 방법론 연구들

이 앞으로 지속되어야 할 것이다.

참고문헌

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, *35*(1), 32 – 58. https://doi.org/10.1177/0265407517718387

Chiou, J., & Spreng, R. A. (1996). The Reliability of Difference Scores: A Re-Examination. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior*, *9*, 158 – 167. https://jcsdcb.com/index.php/JCSDCB/article/view/530

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203774441

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"-or should we? *Psychological Bulletin*, *74*(1), 68 – 80. https://doi.org/10.1037/h0029382

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, *16*(2), 166-178. https://doi.org/10.1037/a0023355

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology*, *8*(2), 172-179. https://doi.org/10.1037/h0034145

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263. https://doi.org/10.2307/2841583

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology*, *44*(7), 673-682. https://doi.org/10.1002/ejsp.2042

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley & Sons Inc. https://doi.org/10.1037/13240-000

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945-960. https://doi.org/10.2307/2289064

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3-25). Laurence Erlbaum Associates. https://doi.org/10.4324/9780203056653

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, *82*, 511-518. https://doi.org/10.1037/h0076767

Jamieson, J. (1994). Correlates of reactivity: Problems with regression based methods. *International Journal of Psychophysiology*, *17*(1), 73-78. https://doi.org/10.1016/0167-8760(94)90057-4

Jamieson, J. (1999). Dealing with baseline differences: Two principles and two dilemmas. *International Journal of Psychophysiology*, *31*(2),

155-161.
https://doi.org/10.1016/S0167-8760(98)00048-8

Jennings, M. A., & Cribbie, R. A. (2016). Comparing Pre-Post Change Across Groups: Guidelines for Choosing between Difference Scores, ANCOVA, and Residual Change Scores. *Journal of Data Science*, *14*(2), 205-230. https://doi.org/10.6339/JDS.201604_14(2).0002

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, *82*(3), 345 – 362. https://doi.org/10.1037/0033-2909.82.3.345

Kim, H. (2019). Propensity Score Analysis in Non-Randomized Experimental Designs: An Overview and a Tutorial Using R Software. *New Directions for Child and Adolescent Development*, *2019*(167), 65-89. https://doi.org/10.1002/cad.20309

Kim, S., & Park, S. W. (2019). An Exploratory Study on the Effectiveness of Educational Donation Programs for Middle School Students. *Korean Journal of Psychology: General*, *38*(3), 301-322. https://doi.org/10.22257/kjp.2019.09.38.3.301

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A Monte Carlo Comparison Study of the Power of the Analysis of Covariance, Simple Difference, and Residual Change Scores in Testing Two-Wave Data. *Educational and Psychological Measurement*, *73*(1), 47-62. https://doi.org/10.1177/0013164412450574

Köhler, C., Hartig, J., & Schmid, C. (2021). Deciding between the Covariance Analytical Approach and the Change-Score Approach in Two Wave Panel Data. *Multivariate Behavioral Research*, *56*(3), 447-458. https://doi.org/10.1080/00273171.2020.1726723

Lee, M-H., & Kim, A. (2008). Development and Construct Validation of the Basic Psychological Needs Scale for Korean Adolescents : Based on the Self-Determination Theory. Korean Journal of Social and Personality Psychology, *22*(4), 157-174. https://doi.org/0.21193/kjspp.2008.22.4.010

Linn, R. L., & Slinde, J. A. (1977). The Determination of the Significance of Change Between Pre- and Posttesting Periods. *Review of Educational Research*, *47*(1), 121 – 150. https://doi.org/10.3102/00346543047001121

Lord, F. M. (1956). *The Measurement of Growth. Educational and Psychological Measurement*, *16*(4), 421-437. https://doi.org/10.1177/001316445601600401

Lord, F. M. (1963). Elementary Models for Measuring Change. In C. W. Harris (Ed.), *Problems in Measuring Change* (pp. 21-38). The University of Wisconsin Press.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304-305. https://doi.org/10.1037/h0025105

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Maris, E. (1998). Covariance adjustment versus gain scores—Revisited. *Psychological Methods*, *3*, 309-327. https://doi.org/10.1037/1082-989X.3.3.309

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577-605. https://doi.org/10.1146/annurev.psych.60.110707.163612

Miyazaki, Y., Kamata, A., Uekawa, K., & Sun, Y. (2022). Bias for Treatment Effect by Measurement Error in Pretest in ANCOVA Analysis. *Educational and Psychological Measurement*, *82*(6), 1130-1152. https://doi.org/10.1177/00131644211068801

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*(3), 622-637. https://doi.org/10.1037/0033-2909.88.3.622

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*(1), 85-86. https://doi.org/10.1037/h0076158

Petscher, Y., & Schatschneider, C. (2011). A Simulation Study on the Performance of the Simple Difference and Covariance-Adjusted Scores in Randomized Experimental Designs. *Journal of Educational Measurement*, *48*(1), 31-43. https://doi.org/10.1111/j.1745-3984.2010.00129.x

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic Methods for Questions Pertaining to a Randomized Pretest, Posttest, Follow-Up Design. *Journal of Clinical Child and Adolescent Psychology*, *32*(3), 467-486. https://doi.org/10.1207/S15374424JCCP3203_15

Rogosa, D. (1995). Myths and Methods: "Myth About Longitudinal Research" Plus Supplemental Questions. In J. M. Gottman (Ed.), *The Analysis of Change* (pp. 3-66). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203763391

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726-748. https://doi.org/10.1037/0033-2909.92.3.726

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the Reliability of the Difference Score in the Measurement of Change. *Journal of Educational Measurement*, *20*(4), 335-343. https://doi.org/10.1111/j.1745-3984.1983.tb00211.x

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*(2), 203-228. https://doi.org/10.1007/BF02294247

Roh, Y., & Chang, J. Y. (2006). The Psychological Impact of Perceived Overqualification in College Graduates: A Longitudinal Analysis. *Korean Journal of Industrial and Organizational Psychology*, *19*(1), 59-84. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART000991801

Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects.

Biometrika, *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage Learning.

Thomas, D. R., & Zumbo, B. D. (2012). Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA: In Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement,* *72*(1), 37-43. https://doi.org/10.1177/0013164411409929

Valente, M. J., & MacKinnon, D. P. (2017). Comparing Models of Change to Estimate the Mediated Effect in the Pretest–Posttest Control Group Design. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 428-450. https://doi.org/10.1080/10705511.2016.1274657

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*(9), 920–925. https://doi.org/10.1016/j.jclinepi.2006.02.007

Van Breukelen, G. J. P. (2013). ANCOVA Versus CHANGE From Baseline in Nonrandomized Studies: The Difference. *Multivariate Behavioral Research*, *48*(6), 895-922. https://doi.org/10.1080/00273171.2013.831743

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*(1), 147–151. https://doi.org/10.1037/0033-2909.109.1.147

Werts, C. E., & Linn, R. L. (1970). A general linear model for studying growth. *Psychological Bulletin*, *73*, 17-22. https://doi.org/10.1037/h0028330

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*(5), 906-919. https://doi.org/10.1037/a0036387

Wright, D. B. (2006). Comparing groups in a before–after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology*, *76*(3), 663-675. https://doi.org/10.1348/000709905X52210

Zimmerman, D. W., & Williams, R. H. (1982). Gain Scores in Research Can Be Highly Reliable. *Journal of Educational Measurement*, *19*(2), 149-154. https://doi.org/10.1111/j.1745-3984.1982.tb00124.x

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 343-351. https://doi.org/10.1111/j.2044-8317.1998.tb00685.x

1차원고접수 : 2024. 02. 23
2차원고접수 : 2024. 08. 08
최종게재결정 : 2024. 09. 12

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 231-260
http://dx.doi.org/10.22257/kjp.2024.9.43.3.231

How to analyze group difference in change: Comparing difference score model and analysis of covariance model

Youngsoo Lee **Hye Won Suk**

Department of Psychology, Sogang University

In various fields of psychology, researchers commonly investigate difference in changes between treatment and control groups by analyzing data gathered before and after interventions. The most widely used analytical methods used in such cases are the difference score model and the analysis of covariance model. However, since these models may produce conflicting outcomes, researchers often get confused when determining the most appropriate method for their studies. Therefore, this study aims to offer an in-depth examination of the theoretical and empirical difference between these models, aiming to furnish guidelines on when to use which method. Initially, we introduce and illustrate each model using an example dataset to showcase their potential divergent analytical outcomes. Subsequently, we scrutinize the debate on the use of difference scores, debunking traditional criticisms grounded in oversimplified assumptions and misunderstandings. We then delve into the implicit assumptions of both models within the framework of causal inference and, drawing upon these assumptions and findings from simulation studies, furnish recommendations for selecting the appropriate method under different participant allocation methods and analytical purposes. This study endeavors to empower researchers in making informed decisions regarding their choice of analytical methods, thereby enhancing the rigor and efficacy of their investigations.

Key words : difference score, ANCOVA, causal inference, treatment effect, Lord's paradox

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 261-295
http://dx.doi.org/10.22257/kjp.2024.9.43.3.261

번안 심리검사 타당화 작업에 대한 체계적 검토: 검수와 보고 관행에 대한 검토와 제언

김 미 림† 임 예 지

고려대학교

본 연구는 한국 심리검사 번안 타당화 연구를 체계적으로 검토하여 국내 심리검사 번안의 검수와 보고에 대해 제언하는 것을 목표하였다. 교육 및 심리검사의 표준과 ITC 검수 지침에 기반하여 2017년부터 2023년까지 출간한 총 107개 번안검사의 타당화 관행을 검토하였으며, 번안검사의 심리측정적 속성에 대한 정보가 충분히 보고되고 있는지 또한 각 연구가 타당도 확보를 위해 적합한 분석을 실시했는지 조사하였다. 검토 결과, 두 편을 제외하고 모든 번안검사 타당화 연구는 신뢰도를 보고하였고 요인분석을 거의 필수적으로 사용하여 구성 타당도를 검증하였음을 알 수 있었다. 따라서 해당 요인분석의 시행과 보고 관행에 대해 보다 면밀히 살펴보고 몇 가지 제언을 함으로써 응용 연구자들에게 도움이 되고자 하였다. 또한 같은 개념에 대하여 연구마다 각기 다른 용어의 사용이 빈번함을 지적하며, 심리학 연구에서 사용하는 학술 용어의 통합을 제언하였다. 체계적 검토 결과에 따른 본 연구의 제언은 번안검사 뿐만 아니라 새로운 검사의 개발 및 타당화를 목표로 하는 연구자들에게도 도움이 될 것으로 기대한다.

주요어 : 번안검사 타당화, 검수 지침, 체계적 검토, 심리측정적 속성, 요인분석

† 교신저자: 김미림, 고려대학교 4단계 BK21 심리학교육연구단, (02841) 서울시 성북구 안암로 145, Tel: 02-3290-2558, E-mail: mirimkim@korea.ac.kr

대다수 사회과학 연구가 잠재적 특성을 수리화하는 작업, 즉 측정을 통하여 양적 연구를 실시한다. 연구의 타당성을 위해서는 각 연구에 알맞은 측정 도구의 사용이 필수적이며, 이에 따라 적합한 측정 도구를 개발하고 도구의 적합성 및 타당성을 검증하는 작업 또한 중요하다. 타당화가 충분히 이루어진 검사 도구는 후속 연구에서도 계속해서 측정 도구로 활용될 수 있으며, 만약 원검사가 외국어로 개발된 것일 경우에는 한국어로 번역 및 국내 문화와 정서에 알맞게 번안하고 해당 번안검사를 타당화한다.

번안검사의 타당화는 검사를 단순히 번역하는 것 이상의 작업이라고 볼 수 있다. 이는 원검사의 언어를 해당 국가의 언어로 번역하는 작업뿐만 아니라 현지 문화권 상황에 더욱 적합하게 하기 위한 노력이 추가적으로 필요하기 때문이다. 부적절한 번안은 검사의 타당도를 저해하고, 부적합한 도구의 사용은 연구 결과를 오도할 수도 있으므로 원검사와 마찬가지로 번안검사의 타당화 작업 역시 필요하다(Clark & Watson, 2019). 이에 따라 국제검사위원회(ITC: International Test Commission)에서는 번역 및 번안 검사를 위한 지침을 제공함으로써 번안 작업의 중요성을 강조하고 올바른 타당화 작업을 소개하고 있다.

서동기와 이순묵이 번역한 한국어판 ITC 지침은(ITC Guidelines for Translating and Adapting Tests; Gregoire, 2018) 크게 여섯 가지 지침으로 이루어져 있다. 첫 번째 지침은 번안 이전에 필요한 선행조건과 관련이 있으며, 원검사에 대한 번안 허가의 필요성을 이야기하고, 원검사와의 언어적/문화적 차이에 따라 발생할 수 있는 여러 상황에 적합한 절차를 설명한다. 두 번째 지침은 언어와 문화를 고려한 번역 설계와 절차에 기반한 검사개발 지침으로 대상 언어와 문화, 그리고 검사에 대한 지식이 풍부한 전문가와 대상 모집단에게 번안 검사 문항, 지침, 그리고 집행을 유사하게 하기 위한 근거 제공을 강조한다. 또한 본격적인 대규모 검사 이전에 파일럿 검사와 그에 대한 문항 분석을 실시한다면 본 검사의 시행에 참고할 수 있는 유용한 정보를 미리 파악하고 추후 확인적 증거로도 활용할 수 있음을 언급하기도 하였다. 세 번째 지침은 검수 방법에 대한 것으로, 타당화 과정에 쓰이는 실제적 자료 분석법(empirical data analysis)과 관련이 있다. 심리측정적 속성(psychometric properties)에 기반한 다수의 분석법을 설명하고, 적합한 신뢰도 및 타당도 확보를 위한 방법을 안내한다. 번역에서의 이슈나 문화 차이에서 발생할 수 있는 문항 비동등성을 나타내는 문항차별기능(differential item functioning)을 살펴보는 분석 또한 언급하고 있으며, 요인분석 등을 통한 검사 구조 동등성 검증을 강조하고 있다. 네 번째와 다섯 번째 지침은 각각 검사의 실시와 해석 단계에서 언어/문화적 차이에서 비롯된 결과의 차이가 발생할 수 있음을 알리고 있으며, 그 차이를 최소화하기 위한 방안을 소개한다. 마지막 지침으로, ITC는 문서화를 강조한다. 번안검사 연구자는 번안 검사와 원검사의 공통점 및 차이점에 대해서 상세하게 문서화하고, 번안검사를 사용하는 실무자들을 위하여 검사 매뉴얼을 제공해야 할 의무가 있음을 강조한다. 번안검사 개발과 관계된 기술보고서와 검사 사용 지침에 대한 설명이 충분해야 추후 번안검사 사용의 타당도를 저해하지 않을 수 있다. 앞서 설명한 전반적인 지침 외에 구체적인 검수와 검사 실시 상황을 알아보고 관련 지침을 파악하고자 하는 연구자들

은 ITC 지침을 필수적으로 참고할 것을 제안한다. 특히 세 번째 지침에서 설명하고 있는 심리측정적 속성과 구체적인 분석법들은 국내 번안검사 연구에 큰 도움이 될 것으로 기대한다.

최근에는 국외에서 개발된 측정도구 및 검사에 쉽게 접근할 수 있기 때문에, 연구자들은 새로운 도구를 빠르게 접하고 이를 번안하여 연구에 이용할 수 있다. 일례로 한국심리학회 산하 학술지에서 출간한 번안 연구의 동향을 살펴보면, 2017년에는 아홉 개의 연구로 비교적 적은 편이었으나 2023년에는 21건으로 증가하여 번안검사에 대한 수요가 증가함을 보였다. 따라서 적합한 번안검사 타당화는 더욱 중요해졌으며 이에 본 연구는 번안검사를 이용하는 연구자들의 타당화 연구 현황을 알아보고 그 적합성을 살펴보는 것을 주요 목표로 하였다. 특별히 국내에 ITC 지침이 소개된 2017년 이후의 번안 심리검사 연구를 대상으로 Thoemmes와 Kim(2011)의 방법을 참고한 체계적 검토(systematic review)를 실시하였고, ITC 검수 지침에 기반하여 번안검사의 정보와 타당화 과정을 살펴보았다. 본 연구는 검토 결과에 대한 독자들의 올바른 이해를 위하여, 번안과 검수 지침을 먼저 소개하고, 기반이 되는 이론적 배경을 살펴보고자 한다. 후반부에는 체계적 검토의 결과를 종합하여 번안 심리검사 연구의 타당화 현황에 대하여 논의하였다. 즉, 번안검사의 심리측정적 속성에 대한 경험적 증거가 충분한지 그리고 증거에 기반한 적합한 판단이 이루어졌는지 살펴보았다. 본 연구는 ITC 지침 중 양적 방법에 의한 검수와 문서화에 대해서 강조하고 있으므로 검토 결과 및 제언은 번안검사 연구자들에게만 국한될 필요가 없으며 새로운 검사의 개발 및 타당화를 목표로 하는 연구자들에게도 도움이 될 것으로 기대한다.

번안

연구자별로 적합한 번안 과정에 대해 다양하게 정의하고 있으나 본 연구는 Beaton 등(2000)과 Wild 등(2005)이 설명한 구조로 이를 설명하고자 한다. 첫째, 일차 번역은 원검사와 번안검사에 해당하는 언어와 문화에 대해 충분한 지식과 이해를 가지고 있는 두 명의 이중 언어 사용자가 실시한다. 둘째, 각 번역 결과가 종합되도록 조율을 한다. 셋째, 번역본을 다시 원검사의 언어로 번역하는 과정, 즉 역번역 과정을 거쳐 더욱 자연스러운 번안 결과를 기획한다. 이때 역번역은 원검사의 언어를 모국어로 사용하는 번역자가 담당하도록 한다. 넷째, 문화적으로 자연스러운 번안을 위해 제삼의 번역자 혹은 전문가 패널을 활용하여 검사 간의 동등성을 확인한다(Brislin, 1970; Gregoire, 2018). 이는 언어적 동등성뿐만 아니라 개념적인 동등성을 확보하기 위함이다. 마지막으로, 예비 검사를 실시하여 검사에 대한 응답자의 이해도를 파악하고, 추가로 내용 타당도에 대한 근거를 확보한다.

검수 지침

연구자는 다양한 경험적 분석을 통해서 번안검사가 원검사와 크게 다르지 않음을 검수하여 검사 동등성을 확립할 필요가 있다. 이와 관련하여, ITC 지침은 연구자가 고려해야 하는 첫 번째 사항으로 충분한 표본크기를 이야기한다. 번안검사의 표본은 원검사와 마찬가지로 검사 대상의 모집단을 대표할 수 있어

야 하며, 분석 모형을 안정적으로 추정할 수 있을 만큼 충분한 크기여야 하기 때문이다. 적합한 표본을 확보했다면, 원검사와의 동등성을 검증하고, 번안검사의 심리측정적 속성을 살펴본다.

검사 목적에 따라, 원검사와 번안검사가 확립해야하는 동립성의 종류는 상이할 수 있다. ITC 지침은 OECD 주관의 국제 학업성취도 평가인 PISA(programme for international student assessment)와 우울척도를 예시로 들고 있다(Gregoire, 2018). PISA 검사는 OECD 국가를 포함한 다양한 국가들의 학생 역량을 평가하고 비교하는 것을 목표하므로, 원검사와 번안검사 간의 검사 양식 동등성을 검수할 필요가 있다. 또한 국가 간 검사점수의 비교를 목적으로 척도체계의 동등화(equating) 및 연계(linking) 작업을 실시한다. 반면, 우울척도는 언어체계가 다른 임상군 간의 우울점수 비교보다는 번안검사가 우울의 측정에 타당한 도구인지, 그리고 측정된 우울 점수가 적은 오차를 갖는지의 여부가 더 중요하다. 따라서 원검사와 번안검사의 척도체계 비교보다는 언어, 혹은 문화에 따른 검사 구조의 동등성을 확인하고 번안검사의 타당성을 확보하기 위한 통계적 접근을 고려한다.

원검사가 신뢰롭고 타당하다는 근거가 충분할지라도, 해당 검사의 개정판에 대해서는 별도의 타당화 작업이 필요하다(AERA, APA, NCME, 2014; Gregoire, 2018). 같은 논리로 번안검사 또한 추가적인 타당화 작업을 실행하여 번안검사 사용에 대한 타당성을 뒷받침할 수 있어야 하며, 연구자는 타당화 과정에서 근거한 검사 정보를 제공할 연구적 책무가 있다.

신뢰도

신뢰도는 같은 검사를 이용하여 동일한 피험자의 구성개념을 반복 측정할 때 검사의 결과가 일관된 정도를 나타낸다. 즉 신뢰도가 낮은 검사는 반복 측정에 따라 상이한 결과가 발생할 수 있기 때문에(강태훈, 김명연, 2023; 성태제, 2002), 연구자들은 변칙적인 상황에도 일정한 검사 점수를 확보하고자 신뢰도가 높은 검사를 사용하고자 한다. 또한 American Educational Research Association(AERA), American Psychological Association(APA), 그리고 National Council on Measurement in Education(NCME) (2014)에서 출간한 교육 및 심리검사를 위한 표준(Standards for Educational and Psychological, Testing)에 따르면 연구자는 검사의 시행마다 신뢰도를 보고해야 하며, 번안 검사의 경우에는 원검사와 번안검사의 신뢰도를 모두 보고할 필요가 있다고 하였다. 이에 대하여, 본 연구는 검사 이론에 기반한 신뢰도의 정의를 소개함으로써 검사의 신뢰도가 거듭 강조되는 이유를 살펴보고자 한다.

연구에서 항상 언급되는 신뢰도의 개념과 신뢰도 지수는 고전검사이론(classical test theory)에 기반하고 있다. 고전검사이론은 관찰점수 X가 진점수(true score) T와 측정오차(measurement error) E로 구성된 확률변수(random variable)라는 개념을 바탕으로 신뢰도의 개념을 정의하였다(Crocker & Algina, 1986). 측정에 오차가 전혀 없다면 검사점수는 항상 일정하겠지만, 현실적으로 측정오차가 없는 검사를 개발하기란 어려운 일이다. 따라서 검사를 개인에게 무한 번 시행할 경우, 검사점수는 시행마다 근소한 차이를 가질 수 있으며, k번째 검사를 통해 관찰된 점수 X_k에 대해

확률 p_k를 갖는 확률분포를 따르게 된다. 확률변수의 기댓값은 확률변수의 평균값으로 표현할 수 있는데, 이를 이용하여 고전검사이론에서는 확률변수 X의 기댓값을 진점수($T = \Sigma_{k=1}^{K} X_k p_k$)라고 정의하고, 관찰점수의 평균값으로 진점수를 표현한다. 또한 관찰점수와 진점수의 차이값, 즉 관찰점수가 진점수에서 벗어난 정도를 측정오차로 정의한다. 식 (1)은 이와 같은 세 가지 개념의 관계를 나타낸다.

$$X = T + E \tag{1}$$

또한 진점수와 측정오차 간에 상관이 없다는 가정에 기반하여 관찰점수의 분산($\rho^{2}{}_{X}$)을 진점수 분산($\rho^{2}{}_{T}$)과 측정오차 분산($\rho^{2}{}_{E}$)의 합으로 나타낼 수 있으며, 이는 식 (2)와 그림 1로 표현할 수 있다. 이때 신뢰도는 관찰점수의 분산에 대해 진점수 분산이 차지하는 비율로 정의한다($\rho_{XX} = \sigma_T^2 / \sigma_X^2$).

$$\rho_X^2 = \rho_T^2 + \rho_E^2 \tag{2}$$

그림 1을 통해 관찰점수 분산 내에서 진점수 분산의 차지하는 비율이 클수록 측정오차의 비율이 작아지는 것을 쉽게 파악할 수 있다. 앞서 설명한 검사점수의 일관도와 연결해 보면, 신뢰도가 높은 검사일수록 측정오차보다 진점수가 차지하는 비율이 높아 일관된 결과를 제공하는 경향이 있다고 볼 수 있다. 따라서 측정오차의 영향력이 적은 것을 증명하기 위해서라도 신뢰도는 계속해서 검증되고 강조되어야 할 검사의 속성이 될 수밖에 없다.

그러나 검사의 높은 신뢰도가 항상 검사의 적합성으로 귀결하는 것은 아니므로 주의가 필요하다(Raykov & Marcoulides, 2011). 신뢰도 지수의 한 가지인 Cronbach's α계수는 검사를 이루는 문항 간 공분산을 이용하여 산출되기 때문에 검사를 이루는 문항 간의 일관성, 즉 내적 일관성을 나타낸다. α계수는 측정 도구의 적합성을 나타내기 위한 특성으로 보고되고 있으나, 검사를 이루는 문항 간의 상관이 높을수록 그리고 문항의 개수가 많아질수록 그 크기가 커지는 경향이 있기 때문에, 높은 신뢰도가 단순히 좋은 검사를 나타내는 것은 아님을 알 수 있다. 단적인 예로, 우울을 측정하기 위하여 지능검사를 사용한 경우를 생각해 보자. 모든 연구자가 지능검사를 통해 우울을 측정하는 것이 적합하지 않음을 알고 있지만, 지능검사를 구성하는 문항 간 상관이 크고, 문항의 개수가 충분히 많다면 연구 목표와 관계없이 연구 도구의 신뢰도는 높게 추정된다. 다시 말해 검사가 연구에 적합한지와는 무관하게, 검사의 신뢰도는 높게 나타날 수 있다. 이러한 예는 신뢰도 이외의 추가적인 근거에 기반하여 검사를 평가할 필요가 있음을 시사한다(Crocker & Algina, 1986). 검사 타당도가 중요한 이유가 바로 여기에 있다.

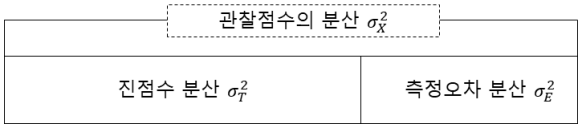


그림 1. 관찰점수, 진점수, 그리고 측정오차 분산의 관계

타당도

최근까지 검사 타당도의 개념은 변화하고 있으며, 이에 따라 검사가 근거해야 하는 타당도의 세부 유형 또한 다양해지고 있다(강태훈, 김명연, 2023). 교육 및 심리검사의 표준(AERA, APA, & NCME, 2014)의 정의에 따르면, 타당도는 검사를 통한 의사결정이 측정(measurement) 혹은 예측(prediction)과 같은 검사의 설계 목적을 충족하는 정도를 의미하며, 그 유형을 크게 내용 타당도(content validity), 준거 타당도(criterion validity), 그리고 구성 타당도(construct validity)로 분류할 수 있다. 그러나 연구 분야에 따라 세 가지 타당도의 하위 유형이 조금씩 상이하기 때문에 본 연구에서는 미국심리학회(APA)의 정의에 따라 세부 유형을 논의하고자 한다. 따라서 예측 타당도(predictive validity)와 공존 타당도(concurrent validity)는 준거 타당도(criterion validity)로, 그리고 수렴 타당도(convergent validity)와 변별 타당도(discriminant validity)는 구성 타당도(construct validity)의 하위 유형으로 분류하였다.

내용 타당도는 타당도 확보를 위한 질적인 근거로, 이를 이용하여 검사의 내용이 측정하고자 하는 구성개념 및 속성과 관계가 있는지에 대해 주관적으로 판단한다(신진아 외, 2021). 간단한 예로, 도박중독을 측정하고자 하는 검사는 중독의 잠재 속성과 관계된 내용으로 구성되어야 한다. 그러나 특별히 도박중독을 나타내는 행동을 기술함으로써 다른 중독(예: 약물중독)과는 차별화할 수 있는 행위를 측정하고, 이에 해당하는 잠재점수를 수량화할 수 있도록 설계해야 한다. 이렇게 심리검사 문항은 대상 구성개념을 나타내는 특수한 행위 및 생각 또한 기술할 수 있어야 하므로, 해당 분야 전문가들의 의견을 수렴하고 검사 문항을 설계함으로써 내용 타당도를 확보할 수 있다.

준거 타당도는 검사와 외부 준거변수와의 상관관계를 살펴봄으로써 검사의 타당도를 검증하는 방식이며, 예측 타당도와 공존 타당도가 이에 포함된다. 예측 타당도가 중요한 상황으로는 선발 현장을 예로 들 수 있다. 인적성 검사에 기반하여 응시자의 채용 여부를 결정할 경우, 검사의 타당도는 검사 점수가 채용 이후 응시자의 업무성과(준거변수)를 얼마나 예측할 수 있는지와 관계가 있을 것이다(Jenkins & Griffith, 2004). 검사 점수와 업무능력 간의 상관이 강할수록 높은 예측 타당도를 의미하여, 검사 점수가 높은 사람들을 우선적으로 선발하는 것에 대한 근거가 된다. 한편, 공존 타당도를 알아보기 위해서는 기존에 널리 쓰이는 검사를 준거변수로 하여 제작 검사와의 상관을 살펴보며, 예측 타당도와 마찬가지로 준거 검사와의 상관이 강할수록 공존 타당도가 높다고 판단한다(신진아 외, 2021). 널리 쓰이고 있는 검사는 반복적으로 타당화 작업을 거친 검사를 의미하기 때문에, 공신력 높은 검사와의 상관이 높은 검사 역시 타당한 검사로 공인할 수 있다는 논리다.

검사는 예측뿐만 아니라 측정을 주요 목적으로 하여 실시할 수 있다. 측정은 구성개념에 대한 응답자의 상태(status)와 위치(standing)에 숫자를 부여하여 수리적으로 추정하는 것을 의미하며(Raykov & Marcoulides, 2011), 구성타당도는 이러한 과정이 적합한지에 대한 근거가 된다. 구성개념은 실재를 관찰할 수 없는 잠재적 속성이기 때문에, 임의의 척도 없이는 구성개념의 차원(dimension)에 대한 개인의 위치를 판단할 수 없다. 따라서 측정을 통

해 개개인의 구성개념에 숫자를 부여함으로써 잠재 속성의 위치를 가늠하고 상대적인 정도를 비교한다. 구성 타당도의 근거는 검사 문항과 구성개념 간의 관계에 기반하며, 검사 점수가 나타내는 것이 실제 측정하고자 하는 대상 구성개념인지, 또한 검사가 구성개념에 최대한 근접한 점수를 추정하였는지를 판단한다(AERA, APA, & NCME, 2014).

구성 타당도는 단 한 가지 통계분석을 통해 평가할 수 있는 개념이 아니며 다양한 근거에 기반하여 종합적으로 평가하게 된다. 구성 타당도를 평가하는 접근 방식은 크게 두 가지로 요약할 수 있는데, 검사를 이루는 문항과 구성개념 간의 관계성을 검증함으로써 타당화하는 것을 내적 접근(internal approach), 외부/준거 검사와의 관계성에 기반하여 구성 타당도를 평가하는 것을 외적 접근(external approach)이라고 한다(Clark & Watson, 2019; Loevinger, 1957; Messic, 1995; Raykov & Marcoulides, 2011). 앞서 강조하였듯이 구성 타당도의 확보를 위해 검사의 다양한 속성을 평가하는 것이 바람직하므로, 두 가지 접근 방식에 모두 근거하여 구성 타당도 개념을 확보할 것을 권고한다.

본 연구는 구성 타당도를 살펴보기 위한 방법으로 요인분석을 다루고자 한다. 요인분석 모형은 구성 타당도를 살펴보기에 적합한 모형 중 하나라고 볼 수 있으며, 국내 심리검사 타당화 연구를 검토한 결과 실제로 가장 빈번하게 사용되고 있는 모형임을 알 수 있었다. 요인분석의 측정모형에 주목하는 경우, 연구자는 요인-문항 간의 관계를 살펴봄으로써 각 문항이 요인(구성개념)을 적절히 측정하는지 평가한다(내적 접근). 이때 문항과 요인 간의 관계가 충분히 커서 잠재점수가 관찰점수를 설명하는 정도가 높을수록 검사 구성 타당도의 적합한 근거가 될 수 있다. 또한 요인분석 모형을 통해 대상 검사와 외부 검사에 해당하는 요인 간 관계를 살펴봄으로써 구성 타당도를 논의할 수 있으며, 이처럼 준거와의 관계에 근거하는 구성 타당도를 특별히 수렴 및 변별 타당도라고 일컫는다(외적 접근). 검사가 대상 구성개념을 잘 측정했다면, 유사한 구성개념을 측정하는 준거검사와 연관성이 높아 수렴하는 경향이 커야 한다. 두 검사가 측정하는 요인 간의 상관이 높을수록, 두 검사의 동질성에 기반한 수렴 타당도의 근거가 된다. 연구자는 수렴 타당도를 확보함과 동시에 변별 타당도에 대해서도 고려할 필요가 있다. 대상 검사와 준거 검사가 측정하는 구성개념이 서로 다를 경우, 두 검사는 서로 연관성이 낮고 서로 변별하는 경향이 커야 한다. 요인 상관이 낮을수록 각 검사가 측정하는 요인 간의 이질성에 기반하여 변별 타당도의 근거를 확보할 수 있다.

심리학 연구에서 언급되는 또 다른 타당도로는 증분 타당도(incremental validity)를 들 수 있다. 증분 타당도는 Sechrest(1954)가 고안한 개념으로, 대상 검사가 기존 검사와 변별이 잘 되는지 그리고 기존 검사와 비교했을 때 종속변수에 대해 추가로 예측하는 면모가 있는지 살핀다.

요인분석 모형

요인분석은 구성개념의 차원성과 문항의 특성을 함께 평가할 수 있기 때문에 심리측정적 속성을 확인하는 주요 분석 방법으로 사용되고 있으며(Brown, 2015), ITC 지침 또한 요인분석 모형을 비중 있게 다루고 있다. 따라서 본 연구는 국내 번안검사 연구에서의 요인분석

관행을 ITC 지침 및 선행 연구로부터의 권고 사항과 비교하고 타당화 작업에 유용한 사항을 논의하고자 하였다. 비교 결과를 기술하기 전에 요인분석 모형에 대한 이론적 배경을 아래에 설명하였다.

탐색적 요인분석과 확인적 요인분석

심리학 연구를 포함한 대다수의 사회과학 연구에서 활용하는 요인분석 모형은 공통요인분석(common factor analysis)으로 탐색적 요인분석 혹은 확인적 요인분석 모형을 의미하며(이순묵, 1994; Brown, 2015), 식 (3)과 같은 선형식으로 나타낼 수 있다.

$$x_{ij} = \lambda_{j1}\xi_{j1} + \lambda_{j2}\xi_{j2} + \ldots + \lambda_{jr}\xi_{jr} + \delta_{ij} \quad (3)$$

x_{ij}는 사람 i가 문항 j에 대해 응답한 문항점수를 의미하며, 이는 첫 번째 공통요인 ξ_{j1}부터 r번째 공통요인 ξ_{jr}, 그리고 고유요인 δ_{ij}와의 선형관계로 표현할 수 있다. 고전검사이론에서 관찰점수의 분산을 진점수와 측정오차 분산의 구성으로 정의했듯이, 공통요인분석 모형은 문항점수의 분산을 공통분산(communality)과 고유분산(unique variance)으로 표현한다(Kline, 2013). 공통분산은 공통요인이 문항점수를 설명하는 정도를 나타내며 고유분산은 공통요인이 설명하지 못하는 문항 특유의 특성이 차지하는 정도를 의미한다. 이때 요인분석은 다수 문항의 공통분산에 관여하는 요인을 추출하고 그 속성을 살펴보는 것을 목표로 한다(Brown, 2015).

요인구조에 대한 가설의 차이

탐색적 요인분석과 확인적 요인분석의 가장 큰 차이는 요인의 개수와 요인-문항 간 관계에 대한 가설 유무에 있다. 탐색적 요인분석은 요인-문항 간의 불명확한 관계에 기반하여 공통분산을 설명하는 적합한 개수의 요인과 요인구조를 탐색하는 것을 목표로 하지만, 확인적 요인분석의 경우 구조에 대한 명확한 가설을 설정하고 이를 검증한다(Brown, 2015). 두 모형의 요인구조 차이는 그림 2를 통하여 살펴볼 수 있다.

그림 2는 여섯 개의 문항과 두 개 요인 간의 관계를 나타내고 있다. 탐색적 요인분석을 나타내는 (a)의 경우, 각 요인이 모든 문항을 설명하고 각 문항이 다수의 요인으로 회귀함으로써 교차 요인부하(crossed-factor loading)를 추정하게 된다. 이는 검사 문항과 요인의 관계에 대한 명확한 가설이 없기 때문에 모든

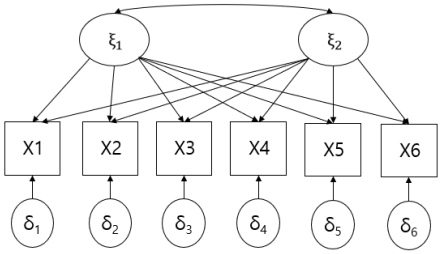
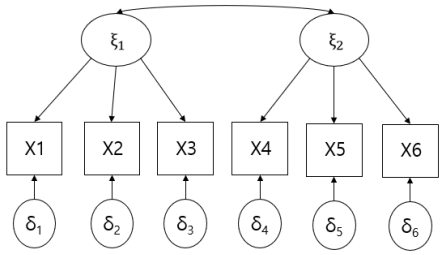


그림 2. (a) 탐색적 요인분석 모형



(b) 확인적 요인분석 모형

요인-문항 관계를 살펴보기 위함이다. 그림 (a)의 예시는 두 개의 요인에 기반한 요인구조를 나타내고 있으나, 실제 탐색적 요인분석은 요인 개수에 대한 모든 경우의 수(예: 여섯 개의 문항에 대해서는 단일 요인부터 6요인까지)에 기반하여 가능한 요인-문항 관계를 살펴보게 된다. 가설이 부재한 경우에 유용하기 때문에 검사 제작의 초기 단계에 실시하여 요인구조를 탐색한다(Flora & Flake, 2017). 반면 확인적 요인분석 모형의 경우, 요인 개수와 요인-문항 관계에 대한 구체적인 가설에 기반하여 모형을 설정하며 그 결과 그림 (b)와 같이 각 문항은 특정 요인에만 회귀한다. 따라서 요인과 문항 간의 유의미한 관계를 예측할 수 있는 경우에 대해서는 해당 요인부하를 추정하고, 그렇지 않은 관계에 대해서는 요인부하를 0으로 고정한다. 추정해야 하는 모수의 수가 적어짐으로써 탐색적 요인분석 모형보다 추정이 용이하다는 장점이 있으며, 확인적 요인분석은 가설에 기반하므로, 요인구조에 대한 탐색 이후에 최종 모형을 확인 및 타당화 하기 위하여 쓰일 수 있다. 즉, 탐색적 요인분석을 통하여 요인구조에 대한 가설을 설정하고, 이를 확인적 요인분석을 통하여 교차 타당화하는 식으로 검사를 타당화하는 것이 일반적이다(Brown, 2015; Flora & Flake, 2017).

요인분석 모형의 설정, 평가 및 수정

요인구조에 대한 두 요인분석 모형의 차이는 모형의 설정(specification), 식별(identification), 그리고 분석 방법에도 차이를 갖게 한다. 이에 대하여 본 연구는 탐색적 요인분석부터 확인적 요인분석의 순으로 각 모형의 설정과 평가를 간략히 설명하고자 하였으며, 표 1에 이를 정리하여 연구자들에게 도움이 되고자 하였다.

탐색적 요인분석은 대략 (1) 요인의 추출, (2) 요인개수 결정, (3) 그리고 요인구조의 회전을 걸쳐 모수를 추정한다. 요인 추출은 추출법에 따라 차이를 보일 수 있는데, 추정법의 예로는 최대우도법(maximum likelihood estimation)과 주축분해법(principal axis factoring)을 들 수 있다(이순묵, 1994; De Winter & Dodou, 2012). 통계 소프트웨어에 따라 추출법에 대한 선택의 폭은 상이할 수 있다. 이때 한 가지 주의할 것은 일부 소프트웨어에서 이용할 수 있는 주성분 분석(principal component analysis)의 활용이다. 특히 SPSS에서는 요인 추출법으로 주성분 분석을 선택할 수 있는데, 주성분 분석은 공통요인분석의 맥락보다는 자료 축소를 연구 목적으로 하는 요인분석 모형이기 때문에 연구자의 주의가 필요하다(이순묵, 1995; Brown, 2015). 공통된 요인의 추출과 그 해석을 목표로 하는 경우에는 주성분 분석 이외의 추정법이 적절하다고 볼 수 있다(Gorsuch, 1990).

탐색적 요인분석은 각 요인개수와 문항-요인 간 관계에 대한 명확한 가설이 없는 채로 수행되기 때문에 두 번째, 요인의 개수 결정 단계에서 연구자의 주관적 판단이 필요하다. 요인 개수의 판단, 즉 모형 결정에 참고할 수 있는 평가지표로는 스크리 도표(Cattell, 1966), 평행분석(parallel analysis; Horn, 1965), MAP test (Velicer, 1976) 등을 들 수 있으며, 소프트웨어에 따라 이용할 수 있는 평가지표에 차이가 있을 수 있다. 소프트웨어 전반에 걸쳐 공통적으로 사용할 수 있는 지표로는 스크리 도표, 평행분석, 그리고 모형 합치도지수(예: χ^2)를 들 수 있으며, 최대우도법을 사용하여 요인을

추출한 경우에 대해 *Mplus*, jamovi, 그리고 R 패키지는 대안적 합치도지수(예: CFI, RMSEA, SRMR)를 제시하고 있어 모형의 평가가 좀 더 수월해졌다.

탐색적 요인분석을 이용한 요인개수의 결정 단계에서 주의해야 할 사항은 다음과 같다. 첫째, 연구자는 카이저 룰(Kaiser-Guttman rule)에만 기반한 요인개수의 결정을 지양해야 한다. 이 지표는 1 이상의 고유치를 갖는 요인의 개수만큼 요인을 추출하는 방식으로, 판단이 직관적이라는 점에서 널리 사용된 적이 있다. 그러나 요인의 개수를 과대 혹은 과소 추정하는 경우가 많아 안정적인 결과를 산출하지 않는다는 점에서 현재는 추천하지 않는 준거로 볼 수 있다(Brown, 2015; Fabrigar, Wegener, MacCallum, & Strahan, 1999). 둘째, 각 평가지표는 서로 다른 결과를 가리키고 있을 수 있으므로 연구자는 다양한 평가지표에 복합적으로 근거하여 요인개수를 결정할 필요가 있다(Brown 2015; O'Connor 2000). 또한 단순히 통계적 결과에만 의존할 것이 아니라 이론에 기반한 해석가능성을 필수적으로 고려하여 요인의 개수를 결정할 필요가 있다.

마지막으로 살펴볼 탐색적 요인분석의 특징으로는 요인구조의 회전을 들 수 있다. 탐색적 요인분석 모형은 유일해(unique solution)가 아닌 다수의 해를 갖는 모형으로, 자료와 추출법이 같다면 동일한 모형 합치도지수를 갖는 여러 개의 요인구조를 산출할 수 있다(Brown, 2015). 이때 요인구조의 회전은 다요인 모형에 대해 특정 요인과 문항 간의 관계를 부각하여 해석이 수월하게 하는 과정으로, 특정 요인과 관계된 요인부하량을 작게 하거나 크게 하여 요인-문항 간 해석이 더욱 용이한 구조를 찾을 수 있도록 한다(Brown, 2015; Flora & Flake, 2017). 연구자는 요인 간 상관에 대해 가설을 세우고 회전법을 선택할 수 있으며, 요인 간 상관이 없다고 가정하는 경우와 있다고 가정하는 경우에 대해 각각 직각회전(요인상관=0)과 사각회전(요인상관≠0)에 해당하는 회전법을 선택하고 요인구조를 산출한다. 일반적으로 사회과학 연구에서는 요인 간 상관이 있음을 가정하므로, 특별한 이론적 근거가 없다면 사각회전법을 사용하는 것이 타당하다.

회전 이후에는 요인과 문항의 적합성을 고려하여 모형을 수정하거나 문항을 삭제하여 최종 요인구조를 결정한다. 먼저 추출한 요인이 적합하지 않다고 판단하는 경우, 연구자는 더 적은 수의 요인을 추출하는 전략을 취할 수 있다. (a)요인에 부하하는 문항의 수가 적어서 해당 요인에 대한 측정모형을 타당화할 수 없을 때, (b)문항 i에 대한 요인 j의 요인부하량 λ_{ij}이 작게 추정되고 공통분산($h_i^2 = \sum \lambda_{ij}^2$)의 크기가 작은 것으로 나타낼 때, 그리고 (c)요인 간 변별이 명확하지 않은 경우가 이에 해당한다. 반대로 검사 문항을 삭제하는 전략이 필요한 때도 있다. (d)다수 요인에 대한 특정 문항의 요인부하량의 크기가 비슷하여 요인-문항 관계를 명확히 할 수 없을 때, 또는 (e)문항의 모든 요인부하량이 [1].3이나 .4보다 작은 경우에는 해당 문항을 삭제하기도 한다. 참고로 각 상황에 대한 구체적인 준거는 통계적 타당화에 근거했다기보다는 응용 연구자들의 경험에 기반하여 마련된 것으로 보인다(Brown, 2015).

확인적 요인분석은 요인구조에 대한 명확한

1) 요인부하량이 0.3, 0.4인 경우, 공통분산은 각각 $0.09 = (0.3)^2$과 $0.16 = (0.4)^2$으로, 요인이 문항의 분산을 9%, 16% 설명한다고 볼 수 있다.

가설에 기반하므로, 탐색적 요인분석과는 달리 요인의 추출 및 요인개수의 결정에 대한 고민을 덜 수 있다는 장점이 있다. 각 문항은 한 개의 요인에만 회귀하는 단순구조에 대한 모수를 추정하므로 회전 과정이 불필요하며, 참조변수에 해당하는 요인부하량을 1로 고정하여 요인에 척도를 부여함으로써 모형을 식별한다. 확인적 요인분석 모형에 대한 평가는 모형 합치도지수를 이용하며, 가설 모형에 대한 합치도지수가 준거 합치도지수에 근접할 때 요인구조가 적합하다고 평가한다. 일반적으로 다수의 합치도지수를 종합적으로 고려하여 모형의 적합성을 판단하며 .95 이상의 상대적 합치도 지수(incremental fit index; 예: CFI, TLI), 0.08 미만의 절대적 합치도 지수(absolute fit index; 예: RMSEA, SRMR)을 나타내는 모형을 적합하다고 판단한다(Hu & Bentler, 1999; Kline, 2013).

확인적 요인분석 또한 탐색적 요인분석과 마찬가지로 요인과 문항의 적합성을 고려하여 모형을 수정하고 최종 모형을 결정하며(표 1 참고), 추가적으로 수정지수(modification index)를 활용하여 모형을 수정하기도 한다. 수정지수는 특정 모수에 대한 제약 조건이 얼마나 부적합한지를 나타내는 지수로 측정모형 내에서 0으로 고정된 가설이 해제될 때 감소할 수 있는 χ^2의 정도로 해석할 수 있다. χ^2의 감소량이 클수록 모형의 합치 정도가 높아지는

표 1. 요인분석 모형의 설정, 평가 및 수정

	탐색적 요인분석	확인적 요인분석
모형설정	요인추출 - 추출: 최대우도법, 주축분해법 요인개수 결정 - 준거: 스크리 도표, 평행검사, 모형합치도 요인구조의 회전 - 사각회전(예: Varimax), 직각회전(예: Promax)	- 이론에 기반하여 요인개수와 요인-문항 간의 관계를 설정
모형평가 및 수정	모형 평가 준거 - 모형합치도 - 요인부하량 및 공통분산 크기	모형 평가 준거 - 모형합치도 - 요인부하량 및 공통분산 크기
	모형 수정 요인개수의 축소 - 공통요인에 해당하는 문항이 적은 경우 - 공통요인에서 비롯된 요인부하량 및 공통분산이 작은 경우 - 요인 간 변별이 어려운 경우 문항 삭제 - 요인부하량이 다수 요인에 대하여 비슷한 경우 - 요인부하량 < 0.3 또는 0.4	모형 수정 문항 삭제 - 한 문항의 요인부하량이 다수 요인에 대하여 비슷한 경우 - 요인부하량 < 0.3 또는 0.4 수정지수 활용

것을 의미하기 때문에 큰 값을 갖는 수정지수를 고려하여 요인분석 모형을 수정한다. 그러나 수정지수의 크기에만 전적으로 의존하는 자료 중심의 분석(data-driven analysis)이 되지 않도록 이론적으로 설명이 가능한 수정지수를 선택하여 모형을 수정하는 것이 필요하다(Whittaker, 2012).

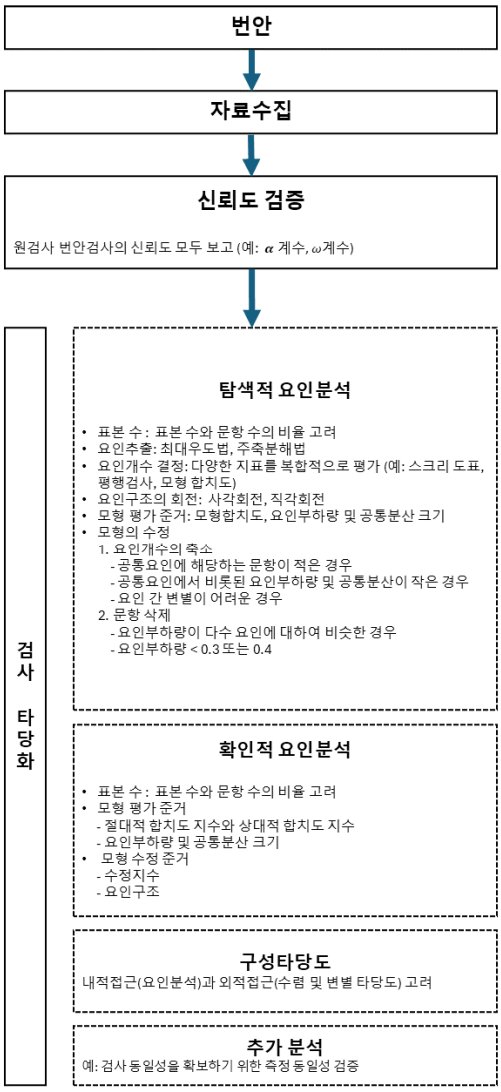


그림 3. 번안검사 타당화 과정

연구자는 요인분석 과정 및 결과와 관련된 주요 사항을 보고할 의무가 있다. 모형의 설정과 분석 과정을 설명하고 준거에 근거한 모형 평가 결과를 보고하며, 참고한 준거에 대해서도 명시하도록 한다. 모형을 수정하는 경우 또한 그 준거와 과정을 자세히 기술하여 수정 모형의 타당성을 뒷받침할 수 있어야 하며, 이렇게 선택된 최종 모형에 대해서는 분석 결과, 즉 요인구조를 보고함으로써 구성타당도에 대한 이해를 용이하게 할 필요가 있다. 앞서 설명한 탐색적 요인분석 모형과 확인적 요인분석 모형의 설정 및 평가 과정은 표 1과 같이 요약할 수 있다. 표 1은 연구자가 각 과정에서 고려해야 하는 대표적인 준거 또한 포함하고 있다. 또한 번안검사 타당화 과정을 구조화하여 그림 3으로 나타내었다.

요인분석과 표본

연구모형의 안정적인 수렴과 추정을 위해서는 충분히 큰 표본에 기반하여 분석할 필요가 있다. 그러나 연구 조건과 분석 모형에 따라 충분한 표본크기가 상이하기 때문에 적정 표본크기에 대한 획일적 기준을 마련하는 것은 어려운 일이다. 일반적으로 추정해야 하는 모수가 많을수록, 자료의 분포가 비정규성을 가질수록 근사 정규성에 기반하여 모수를 추정하기 위해 더 큰 표본이 필요하며, 비정규성을 통제하기 위한 추정법(예: 가중 최소 제곱법; weighted least square method)을 사용하는 경우에도 최대우도법을 사용한 경우보다 더 큰 표본이 필요하다(Kline, 2011). 같은 맥락으로, 요인분석도 표본의 크기가 클수록 작은 표본오차와 정확한 요인부하량을 산출할 수 있으므로(Browne, 1968) 충분한 크기의 표본에 기반

하여 검사를 타당화하는 것이 중요하다.

탐색적 요인분석의 경우 연구자는 *N:p* 비율, 즉 문항의 수(*p*)에 대비한 최소 표본크기(*N*)를 고려하여 적정 표본크기를 설계할 수 있다. 예를 들어 Everitt(1975)는 *N:p* 비율이 최소 10 이상의 값을 가질 것을 제안하였고, 이는 1개의 문항에 대해 표본크기가 10배 이상이어야 함을 의미한다. *N:p* 비율에 대한 획일적인 기준은 없지만, 이순묵(1994)에 따르면 해당 비율이 5 미만인 경우에 대해서 여러 선행 연구가 우려를 나타냈다고 하므로 연구자들은 이를 참고하길 바란다. 그밖에 요인분석에 필요한 절대적인 표본크기에 대한 연구도 있다. MacCallum, Widaman, Zhang 그리고 Hong (1999)은 추정된 요인구조의 복잡도에 따라 필요한 표본크기가 다름을 확인하였으며, 요인 간 변별이 수월하여 추출된 요인의 수가 적고, 각 요인 당 부하하는 문항이 3-4개인 경우에는 최소 300 이상의 표본이 필요하고, 그렇지 않은 경우에는 500 이상의 표본에 기반해야 안정적인 추정이 가능하다고 하였다. 그러나 실제 분석 이전에 연구자가 요인구조에 대해 판단하기는 어려우므로 500 이상의 표본을 확보하는 것이 타당한 것으로 보인다. 최근 연구도 이와 같은 맥락을 공유한다. Auerswald와 Moshagen(2019)은 공통분산과 요인부하량이 낮을 때, 요인 간 상관이 높을 때, 그리고 요인 당 부하하는 문항의 수가 적을 때 요인개수를 결정하는데 어려움이 더 커지는 경향이 있음을 밝혔으며, 이에 따라 요인개수 결정 준거가 서로 불일치한 결과를 지지할 경우에는 500 이상의 표본에 기반하여 더욱 안정적인 요인 추출을 계획할 것을 제시하였다.

확인적 요인분석을 실시할 경우에는 *N:q* 비율을 참고할 수도 있다. 추정하는 모수의 개수(*q*)에 대비하여 최소 표본의 크기(*N*)를 고려하는 방식이며, Jackson(2003)은 이상적인 *N:q* 비율을 20:1, 즉 한 개 모수를 추정할 때 필요한 표본은 20이라고 제안하였다. Kline(2011)은 최소 비율로 10:1을 언급하였으며, 한 모수 당 최소 10배 이상의 표본이 갖추어지지 않으면 표본크기가 작을수록 분석 결과에 대한 명확성이 줄어들 것임을 이야기하였다. 추가적으로, 구조방정식모형의 분석을 위해서는 200 이상의 표본이 필요하다는 선행 연구가 있으며(Barrett, 2007; Cattell, 1978), 다요인 모형에 대해서는 200, 500, 그리고 1,000의 표본크기를 작은 크기, 중간 크기, 그리고 큰 크기의 표본으로 고려할 수 있으니(Li, 2016), 연구자들은 이를 참고하여 자료수집과 분석을 계획하는 것이 필요하다.

실증 자료를 이용한 분석은 표집 오차(sampling error)를 마주하기 마련이며, 요인분석 또한 표집 오차의 영향을 받을 수 있다. 즉, 현재의 결과는 다른 표본을 이용하여 재분석한 결과와는 동일하지 않을 수 있기 때문에 반복 검증을 통한 교차 타당도의 확보가 중요하다(Flora & Flake, 2017). 만약 같은 표본을 이용하여 반복적으로 검증하는 경우에는 동일한 결과를 도출할 가능성이 높아지므로 교차 검증의 의미가 퇴색할 수 있다. 많은 연구자들이 표집의 어려움 때문에 동일하게 표집한 표본을 무작위로 나누어 요인분석 결과를 반복적으로 확인하기도 한다. 그러나 이와 같은 자료 이용법에 대한 명확한 이론적 근거는 찾아볼 수 없으며 연구자들의 경험에 근거한 것으로 보인다. 동일 검사에 대한 요인분석일지라도 다른 표본에 기반한 재분석은 언제나 필요한 과정이며(Cham, Hughes, West, & Im, 2014), 번안검사의 타당화가 중요한 이유가 여

기에 있다.

결과보고

앞서서 요인분석 모형의 설정, 평가, 그리고 수정과 관련된 구체적인 정보를 보고해야함을 명시한 바 있다. 마찬가지로 최종 요인구조를 도출한 후에는, 그 결과를 상세히 보고하고 해석하여 후속 연구자들이 검사를 사용하고 검사 타당도를 재검토할 수 있도록 해야 한다. 이를 위해 연구자는 (1) 최종 요인구조, (2) 모형합치도, (3) 그리고 최종 검사 문항을 보고하도록 한다. 최종 요인구조에는 요인부하량과 요인 간 상관이 포함되며 문항-요인 간의 관계와 요인-요인 간의 관계를 살펴볼 수 있도록 하기 위함이다. 또한 최종 모형과 관계된 합치도 지수를 보고함으로써 모형 적합성에 대한 근거를 제공하고, 수정지수를 이용한 경우에는 해당 지수와 관계된 통계적/이론적 근거 또한 보고하도록 한다. 마지막으로, 최종 검사를 이루는 문항을 보고하여 다른 연구자들도 타당한 검사를 이용할 수 있도록 돕고, 후속 연구가 타당화 작업을 지속할 수 있도록 독려할 필요가 있다.

번안 타당화에서의 요인분석

번안 타당화의 맥락에서도 요인분석 모형은 언어와 문화에 따른 검사 구조의 동일성을 확인하려는 방법으로 활용된다. 교육 및 심리검사의 표준(AERA, APA, & NCME, 2014)은 원검사의 축소판 검사를 개발하거나 원검사를 번안, 혹은 수정하는 경우에 요인구조를 재평가할 것을 강조하였는데, 이는 모집단이 변경되거나 원검사에 변경이 이루어지는 모든 검사에 대하여 요인구조를 재검증하는 것이 필요함을 시사하는 것으로 볼 수 있다. ITC 지침의 경우에도 번안 검사의 동일성 확인을 위하여 구체적으로 탐색적 요인분석, 확인적 요인분석 등을 이용할 것을 제안한 바 있다. 최근에는 검사 동일성을 확보하기 위한 측정 동일성(measurement invariance)의 성립이 중요해지고 있으며(Chen et al. 2020; Flora et al., 2017), 이 또한 요인분석을 통해 검증할 수 있는 심리측정적 속성이다. 요인분석 모형에 기반한 측정 동일성에 대해서는 김미림(2023), 주영신과 장승민(2023)의 연구를 참고하길 바란다.

앞서 설명한 이론적 배경을 바탕으로, 다음 장에서는 한국심리학회 산하 학술지에서 출간된 번안검사 타당화 연구에 대해서 살펴보도록 한다. 본 연구는 첫째, 출간된 번안검사 연구 결과를 바탕으로 번안검사 연구자가 검사의 심리측정적 속성(신뢰도와 타당도)을 어떻게 다루고 있는지 그 현황을 알아보았고 둘째, 타당화 작업에 쓰인 요인분석의 시행과 보고를 살펴봄으로써 심리검사 번안 연구에서의 요인분석 관행을 알아보았다.

방 법

연구 대상 및 자료수집

한국심리학회(https://www.koreanpsychology.or.kr)에 출판되는 연구 논문은 심리학회 산하의 17가지 하위분야에 따라 각기 다른 세부 주제를 다루고 있다. 본 연구는 전반적인 한국 심리검사 번안연구의 관행을 살펴보고자 하였으므로, 모든 하위분야의 연구를 대상으로 체계적 검토를 실시하였다. 구체적으로, 2017년부터

2023년까지 한국 심리학회지에 게재된 연구 논문 중, 척도 번안 및 타당화를 목적으로 하는 연구 논문을 검토의 대상으로 포함하였다. 검토 대상 연구의 색출을 위하여 해당하는 연도의 심리학회 산하 학회지 문헌들을 모두 검토하였다. 구체적으로, 각 권호에 실린 문헌의 제목과 키워드에 '척도', '번안', '번역', '타당화' 들어간 논문들을 우선적으로 선별했고, 추가로 이에 해당하는 연구의 초록을 검토하여 저자가 척도를 개발한 경우와 직접 번안을 하지 않는 경우를 제외하고, 척도 번안 및 타당화를 연구 주제로 한 논문들만 색출하였다. 그 결과 표 2에 보고한 것과 같이 한국 심리학회지의 하위 13개의 학술지로부터 총 107편의 논문을 체계적 검토의 대상 연구로 선정하였다.

표 2. 논문 분류

논문 하위 영역	N(Proportion)
건강	21(23%)
문화 및 사회문제	12(11%)
발달	5(5%)
법	1(1%)
사회 및 성격	4(4%)
산업 및 조직	9(8%)
상담 및 심리치료	18(17%)
여성	4(4%)
일반	3(3%)
임상심리 연구와 실제	11(10%)
중독	1(1%)
학교	5(5%)
Korean Journal of Clinical Psychology	13(12%)
합계	107(100%)

자료 코딩과 분석 방법

본 연구는 선정된 107개의 논문을 검사 정보, 검사의 심리측정적 속성, 그리고 연구에서 실시한 요인분석을 검토하고 분석하였다. 검사 정보로는 번안검사와 원검사와의 차이 여부와 최종 번안검사 문항의 보고 여부를 확인하였고, 심리측정적 속성으로는 각 연구의 신뢰도와 타당도 정보를 정리함으로써 107개 번안검사의 신뢰도 분포와 연구가 근거한 타당도의 종류를 살펴보았다.

또한 요인분석 모형이 번안검사 연구에서 빈번하게 사용되고 있음을 확인하고, 탐색적 요인분석과 확인적 요인분석으로 나누어 번안 타당화를 위한 요인분석의 시행과 결과에 포함된 정보를 분석하였다. 먼저 각 연구에서의 표본크기와 문항 대비 표본크기의 비율($N:p$)을 알아봄으로써 요인분석에 사용한 표본이 충분하였는지 살펴보고, 모형 적절성 판단과 최종 모형의 선택 및 수정을 위해 근거한 준거를 검토하였다. 또한 최종 모형에 기반하여 요인부하량과 요인 간 상관 정보를 고지하였는지를 포함하여 번안검사 타당도에 대한 충분한 정보를 제공하였는지도 확인하였다. 체계적 검토를 통해 산출한 모든 통계량은 전체 107개 연구에 기반하였으나, 각 연구에서 다수의 값을 보고했을 경우(예: 하위 요인에 대한 신뢰도 분석을 추가로 실시하여 보고), 해당 연구 내의 평균값을 사용하여 전체 통계량 분석에 활용하였다. 이와 같은 체계적 검토 전략에 기반하여 분석한 결과는 다음 장에 보고하였다.

결 과

검사 정보

107개의 번안검사 중 총 75편의 연구가 원 검사와 최종 번안검사의 요인구조나 문항이 다르다고 보고하였으며, 이 중 14편의 연구는 번안의 유연성을 위해(예: 국내 정서에는 적합하지 않은 문항: N = 5) 또는 통계적 기준을 충족하기 위해(예: 문항이 정규성 기준을 충족하지 못함: N = 11) 요인분석 이전에 삭제한 문항이 있다고 밝혔다. 또한 총 95편의 연구가 최종 번안검사의 문항을 보고했으나, 12편의 연구는 보고하지 않은 것으로 나타났다.

검사도구의 신뢰도 및 타당도

표 3은 검토 대상 연구가 다룬 심리측정적 속성을 나타낸다. 107개의 번안 검사 중 총 105편의 연구가 검사 문항의 내적 일치도를 나타내는 Cronbach's α를 신뢰도로 보고하였으며, 105편의 연구에 기반한 번안검사의 신뢰도 평균은 .87이었다(최댓값=.98, 최솟값=.55). 신뢰도의 계산은 각 연구가 보고한 전체 검사의 α에 기반하였으며, 각 하위 척도의 α만을 보고 한 경우에는 하위 척도의 평균값을 전체 검사의 α로 고려하였다. 더불어, 총 28편의 연구가 검사-재검사 신뢰도를, 6편의 연구는 반분 신뢰도를 확인하였다. 이 외에도 6편의 연구에서 ω 계수를 보고하였고 해당 신뢰도 평균은 .88이었다(최댓값=.97, 최솟값=.77).

번안 검사는 구성개념에 대한 측정을 주목적으로 하므로 구성 타당도의 확인이 필수적이다. 107편의 모든 연구가 구성 타당도를 확인하였다고 보고하였으며, 대다수의 연구가 요인분석을 기반으로 문항과 요인 간 관계를 살펴봄으로써 내적 접근에 기반하여 타당도를 평가한 것으로 보인다. 107편의 연구 중 추가로 외적 접근에 기반한 구성 타당도를 확인한 연구도 있었으며, 수렴 타당도와 변별 타당도를 확인한 연구는 각각 83편과 63편이었다. 그 밖에, 준거 타당도를 살펴본 연구가 63편이었으며, 특별히 준거 타당도의 일종인 공존 타당도와 예측 타당도를 조사했다고 명시한 연구는 각각 28편과 6편이었다. 또한 증분 타당도와 내용 타당도를 확인한 연구는 각각 18편과 7편으로 나타났다. 이와 같은 결과는 연구자의 주관적인 분류에 기반한 것이 아니라 각 번안연구에서 언급하고 정의한 타당도에 근거하여 검토하고 분석한 결과임을 밝혀둔다.

표 3. 검사도구의 신뢰도 및 타당도 보고 현황

심리측정 속성	N
신뢰도	107
Chronbach's α	105
검사-재검사 신뢰도	28
반분 신뢰도	6
ω	6
구성 타당도	107
수렴 타당도	83
변별 타당도	63
준거 타당도	63
공존 타당도	28
예언 타당도	6
증분 타당도	18
내용 타당도(문화적 타당도)	7

요인분석의 시행과 표본크기

대부분의 번안 및 타당화 연구가 구성 타당

도의 확인을 위해 탐색적 요인분석과 확인적 요인분석을 사용하였다고 밝혔다. 탐색적 요인분석과 확인적 요인분석을 모두 실시한 연구의 경우 70편이며 반면 한 가지 요인분석만 실시한 연구는 34편(탐색적 요인분석 9편, 확인적 요인분석 25편)이었다. 각 요인분석 모형에 따라 살펴보면 총 79편의 연구가 탐색적 요인분석을, 95편의 연구가 확인적 요인분석을 통해 타당도를 검증하였다고 볼 수 있다. 탐색적 및 확인적 요인분석을 모두 실시하지 않은 연구는 3편이었으며, 상관분석을 통해 구성 타당도를 확인했다고 보고하였다.

각 요인분석에 사용된 표본크기를 표 4에 요약 및 정리하였다. 분석 결과, 평균적으로 탐색적 요인분석과 확인적 요인분석에 사용된 표본크기는 약 283와 407로 확인적 요인분석에 사용된 표본크기가 큰 것으로 나타났다. 표본크기의 최솟값은 탐색적 요인분석의 경우 100, 그리고 확인적 요인분석의 경우 124로 그 크기가 비슷하였으나 최댓값은 각각 846과 2,478로 큰 차이를 보이는 경향이 있었다. 추가로, 본 연구는 문항 개수 대비 표본크기, 즉 *N:p*를 확인함으로써 모형의 복잡도 대비 표본크기를 검토하였으며, 평균적으로 탐색적 요인분석(16.72)보다 확인적 요인분석(30.56)의 *N:p* 비율이 높았다.

탐색적 요인분석 세부 사항

탐색적 요인분석을 실시한 79편의 연구 결과에 기반하여 모형의 설정, 평가, 그리고 수정과 관계된 사항을 표 5에 정리하였다. 요인추출법으로는 최대우도법(n = 38)과 주축분해법(n = 29)이 가장 많았으며, 주성분 분석과 가중 최소 제곱법을 사용하였다고 보고한 연구가 각각 5편, 3편, 그리고 추출법을 보고하지 않은 연구가 4편 있었다. 회전 방법을 보고한 연구는 79편 중 74편으로 사각회전(n = 65)을 적용한 연구가 직각회전(n = 7)을 적용한 연구보다 많았다. 사각회전 중에서는 직접 오블리민(n = 38), 프로맥스(n = 16), Geomin(n = 7) 순서로 빈번히 사용되었으나, 구체적으로 어떤 사각회전을 택하였는지 명시하지 않은 연구도 4편 있었다. 직각회전의 경우 4편의 연구가 배리맥스를 적용하였다고 하였으며, 2편의 연구의 경우 회전법을 중복으로 보고하는 등 사용한 회전법이 명확하지 않기에 기타로 분류하였다. 나머지 5편의 연구는 회전법

표 4. 요인분석에 사용된 표본크기

	Mean	SD	Minimum	Maximum
탐색적 요인분석 (79편)				
표본크기	282.88	126.92	100	846
N:p	16.72	15.47	1.06	112.60
확인적 요인분석 (95편)				
표본크기	406.64	349.28	124	2,478
N:p	30.56	52.22	4.63	495.60

참조. N:p는 표본크기:문항개수의 비율을 나타냄.

표 5. 탐색적 요인분석 모형의 설정과 평가(*N*=79)

보고 사항	*N*
요인 추출법	총 75편 보고
최대우도법	38
주축분해법	29
주성분분석	5
가중 최소 제곱법	3
회전 방법	총 74편 보고
사각회전	65
직접 오블리민(Direct Oblimin)	38
프로맥스(Promax)	16
Geomin	7
직각회전	7
배리맥스(Varimax)	5
기타	2
모형의 평가 및 요인개수 결정 준거	
고유치	61
스크리 도표	57
해석 가능성	31
설명량	23
평행검사	21
모형 합치도	13
요인부하량	5
요인 내 최소 문항 수	2
모형의 수정 및 문항 선택 준거	
요인부하량	62
요인부하량 ≥.30	18
요인부하량 ≥.40	32
요인부하량 ≥.50	6
요인부하량 ≥.60	1
기타 (구체적인 준거 미보고)	3
교차부하	38
공통분	13
이론적 근거	9
문항 변별도(총점과 문항 간의 상관)	7
신뢰도 증감	5
모형합치도	2

을 보고하지 않았다.

각 번안검사 연구는 다양한 준거를 이용하여 모형을 평가하고 요인개수를 결정하였으며, 평균적으로 약 3개(평균 = 2.56)의 준거를 사용하는 것으로 나타났다. 표 5를 살펴보면 고유치(*n* = 61), 스크리 도표(*n* = 57), 해석 가능성(*n* = 31), 설명량(*n* = 23), 평행검사(*n* = 21), 모형합치도(*n* = 13), 요인부하량(*n* = 5), 그리고 요인 내 최소 문항 수(*n* = 2) 순으로 빈번하게 고려하여 요인개수를 결정하였음을 알 수 있다. 하나의 준거만을 살펴보고 모형을 결정한 연구는 11편이었다.

모형의 수정 및 문항 선택을 위한 준거로는 요인부하량(*n* = 62)을 가장 많이 참고한 것으로 나타났으며, 부하량 θ.40 이상을 보이는 문항을 타당한 문항으로 판단한 연구가 가장 많았다(*n* = 32). 이외에도 문항의 교차부하 여부(*n* = 38), 공통분산의 크기(*n* = 13), 이론적 근거(*n* = 9), 그리고 문항 변별도(*n* = 7), 문항의 포함에 따른 신뢰도의 증감 여부(*n* = 5), 모형 합치도(*n* = 2)에 근거하여 적합하지 않은 문항을 삭제하고 최종 모형을 추정하였음을 알 수 있었다.

확인적 요인분석 모형의 설정과 평가

확인적 요인분석을 실시한 연구는 총 95편이었다. 각 연구는 구조방정식모형의 모형합치도에 기반하여 모형을 평가하였으며, 95개 번안검사의 최종 요인구조에 대한 모형 합치도지수의 정보를 요약하면 표 6과 같다. 선행연구에서 제시한 준거(RMSEA, SRMR <0.08; CFI, TLI > .95)에 적합하지 않은 모형합치도 값을 보이는 모형도 있었는데, 다수의 합치도지수를 모두를 고려한 연구도 있는 반면 그중

표 6. 확인적 요인분석 세부 사항(N=95)

	N	Mean	*SD*	Minimum	Maximum
모형의 평가					
RMSEA	94	.07	.02	.01	.19
CFI	95	.92	.05	.56	.99
TLI	86	.91	.06	.54	.99
SRMR	41	.06	.01	.03	.08
모형의 수정 및 문항 선택 준거					
모형 간 모형합치도 비교			44		
수정지수 검토			19		
요인부하량			15		
이론적 배경			13		
ω 계수			3		
다중상관자승			2		
표준잔차행렬			2		
요인 간 상관계수			1		

일부에만 기반하여 모형의 적절성을 판단한 경우도 있었기 때문이다.

탐색적 요인분석과 마찬가지로 확인적 요인분석도 모형의 수정과 문항 선택을 위한 판단이 필요하다. 검토 결과, 44편의 연구가 모형 간 모형합치도 비교를 통해 모형의 수정 여부를 선택하였고, 수정지수를 이용한 연구가 19편이었다. 또한 최종 문항의 선택과 관련해서는 이론적 배경과 요인부하량을 고려한 연구가 각각 15, 13편, ω 계수, 즉 신뢰도를 고려한 연구가 3편이었다. 이 외에도 다중상관자승과 표준잔차행렬을 토대로 문항을 선택했다고 보고한 연구는 각각 2편씩 있었는데, 요인부하량의 크기에 따라 증감하는 공통분산과 고유분산을 고려한 것으로 요인부하량을 살피는 것과 같은 맥락이라고 볼 수 있다. 그밖에 모형 수정을 위해 요인 간 상관을 고려한 연구는 1편이었다.

최종 요인구조의 보고 요인부하량과 요인 간 상관계수 보고

연구자는 최종 요인분석 모형에 기반하여 측정모형의 정보를 보고할 필요가 있으며, 문항-요인 간의 관계를 나타내는 요인부하량과 요인 간 상관계수(다요인 모형의 경우)를 보고한다. 본 연구는 검토 대상 연구가 보고한 최종 요인분석 모형의 정보를 살펴보고 그 결과를 표 7에 정리하였다.

먼저 요인부하량 정보를 살펴보면, 탐색적 요인분석을 사용한 사례(n=79) 중 78편의 연구가 요인부하량을 표나 그림을 이용하여 보

표 7. 측정모형의 보고

요인분석	*N*
탐색적 요인분석(79편)	
요인부하량 보고	78
요인 간 상관 보고 (1요인:11편, 다요인:68편)	
상관계수와 유의성	19
상관계수만 보고	2
모두 보고하지 않은 논문	47
확인적 요인분석(95편)	
요인부하량 보고	75
요인 간 상관 정보 (1요인:9편, 다요인:82편 고차요인:4편)	
상관계수와 유의성	50
상관계수만 보고	17
모두 보고하지 않은 논문	15
기타	7

고하였고, 1편의 연구는 관련 정보를 보고하지 않았다. 확인적 요인분석(n=95)의 경우에는 75편의 연구가 요인부하량을 보고하였으나 20편의 연구는 보고하지 않았다. 요인 간 상관의 보고 여부는 단일요인 사례(탐색적 요인분석: 11건, 확인적 요인분석: 9건)를 제외하고 다요인을 분석한 사례에 대해서만 그 현황을 살펴보았다. 68개의 탐색적 요인분석 모형에 대하여, 요인 간 상관계수와 그 유의성을 모두 보고한 연구는 19편, 상관계수만 보고한 연구는 2편이었으며, 모두 보고하지 않은 연구는 47편으로 가장 많이 나타났다. 82개의 다요인 확인적 요인분석 모형에 대해서는 요인 상관계수와 그 유의성을 모두 보고한 연구가 50편으로 가장 많았고, 상관계수만 보고한 경우는 17편, 그리고 모두 보고하지 않은 연구는 15편이었다. 요인 간 상관계수를 보고했으나, 어떤 요인분석에서 기반한 결과인지 명시하지 않은 경우는 기타로 분류하였고 7편의 논문이 이에 해당하였다.

검토 대상 중, 김광태 등(2023)의 연구는 앞서 살펴본 사항들을 전반적으로 다루고 다양한 근거에 기반하여 번안검사를 타당화한 연구로 볼 수 있다. 번안, 신뢰도 및 타당도 검증을 위한 절차를 자세히 기술하고 있으며 의사 결정을 위해 근거한 준거와 연구 결과에 대해서도 구체적으로 보고하고 있다. 특별히 준거를 복합적으로 검토하여 의사를 결정하고, 이에 대해 구체적으로 보고하고 있다는 면을 주목할 필요가 있다.

논의 및 제언

본 연구는 국내 심리검사 번안의 검수와 보고에 대해 제언하는 것을 목표로 2017년부터 2023년까지의 한국 심리검사 번안 타당화 연구를 대상으로 체계적 검토를 실시하고 타당화 현황을 살펴보았다. 먼저 교육 및 심리검사의 표준과 ITC 검수 지침에 기반하여, 번안검사의 심리측정 속성에 대한 정보가 충분히 보고되고 있는지 또한 각 연구가 타당도 확보를 위해 적합한 분석을 실시했는지 조사하였다. 특별히 번안검사 타당화 검수 작업에서 사용된 요인분석의 시행과 결과보고에 대해 면밀히 살펴보았다.

요약하자면, 검토 대상 중 두 편을 제외한 모든 번안검사 타당화 연구가 신뢰도를 보고하였고 측정의 타당성을 근거하기 위해 구성 타당도를 검증하였음을 알 수 있었다. 외부검사와의 관계성에 기반한 구성 타당도, 즉 수

렴 타당도와 변별 타당도를 추가로 검증한 연구는 각각 83, 63편이었다. 준거변수와의 관계에 기반하여 준거 타당도를 고려한 연구는 63편으로, 이 중 28편과 6편의 연구가 구체적으로 각각 공존 타당도와 예언 타당도를 살펴보았음을 언급했다. 검사 번안에 대한 비교 문화 심리학 분야의 관심이 높아지고 있는데(Gregoire, 2018), 이는 검사를 문화적 배경에 더욱 적합한 검사로 각색하고 활용해야 하기 때문이다. 이에 따라 7편의 연구가 문항내용의 타당성과 문화적 적합성에 근거하여 내용 타당도를 참작하였음을 직접 보고하였다. 좀 더 세부적인 검토 결과에 기반한 논의와 제언은 신뢰도, 타당도, 그리고 요인분석 순으로 아래에 언급하고자 한다.

대다수의 연구가 Cronbach's α를 신뢰도로 보고한 것과 관련하여, 본 연구는 α 계수에 대한 지나친 의존에서 벗어날 필요가 있다는 점을 언급하고자 한다. α 계수는 다수 문항이 한 검사를 구성할 때 산출하는 지수로, 전체 검사점수의 분산에 대비한 문항 간 공분산 평균[2]의 크기에 기반하여 산출한다(Geldhof, Preacher, & Zyphur, 2014). 이는 문항 간 내적 일관도를 나타내는 지수로 전체 분산에서 공통요인에 기반한 문항 간 공분산이 차지하는 비율로 해석할 수 있으므로 신뢰도를 의미한다. 그러나 엄밀히 말한다면 α 계수는 타우동형(tau-equivalent) 검사 조건을 만족하지 않는 검사에 대해 제약이 있다. 타우동형 검사 조건을 충족하는 검사는 모든 문항의 표준화 요인부하량이 거의 같은 크기여야 하지만 이를 만족하는 검사는 현실에서 찾아보기 어렵기 때문이다. 또한 타우동형 검사 조건을 만족하지 않는 경우에 α 계수는 과소 추정되기 때문에(McNeish, 2018; Raykov & Marcoulides, 2011) 선행연구는 α 계수의 대안으로 McDonald (1999)의 ω 계수를 언급한다. 본 연구가 검토한 번안연구 중에도 과소추정 문제로 ω 계수를 보고한 연구를 살펴볼 수 있었다(정서영, 박희웅, 손우영, 2023). ω 계열의 신뢰도는 타우동형 검사보다 제약이 적은 동일구조(congeneric) 검사에 기반하고 있으며 모든 문항의 요인부하량이 동일하지 않아도 산출할 수 있다(이순묵, 조은경, 1998; McNeish, 1995). 그러나 아직까지 α 계수의 보고가 주를 이루고 있는 것으로 보이며, 한국심리학회 산하 학술지에서도 ω 계수를 보고한 연구는 적은 편인 것으로 나타났다. 과거에는 소프트웨어의 접근성을 이유로 α 계수의 보고가 주를 이룬 것으로 보이나 현재는 R, jamovi와 같은 소프트웨어에서 쉽게 산출할 수 있으므로 연구자들은 ω 계수의 보고를 활발히 할 필요가 있다. 본 연구는 ω 계수를 소개하는 것에서 글을 마치지만, ω 계열의 신뢰도를 자세히 알고 싶은 연구자는 신재은과 이태헌(2017), 그리고 McNeish(1995)의 연구를 참고하기 바란다.

본 연구는 앞서 교육 및 심리검사의 표준과 미국심리학회의 정의를 따라 구성 타당도와 준거 타당도를 중점적으로 설명하였다. 그러나 증분 타당도를 검증하고 보고한 연구의 수(18편)가 적지 않은 것을 확인하고 후속 연구자들을 위해 설명을 보충하고자 한다. 증분 타당도는 예측력을 논의하는 평가 현장에서 활용되고 있으며(Clark & Watson, 2019; Hunsley & Meyer, 2003), 인사 혹은 임상 현장에서 근거하는 타당도의 종류로 볼 수 있다. Hunsley

2) $\alpha = (p^2\overline{\sigma_{ij}})/\sigma_X^2$, p = 검사 문항 수; $\overline{\sigma_{ij}}$ = 문항 i와 j 간의 공분산 평균; σ_X = 척도 점수의 분산

와 Meyer(2003)는 증분 타당도가 구성 타당도의 추가 근거로 활용되고 있음을 언급한 바 있다. 예를 들어, 기존 풀 배터리 검사(full battery test)에 새로운 소검사나 문항을 추가하는 경우, 개정한 검사가 기존 검사보다 더 많은 정보를 제공해야 증분된 타당도를 갖췄다고 볼 수 있다.

다음으로 요인분석 모형의 선택에 관하여 논의하고자 한다. 번안검사는 선행 타당화 작업을 거친 원검사를 바탕으로 하지만, 번안에 따른 검사의 개정이 이루어지므로 별도의 타당화 작업이 필요하다(AERA, APA, NCME, 2014; Gregoire, 2018). 원검사와 번안검사의 심리측정적 속성은 다를 가능성이 높아서 번안검사에 대한 추가 검증이 필수적이기 때문이다. 연구자는 언어적/문화적 차이에 의해 번안검사의 요인구조가 원검사와 다를 수도 있다는 점에 유의하고, 이를 반영한 검수 방법을 택해야 한다. 실제로 본 연구가 검토했던 107편의 번안검사 타당화 연구 중 75편의 연구가 원검사와 번안검사의 요인구조 간에 차이가 있음을 보고하였다. 이러한 맥락에서 볼 때, 연구자는 번안검사가 원검사와 다른 요인구조를 가질 수 있음을 가정하고 타당화 작업을 시행하는 것이 적합한 것으로 보인다. 총 107편의 번안연구 중 탐색적 요인분석을 활용하여 번안검사의 요인구조를 새롭게 탐색하고 확인적 요인분석을 통해 이를 교차 타당화한 연구는 70편이었다. 34편의 연구는 두 가지 요인분석 중 한 가지만 이용하여 요인구조 모형을 결정하였으나, 원검사와 다른 요인구조를 예상하는 경우에는 타당도에 대한 더욱 충분한 근거를 확보하기 위하여 탐색 및 확인적 요인분석을 모두 실시하는 것이 적절할 수도 있다.

체계적 검토를 통하여 요인분석이 번안검사 타당화 연구에서 거의 필수적으로 사용하고 있는 방법임을 다시 한번 확인할 수 있었다. 그러나 국내 번안 타당화 연구 중 일부는 요인분석과 관계된 사항을 충분히 보고하지 않고 있었기에, 응용 연구자들에게 도움이 되고자 몇 가지 제언을 하고자 한다. 첫째, 검사 타당화에 요인분석을 사용하는 연구자는 안정적인 추정을 위해 충분한 크기의 표본을 확보할 필요가 있다. 검토 결과를 살펴보면, 탐색적 요인분석의 경우 *N:p*의 최소값은 1.06로 선행연구에서 제시한 문항 수 대비 5배의 표본크기에 한참 미치지 못하였으며, 최소 표본크기는 100으로 선행연구가 제시한 최소 표본크기(300)보다 매우 작았다. 확인적 요인분석은 좀 더 충분한 표본크기에 기반했으나 역시 주의할 필요가 있는 것으로 보인다. *N:p*의 최소값이 4.63로 이는 Kline이 제시한 10보다 작은 비율이었으며, 최소 표본크기는 124로 일반적으로 구조방정식 모형에서 고려하는 작은 표본크기(200)보다 작았다. 연구자들은 연구설계 단계에서 문항/모수의 수를 고려하여 충분한 표본크기를 계획해야 하며, 이를 확보하기 위해 노력해야 함을 다시 한 번 강조한다.

둘째, 요인분석 이후에는 최종 요인구조를 보고함으로써 모형의 타당성을 뒷받침하고 다른 연구자들의 이해를 도울 필요가 있으나, 이에 대한 보고가 충분하지 않은 경우가 있음을 확인할 수 있었다. 요인부하량은 문항에 대한 요인의 공통분산을 확인할 수 있는 정보이며 요인 간 상관계수는 추출된 요인이 이론에 적합한지의 여부, 또는 요인의 추출이 과다하지는 않은지의 여부(예: 요인 간 상관이 너무 큰 경우)를 판단할 수 있는 정보다. 그러나 모든 연구가 이에 해당하는 결과를 보고하

지는 않았으며 특히, 요인 간 상관의 경우 탐색적 요인분석은 19편, 그리고 확인적 요인분석의 경우 50편의 연구만 상수계수와 그 유의성을 보고하였다. 번안검사의 타당성을 주장하고 후속 연구자들에게 검사 사용을 권장하기 위해서는 필요한 정보를 충분히 보고하는 연구자들의 노력이 필요하다.

셋째, 탐색적 요인분석의 요인구조의 회전과 요인개수의 결정 과정에 대해서 추가적으로 제안하고자 한다. 본 연구에서 검토한 결과 탐색적 요인분석을 실시한 79건의 사례 중 6건이 직각회전을 실시하였으나 그 이유에 대한 설명이 충분하지 않은 경우가 대부분이었다. 그러나 사회과학 연구에서는 요인 간 상관이 0이라는 가설이 적합하지 않은 경우가 많으므로(이순묵, 1994), 요인 간 상관을 0으로 가정하고 직각회전을 실시한 것에 관해 설명이 필요하다. 추가로, 요인개수에 대한 결정을 단일 준거에만 기반하는 것은 지양할 필요가 있음을 강조하고자 한다. 본 연구의 검토 결과, 단일 준거에 기반한 연구가 총 11편 있었으며 그 중 카이저 룰에만 기반하여 요인모형을 결정한 사례가 2편임을 확인하였다. 공통적으로 선행연구들은 여러 개의 준거를 함께 고려한 요인개수 결정을 추천하고 있으며, 특히 카이저 룰은 단일 준거로의 사용이 적합하지 않다는 것이 반복적으로 확인되었으므로 (Brown, 2015) 연구자들의 주의가 필요하다.

넷째, 준거값에 가깝지 않은 합치도지수에 기반하여 확인적 요인분석 모형을 평가한 사례들이 있어 이에 대해 언급하고자 한다. 최종 모형의 CFI와 TLI의 최솟값이 각각 .56과 .54로 굉장히 낮은 연구 사례와 RMSEA의 최댓값이 .19로 비교적 높은 사례가 있었다. 모형의 적합성은 다수의 모형 합치도지수에 기반하여 판단하므로, 연구자는 일부 합치도지수가 적합하지 않아도 적합한 나머지에 기반하여 전반적인 평가를 실시했을 수도 있다. 그러나 CFI와 TLI는 대상 모형이 가장 제약이 많은 모형과 비교했을 때 상대적으로 얼마나 나은지를 나타내는 상대적 합치도 지수이며, RMSEA와 SRMR은 자료와 모형 간에 합치하는 정도를 살피는 절대적 합치도 지수로서 모형의 서로 다른 면모를 평가한다. 따라서 연구자는 전반적 모형 평가를 위해서 절대적 합치도 지수와 상대적 합치도지수 모두 적정 준거점수를 만족하고 있는지 살펴볼 필요가 있음을 강조하고자 한다. 다양한 모형 합치도 지수를 종합적으로 판단하는 방법은 김수영(2017), 홍세희(2000), 그리고 Kline(2013) 연구를 통해 살펴볼 수 있다.

본 연구는 번안검사 연구가 주로 요인분석을 사용하고 있는 것을 확인함과 동시에 타당도 확보를 위한 통계모형이 전통적인 요인분석에만 머물러 있다는 점을 알 수 있었다. ITC 지침 또한 이미 집단 간 동일한 번안검사의 타당화에 대해서 강조하고 측정 동일성의 확보를 위한 검증법을 자세히 기술한 바 있으나 국내 107편의 연구 중에 측정 동일성을 검증한 연구는 단 12편이었다. 더욱 다양한 각도에서의 타당도 검증을 실시하고 이에 기반하여 풍부한 근거를 확보하는 국내 연구자들의 노력이 필요한 것으로 보인다.

마지막으로 심리학 연구에서 사용하는 학술용어의 통합을 제언한다. 한국심리학회에서는 영문 용어에 해당하는 한글 용어를 제공함으로써 연구자들의 공통된 용어 사용을 지향하고 있는 것으로 보이지만, 검토 결과 같은 개념에 대해 서로 다른 용어를 사용하는 경우가 빈번한 것을 확인하였다. 예를 들어, 심리학

용어사전(https://www.koreanpsychology.or.kr/psychology/term.html)에는 구성 타당도로 등재된 용어를 구인 타당도로, 공존 타당도로 등재된 용어를 공인 타당도로 보고하는 사례들을 들 수 있다. 또한 사전에 등록되지 않았으나 이미 널리 쓰이고 있는 개념의 경우에는 연구자끼리 서로 다른 용어를 빈번하게 사용했다. 그 예로는 orthogonal rotation에 대한 직각회전과 직교회전, content validity에 대한 내용 타당도와 문화적 타당도를 들 수 있다. 이에 따라 본 연구에서 체계적 검토를 실시하는 과정에서도 용어에 해당하는 원개념을 추적하기 위해 추가로 시간을 소비해야만 했는데, 다른 연구자들도 동일한 어려움을 겪을 수 있을 것으로 예상한다. 따라서 이와 같은 혼동을 피하기 위해서라도 한국심리학회 산하의 학술지에서 출판하는 연구논문은 이미 등재된 용어를 사용할 필요가 있으며, 새로운 용어는 계속해서 사전에 등재하는 등의 노력이 필요해 보인다.

마무리하며, 본 연구는 검사 번역/번안을 위한 ITC 지침에 기반하여 국내 번안 심리검사 타당화 연구의 현황을 검토하고 번안검사 타당화 연구가 고려해야 할 사항에 대해 제언함으로써 연구자들에게 도움이 되고자 하였다. 그러나 새로운 검사의 타당화 과정 또한 번안이라는 과정을 제외하고는 번안검사 타당화 연구와 마찬가지로 심리측정적 속성의 검증을 중요시하므로, 본 연구의 검토 결과 및 제언은 번안검사 뿐만 아니라 새로운 검사의 개발 및 타당화를 목표로 하는 연구자들에게도 도움이 될 것으로 기대한다. 체계적 검토 대상에 포함된 우수한 연구들[3] 또한 참고문헌에 표기하였으므로, 실제 사례를 살피는데 도움이 될 것으로 본다.

참고문헌

*강규림, 현명호 (2022). 한국판 실존불안 질문지 타당화 및 관련 변인 탐색. 한국심리학회지: 임상심리 연구와 실제, 8(4), 683-711. https://doi.org/10.15842/CPKJOURNAL.PUB.8.4.683

*강병은, 신현숙 (2017). 청소년 자의식 척도의 타당화. 한국심리학회지: 학교, 14(1). 105-128. https://doi.org/10.16983/kjsp.2017.14.1.105

*강수경, 김해미, 정미라 (2017). 한국판 태아애착 척도(MAAS)의 타당화 연구. 한국심리학회지: 여성, 22(2), 89-112. https://doi.org/10.18205/kpa.2017.22.2.001

강태훈, 김명연 (2023). 이해하기 쉬운 교육평가. 서울: 박영story.

*강지훈, 김재웅, 석정호, 구본훈, 류진선, 신현경, 윤석호, 홍혜정, 최현정 (2023). Validation of the Korean version of the Borderline Symptom List Short Version (K-BSL-23). Korean Journal of Clinical Psychology, 42(2), 23-34. http://doi.org/10.15842/KJCP.PUB.42.2.23

*강효신, 김빛나 (2021). 미래시간조망 척도의 탐색적 및 확인적 요인분석과 연령 집단 비교. 한국심리학회지: 발달, 34(1), 63-80. https://doi.org/10.35574/KJDP.2021.3.34.1.63

*권두리, 신나나 (2021). 유아기 일상의 실행기

3) 참고문헌 목록에 표기된 *표기는 본 연구가 검토한 번안 타당화 연구이므로, 구체적인 예시가 필요한 연구자는 해당 문헌을 확인하기 바람.

능 척도(REEF)의 타당화 연구. 한국심리학회지: 발달, 34(4), 67-90.
https://doi.org/10.35574/KJDP.2021.12.34.4.67

*권은정, 김율리, 김미리혜, 곽경화, 양재원 (2023). 한국판 ICD-11 성격장애 심각도 평가 (PDS-ICD-11)의 타당화. 한국심리학회지: 건강, 28(1), 205-226.
https://doi.org/10.17315/kjhp.2023.28.1.0

*권혁진, 권석만 (2017). 한국판 자해기능 평가지(The Functional Assessment of Self-Mutilation)의 타당화 연구: 대학생을 중심으로. . 한국심리학회지: 임상심리 연구와 실제, 3(1), 187-205.

*김경아, 장혜인 (2022). 모호한 상실의 측정: 북한이탈주민 경계모호성 척도(BAS-NK)의 예비 타당화 연구. *한*국심리학회지: 일반, 41(4), 349-386.
https://doi.org/10.22257/kjp.2022.9.41.4.349

*김경훈, 전영민, 이길전 (2018). 한국어판 단도박변화촉진활동척도(K-POCS-G)의 타당화 연구. 한국심리학회지: 중독, 3(1), 1-18.
https://doi.org/10.23147/ADDICTPSY.PUB.3.1.1

*김광태, 이혜원, 손영우 (2023). 다차원적 조용한 사직 척도(MQQS) 타당화 연구. 한국심리학회지: 산업 및 조직, 36(4), 557-583.
https://doi.org/10.24230/kjiop.v36i4.557-583

*김다혜, 안정광 (2021). 한국판 사회불안 안전행동 질문지(SBQ) 타당화 연구. Korean Journal of Clinical Psychology, 40(3), 280-298.
https://doi.org/10.15842/kjcp.2021.40.3.009

*김도희, 김희정, 정주리 (2022). 한국판 청소년용 플로리시 척도 타당화 연구: EPOCH 모델을 중심으로. 한국심리학회지: 학교, 19(3), 187-213.
https://doi.org/10.16983/kjsp.2022.19.3.187

김미림 (2023). 영과잉 자료의 측정동일성 검증: 다집단 확인적 요인분석과 2-부분 요인분석 모형의 적용과 비교. 교육평가연구, 36(3), 445-472.
http://dx.doi.org/10.31158/JEEV.2023.36.3.445

김수영 (2016). 구조방정식 모형의 기본과 확장: Mplus 예제와 함께. 학지사.

*김승민, 박 경 (2017). 한국판 양육 수용행동 척도(Korean Parental Acceptance Questionnaire: K-4-PAQ) 타당화 연구. 한국심리학회지: 건강, 22(3), 531-549.
https://doi.org/10.17315/kjhp.2017.22.3.004

*김유나, 안정광 (2023). 한국판 겸손 반응 척도(K-MRS) 타당화 연구. 한국심리학회지: 사회 및 성격, 37(2), 215-235.
https://doi.org/10.21193/kjspp.2023.37.2.005

*김은하, 김도연, 박한솔, 김수용, 김지수 (2017). 한국어판 정당한 세상에 대한 믿음 척도(Belief in a Just World Scale: K-BJWS)의 타당화. 한국심리학회지: 상담 및 심리치료, 29(3), 689-710.
https://doi.org/10.23844/KJCP.2017.08.29.3.689

*김은하, 김현지 (2020). 한국판 남성에 대한 양가적 태도 척도 타당화 연구. 한국심리학회지: 문화 및 사회문제, 26(4), 525-549.
https://doi.org/10.20406/kjcs.2020.11.26.4.525

*김준현, 유금란 (2023). 한국판 남성 동성애자의 이성애불편감(Heterophobia) 척도(K-HGM)의 타당화. 한국심리학회지: 문화 및 사회문제, 29(1), 25-51.
https://doi.org/10.20406/kjcs.2023.2.29.1.25

*김지은, 안현의 (2021). 한국판 조직배반 척도의 타당화: 조직 내 성폭력 피해를 중심

으로. 한국심리학회지: 여성, 26(2), 99-121.
https://doi.org/10.18205/KPA.2021.26.1.005

*김홍주, 김은영 (2018). 한국판 청소년 성찰기능척도 타당화. 한국심리학회지 : 상담 및 심리치료, 30(2), 297-316.
https://doi.org/10.23844/kjcp.2018.05.30.2.297

*남기은, 이선희 (2022). 다차원 일중독 척도 (Multidimensional Workaholism Scale, MWS) 타당화 연구. 한국심리학회지: 산업 및 조직, 35(1), 65-87.
https://doi.org/10.24230/kjiop.v35i1.65-87

*남지선, 박형인. (2021). 분리된 관심의 척도 타당화 및 직무열의와의 관계 연구. 한국심리학회지: 산업 및 조직, 34(4), 629-662.
https://doi.org/10.24230/kjiop.v34i4.629-662

*류지영, 신희천, 김은하 (2020). 한국판 기본 심리적 욕구에 기반한 대인관계 행동 척도 (IBQ)의 타당화. 한국심리학회지: 상담 및 심리치료, 32(3), 1203-1224.
https://doi.org/10.23844/KJCP.2020.08.32.3.1203

*문선현, 엄진섭, 최원일, AllysonBrothers, 노수림 (2023). 한국판 연령 관련 변화에 대한 인식 척도 타당화: 중년 및 노인을 대상으로. 한국심리학회지: 건강, 28(3), 789-915.
https://doi.org/ 10.17315/kjhp.2023.28.3.010

*문영국, 이종현 (2021). 직무배태성 개념의 확장: 가정배태성 척도 타당화 연구. 한국심리학회지: 산업 및 조직, 34(4), 723-750.
https://doi.org/10.24230/KJIOP.V34I4.723-750

*문찬기, 이세라 (2023). Validating Korean Narcissistic Admiration and Rivalry Questionnaire (NARQ): Its Relations with Big Five, Self-esteem, NPI and Benign and Malicious Envy. Korean Journal of Clinical Psychology, 42(3), 45-57.
https://doi.org/10.15842/kjcp.2023.42.3.001

*박가현, 김시형, 이동훈 (2020). 한국판 정서에 대한 신념 척도의 타당화. 한국심리학회지: 건강, 25(1), 97-114.
https://doi.org/10.17315/kjhp.2020.25.1.006

*박경우, 장혜인 (2021). 한국어판 강박적 성행동 장애 척도(K-CSBD-19)의 타당화 연구. 한국심리학회지: 건강, 26(5), 859-879.
http://dx.doi.org/10.17315/kjhp.2021.26.5.002

*박경우, 유현종, 장혜인, 이상규, 이은지 (2022). 한국어판 음란물 사용동기 척도 (K-PUMS)의 타당화 연구. 한국심리학회지: 건강, 27(5), 763-788.
https://doi.org/10.17315/kjhp.2022.27.5.003

*박기라, 김소은, 양은주 (2018). 한국판 대학생 진로타협 경향성 척도의 타당화 연구. 한국심리학회지: 학교, 15(2), 155-175.
https://doi.org/10.16983/kjsp.2018.15.2.155

*박도담, 유성경 (2019). 한국판 성적 지향 마이크로어그레션 척도 (Korean version of the Sexual Orientation Microaggressions Scale; K-SOMS) 타당화. 한국심리학회지: 상담 및 심리치료, 31(3), 899-927.
https://doi.org/10.23844/kjcp.2019.08.31.3.899

*박선영, 이종은, 이정애, 오강섭 (2023). 한국판 노인불안척도(K-GAS)의 신뢰도 및 타당도 연구. *한*국심리학회지: 임상심리 연구와 실제, 9(2), 235-251.
https://doi.org/10.15842/CPKJOURNAL.PUB.9.2.235

*박세란 (2022). 따뜻함과 안전 초기기억 척도 타당화. 한국심리학회지 : 상담 및 심리치료, 34(2), 413-431.

https://doi.org/10.23844/kjcp.2022.05.34.2.413

*박세란 (2022). 자비적 참여행동 척도 타당화 연구. 한국심리학회지: 건강, 27(3), 563-587.
https://doi.org/10.17315/kjhp.2022.27.3.007

*박용욱, 설정훈, 최진수, 이혜주, 손영우 (2022). 다차원적 일 의미감 척도(CMWS) 타당화 연구. 한국심리학회지: 산업 및 조직, 35(2), 213-245.
https://doi.org/10.24230/kjiop.v35i2.213-245

*박정민, 안현의 (2022). 한국판 과거 연애 관계 사고 척도(Positive and Negative Ex-Relationship Thoughts Scale) 타당화. 한국심리학회지: 문화 및 사회문제, 28(4), 627-659.
https://doi.org/10.20406/kjcs.2022.11.28.4.627

*박주하, 양수진 (2022). 미취학 자녀 양육모를 대상으로 하는 한국판 코로나19 스트레스 척도 타당화. 한국심리학회지:발달, 35(3), 1-19.
https://doi.org/10.35574/KJDP.2022.9.35.3.1

*박홍석, 이정미 (2018). 한국판 상황적 자기인식척도(K-SSAS)의 타당화 연구. 한국심리학회지: 학교, 15(2), 177-196.
https://doi.org/10.16983/kjsp.2018.15.2.177

*배라영, 최지영 (2018). 한국판 관계적 공격성 척도의 타당화 연구: 대학생을 대상으로. 한국심리학회지: 상담 및 심리치료, 30(1), 55-79.
https://doi.org/10.23844/kjcp.2018.02.30.1.55

*서원진, 이수민, 김미리혜, 김율리, 김경희, Chad Ebesutani, 김다미, 황보인, 도현정, 박유진 (2018). 한국판 체중 걱정 척도(Weight Concern Scale)의 타당화. 한국심리학회지: 건강, 23(4), 925-938.
https://doi.org/10.17315/kjhp.2018.23.4.006

*서자경, 이기학 (2017). 한국판 다면적 성실성 척도(K-CCS)의 타당화 연구. 한국심리학회지: 사회 및 성격, 31(4), 51-78.
https://doi.org/10.21193/kjspp.2017.31.4.003

*서장원 (2021). 개정판 데이트 폭력 질문지의 타당화 연구. 한국심리학회지: 건강, 26(1), 109-123.
https://doi.org/10.17315/kjhp.2021.26.1.007

*서종한 (2022). 면담기반 사이코패시 성격 종합평가-기관평가척도(CAPP-IRS) 도구 타당화 연구. 한국심리학회지: 사회 및 성격, 36(1), 25-46.
https://doi.org/10.21193/kjspp.2022.36.1.002

*손옥선, 김진숙 (2021). 한국판 매우 민감한 사람 척도(K-HSPS-18)의 재타당화. 한국심리학회지: 상담 및 심리치료, 33(3), 1049-1075.
https://doi.org/10.23844/kjcp.2021.08.33.3.1049

*손은솔, 송현주 (2022). 한국형 부모의 포괄적 식사양육 실행척도(K-CFPQ)의 타당화 연구. 한국심리학회지:발달, 35(4), 71-92.
https://doi.org/10.35574/KDJP.2022.12.35.4.71

*송연주, 하문선. (2020). 한국 단축형 사랑중독 척도(Korean-Love Addiction Questionnaire-Short Form: K-LAQ-SF) 타당화 연구. 한국심리학회지: 문화 및 사회문제, 26(4), 501-524.
https://doi.org/10.20406/kjcs.2020.11.26.4.501

*송원영 (2021). 한국판 성적 이미지 기반 학대 통념 수용 척도(K-SIAMA)의 타당화. 한국심리학회지: 여성, 26(3), 185-200.
https://doi.org/10.18205/kpa.2021.26.1.009

*신문혜, 이지연 (2020). 한국판 긍정도식척도(YPSQ)의 타당화. 한국심리학회지: 상담

및 심리치료, 32(3), 1125-1151.
https://doi.org/10.23844/kjcp.2020.08.32.3.1125

*신성만, 윤지혜, 조요한, 고은정, 박명준 (2018). 예일음식중독척도 2.0(Yale Food Addiction Scale 2.0) 국내 타당화 연구. 한국심리학회지: 여성, 23(1), 25-49.
https://doi.org/10.18205/KPA.2018.23.1.002

신재은, 이태헌 (2017). 쌍요인(Bifactor) 모형을 이용한 심리척도의 측정적 속성 연구방법 개관. 한국심리학회지: 일반, 36(4), 477-504.
https://doi.org/10.18205/kpa.2018.23.1.002

신진아, 시기자, 성태제 (2021). 검사제작과 분석. 서울: 학지사.

*안명희, 정유선 (2023). 한국판 정신화된 정서성 척도(K-MAS)의 타당화 연구. 한국심리학회지: 상담 및 심리치료, 35(2), 413-440, https://doi.org/10.23844/kjcp.2023.05.35.2.413

*안선미, 현영권 (2023). 대학생을 대상으로 한 비판적 의식 척도 타당화. 한국심리학회지: 문화 및 사회문제, 29(4), 595-616, https://doi.org/10.20406/kjcs.2023.11.29.4.595

*안선영, 황순택 (2019). 한국판 DSM-5 성격질문지 정보제공자보고형(K-PID-5-IRF)의 신뢰도와 타당도. Korean Journal of Clinical Psychology, 38(1), 82-101.
https://doi.org/10.15842/kjcp.2019.38.1.007

*안재경, 최이문 (2020). 한국판 자유의지와 결정론 척도(Free will and Determinism Plus: FAD+)의 타당화 연구. 한국심리학회지: 법, 11(2), 191-210.
https://doi.org/10.53302/KJFP.2020.07.11.2.191

*양나연, 이수정 (2018). 한글판 의미 만들기 척도(K-MMS)의 타당화. 한국심리학회지: 상담 및 심리치료, 30(1), 81-100,
https://doi.org/10.23844/kjcp.2018.02.30.1.81

*양진원, 권석만 (2021). 한국판 보상적 섭식 욕구 척도의 타당화. 한국심리학회지: 건강, 26(6), 985-1003.
https://doi.org/10.17315/kjhp.2021.26.6.002

*양진원, 권석만 (2022). 한국판 음식기대척도(AEFS)의 타당화. 한국심리학회지: 임상심리 연구와 실제, 8(2), 249-271.
https://doi.org/10.15842/CPKJOURNAL.PUB.8.2.249

*오주용, 안도연. (2023). 한국어판 죽음태도 척도 개정판(DAP-R)의 타당화. 한국심리학회지: 임상심리 연구와 실제, 9(2), 345-371.
https://doi.org/10.15842/CPKJOURNAL.PUB.9.2.345

*유지혜, 설경옥 (2018). 한국판 물질주의척도의 타당화 연구. 한국심리학회지: 문화 및 사회문제, 24(3), 385-410.
https://doi.org/10.20406/kjcs.2018.8.24.3.385

*윤정미, 김진영 (2019). 한국판 청소년용 지각된 스트레스 척도의 타당화 연구. 한국심리학회지: 건강, 24(3), 569-586.
https://doi.org/10.17315/kjhp.2019.24.3.003

*이덕희, 남슬기, 이동훈 (2021). 초월적 시간관 척도(TFTPS) 타당화. 한국심리학회지: 건강, 26(6), 961-983.
https://doi.org/10.17315/kjhp.2021.26.6.001

*이덕희, 김성현, 정다송, 이동훈 (2023). 자살사고 속성 척도(Suicidal Ideation Attributes Scale; SIDAS) 타당화 연구. 한국심리학회지: 문화 및 사회문제, 29(1), 1-23.
https://doi.org/10.20406/kjcs.2023.2.29.1.1

*이동현, 김향숙 (2021). 고통 과잉 감내 척도의 타당화 연구: 한국 대학생 집단을 대

상으로. Korean Journal of Clinical Psychology, 40(2), 143-155. https://doi.org/10.15842/kjcp.2021.40.2.003

*이동훈, 엄희준, 이덕희 (2022). 트라우마와 사별 경험에 대한 개인의 의미통합 척도 (K-ISLES): 한국판 타당화를 위한 탐색적 연구. 한국심리학회지: 상담 및 심리치료, 34(3), 719-744. https://doi.org/10.23844/kjcp.2022.08.34.3.719

이동훈, 엄희준, 이덕희 (2022). 한국판 트라우마 분노반응척도-5(DAR-5-K)의 종단 타당화 연구. 한국심리학회지: 일반, 41(2), 133-161. https://doi.org/10.22257/kjp.2022.6.41.2.133

*이보라, Eunjoo Kim (2019). 한국판 일유인가 척도 타당화 연구. 한국심리학회지: 상담 및 심리치료, 31(1), 283-302. https://doi.org/10.23844/kjcp.2019.02.31.1.283

*이선영, 안현의 (2020). 한국판 가족돌봄의무 척도(Filial Responsibility Scale-Adult)의 타당화. 한국심리학회지: 문화 및 사회문제, 26(3),259-282. https://doi.org/10.20406/kjcs.2020.8.26.3.259

*이세라, 신현균 (2018). 한국판 모멸감 척도 (K-HI)의 타당화 연구. Korean Journal of Clinical Psychology, 37(1), 119-129. https://doi.org/10.15842/kjcp.2018.37.1.010

*이소정, 김은하 (2022). 한국판 동성애자/양성애자 다차원 정체성 척도 타당화. 한국심리학회지: 문화 및 사회문제, 28(2), 133-161. https://doi.org/10.20406/kjcs.2022.5.28.2.133

이순묵 (1994). 요인분석의 관행과 문제점. 한국심리학회지: 산업 및 조직, 7(1), 1-27.

이순묵 (1995). SPSS를 사용한 공통요인분석의 문제점. 교육평가연구, 8(1), 5-33.

이순묵, 조은경 (1998). 고전검사이론의 가정검증: 불필요한가 또는 불가능한가. 교육평가연구, 11(2), 98-109.

*이순행, 이희연, 정미라 (2018). 한국판 성인 놀이성 척도(K-APTS) 타당화 연구. 한국심리학회지: 건강, 23(2), 397-425. https://doi.org/10.17315/kjhp.2018.23.2.006

*이승민, 이혜진 (2021). 한국판 신체자비 척도 (BCS)의 타당화. 한국심리학회지: 건강, 26(5), 835-858. https://doi.org/10.17315/kjhp.2021.26.5.001

*이은진, 최보윤, 한주옥 (2022). 한국판 다차원적 문화적 겸손성 척도 타당화 연구. 한국심리학회지: 상담 및 심리치료, 34(2), 461-491. https://doi.org/10.23844/kjcp.2022.05.34.2.461

*이재은, 정보영 (2021). 한국판 칼레이도스코프 경력태도 측정도구 타당화 연구. 한국심리학회지: 산업 및 조직, 34(1), 51-79. https://doi.org/10.24230/kjiop.v34i1.51-79

*이종환, 임종민, 장문선 (2018). 주제통각검사를 활용한 사회인지와 대상관계 척도 타당화 연구. Korean Journal of Clinical Psychology, 37(4), 540-557. https://doi.org/10.15842/kjcp.2018.37.4.007

*이주원, 유정아, 송원영 (2022). 한국판 단축형 어둠의 성격 4요소(SD4-K) 척도의 타당화 연구. 한국심리학회지: 건강, 27(6), 999-1023. https://doi.org/10.17315/kjhp.2022.27.6.010

*이현지 & 김가림, & 권유리, & 신윤정 (2021). 한국형 트랜스젠더에 대한 태도 및 믿음 척도(K-TABS) 타당화 연구. 한국심리학회지: 상담 및 심리치료, 33(3), 1077-1108,

https://doi.org/10.23844/kjcp.2021.08.33.3.1077

*임선영 (2023). 외상 후 성장 척도 확장판 (PTGI-X)의 심리측정적 속성: 한국판 척도의 요인구조와 유용성 재검토. 한국심리학회지: 임상심리 연구와 실제, 9(1), 161-181. https://doi.org/10.15842/CPKJOURNAL.PUB.9.1.161

*임소희, 황순택, 권해수, 김지혜, 박은영, 박중규, 이수정, 이은호, 홍상황 (2018). PAI-A(Personality Assessment Inventory for Adolescent) 재표준화 연구. 한국심리학회지: 임상심리 연구와 실제, 4(3), 435-454. https://doi.org/10.15842/CPKJOURNAL.PUB.4.3.435

*임형민, 이슬아, 권석만 (2021). 한국판 미래 심상과제의 타당화. Korean Journal of Clinical Psychology, 40(1), 103-113. https://doi.org/10.15842/kjcp.2021.40.1.008

*임혜선, 김정윤, 홍혜영 (2023). 한국판 공격자 동일시 척도(K-IAS)의 타당화. 한국심리학회지: 상담 및 심리치료, 35(3), 997-1022. https://doi.org/10.23844/kjcp.2023.08.35.3.997

*장은영 (2021). 한국판 도덕손상 사건 척도 및 도덕손상 경험 척도 개발 연구. 한국심리학회지: 일반, 40(3), 301-327. https://doi.org/10.22257/kjp.2021.9.40.3.301

*장지윤, 안현의 (2018). 한국판 동성애자/양성애자 긍정적 정체성 척도 타당화. 한국심리학회지: 상담 및 심리치료, 30(2), 273-295. https://doi.org/10.23844/kjcp.2018.05.30.2.273

*전영민, 이정임, 박은경 (2018). 치료효과 평가도구로서의 한국판 도박회복추적척도 (GFS-K) 타당화. Korean Journal of Clinical Psychology, 37(2), 225-235. https://doi.org/10.15842/kjcp.2018.37.2.009

*정경미, 김수연 (2017). 한국형 아동 섭식행동 질문지(K-CFQ)의 타당화 연구. 한국심리학회지: 건강, 22(2), 317-338. https://doi.org/10.15842/10.17315/kjhp.2017.22.2.006

*정경미, 양윤정, 정승민, 이경숙, 박진아 (2019). 한국판 부모 양육스트레스 검사 4판 단축형 (K-PSI-4-SF)의 표준화 연구. 한국심리학회지: 건강, 24(4), 785-807. https://doi.org/10.17315/kjhp.2019.24.4.001

*정다송 & 이보라, & 이덕희, & 이동훈 (2023). 사별 경험 이후 사회적 의미 만들기 척도 타당화 연구. 한국심리학회지: 상담 및 심리치료, 35(4), 1361-1397. https://doi.org/10.23844/kjcp.2023.11.35.4.1361

*정서영, 박희웅, 손우영 (2023). 한국판 세대 친화적 조직문화척도(K-WICS) 타당화 연구. 한국심리학회지: 문화 및 사회문제, 29(4), 429-453. https://doi.org/10.20406/kjcs.2023.11.29.4.429

*정수인, 안현의 (2019). 한국판 성인용 놀이성 척도의 타당화. 한국심리학회지: 문화 및 사회문제, 25(4), 353-375. https://doi.org/10.20406/kjcs.2019.11.25.4.353

*정주리. (2021). 한국판 제로섬 신념 척도 타당화 연구. 한국심리학회지: 문화 및 사회문제, 27(3), 285-303. https://doi.org/10.20406/kjcs.2021.8.27.3.285

*조성훈, 권정혜 (2017). 한국판 인터넷 게임장애 척도의 타당화. Korean Journal of Clinical Psychology, 36(1), 104-117. https://doi.org/10.15842/KJCP.2017.36.1.010

*조수현 (2020). 한국판 직무소진평가척도 (K-BAT) 타당화를 위한 예비연구. 한국심리학회지: 산업 및 조직, 33(4), 461-499. https://doi.org/10.24230/KJIOP.V33I4.461-499

*조은호, 최지은, 한지수, 허윤성, 김현경 (2023). 한국형 미디어 리터러시 척도 (K-NMLS)와 한국형 미디어 리터러시 단축형 척도(K-NMLS-SF)의 타당화. 한국심리학회지: 발달, 36(4), 45-64. https://doi.org/10.35574/KJDP.2023.12.36.4.45

*조혜현, 현명호 (2023). 한국판 뒤셀도르프 건강음식집착 척도(K-DOS)의 타당화. 한국심리학회지: 건강, 28(2), 561-579. https://doi.org/ 10.17315/kjhp.2023.28.2.015

*조호진, 이풍가, 마 홍, 옥지수 (2022). 한국어판 단축형 어둠의 3요소 척도(K-Dirty Dozen)의 타당도와 요인구조 연구. 한국심리학회지: 산업 및 조직, 35(2), 299-326, https://doi.org/10.24230/kjiop.v35i2.299-326

주영신, 장승민 (2023). 정렬법을 이용한 범주형 자료의 근사 측정동일성 분석. 한국심리학회지: 일반, 42(2), 119-140. http://dx.doi.org/10.22257/kjp.2023.6.42.2.119

*주현영, 김나래 (2023). 한국판 실수 반추 척도(K-MRS)의 타당화 연구: 대학생을 대상으로. 한국심리학회지: 상담 및 심리치료, 35(3), 971-994. https://doi.org/10.23844/kjcp.2023.08.35.3.971

*지은혜, 조용래, 김선영 (2022). 코로나 19 스트레스 척도(COVID Stress Scales, CSS)의 타당화: 한국 성인표본을 대상으로. 한국심리학회지: 임상심리 연구와 실제, 8(4), 659-681,https://doi.org/10.15842/CPKJOURNAL.PUB.8.4.659

*차윤지, 이은호, 황순택, 홍상황, 김지혜. (2020). 한국어판 개정된 아동발현불안척도 2판의 심리측정적 속성에 대한 연구. Korean Journal of Clinical Psychology, 39(3), 203-214. https://doi.org/10.15842/KJCP.2020.39.3.001

*최영환, 이은호, 황순택, 홍상황, 김지혜. (2020). 한국어판 벡자살사고척도의 신뢰도 및 타당도 연구: 일반 성인 집단을 대상으로. Korean Journal of Clinical Psychology, 39(2), 111-123. https://doi.org/10.15842/KJCP.PUB.39.2.111

*최재광, 한지현, 김민범, 송원영 (2023). 한국판 노모포비아 척도 타당화. 한국심리학회지: 건강, 28(2), 581-600, https://doi.org/10.17315/kjhp.2023.28.2.016

*최정금, 김명소 (2018). 한국판 직무 번영감 척도의 타당화 연구. 한국심리학회지: 산업 및 조직, 31(3), 715-739. https://doi.org/10.24230/KSIOP.31.3.201808.715

*최지선, 이정미 (2018). 한국판 균형적 시간조망척도(K-BTPS)의 타당화 연구. 한국심리학회지: 건강, 23(3), 701-720. https://doi.org/10.17315/kjhp.2018.23.3.006

*최환규, 이정미 (2017). 한국판 일의 의미 척도(K-WAMI)의 타당화 연구. 한국심리학회지: 사회 및 성격, 31(4), 1-25. https://doi.org/10.21193/kjspp.2017.31.4.001

*표소휘, 양은주 (2020). 한국판 학생 진로구성 척도(Korean Student Career Construction Inventory)의 타당화 연구: 초기 성인기를 중심으로. 한국심리학회지: 학교, 17(2), 145-164. https://doi.org/10.16983/kjsp.2020.17.2.145

*표지은, 안정광 (2021). 한국판 사후처리 특성 및 상태 척도(PEPI) 타당화. 한국심리학회

지: 임상심리 연구와 실제, 7(3), 283-308. https://doi.org/10.15842/CPKJOURNAL.PUB.7.3.283

*허민희, 황순택, 박광배, 유성은, 김지혜, 박중규, 이은호, 홍상황. (2021). 한국판 밀론 다축 임상성격검사 4판의 신뢰도와 타당도. Korean Journal of Clinical Psychology, 40(1), 91-102. https://doi.org/10.15842/kjcp.2021.40.1.007

*허주연, 황성훈 (2023). 한국판 Perth 정서 반응성 척도의 타당화. 한국심리학회지: 임상심리 연구와 실제, 9(3), 425-453. https://doi.org/10.15842/CPKJOURNAL.PUB.9.3.425

*현혜민, 박기환 (2018). 한국판 저장 척도 개정판의 타당화 연구. 한국심리학회지: 건강, 23(3), 721-738. https://doi.org/10.17315/kjhp.2018.23.3.007

홍세희 (2000). 특별기고: 구조 방정식 모형의 적합도 지수 선정기준과 그 근거. 한국심리학회지: 임상, 19(1), 161-177.

*홍혜정, 박중규 (2019). 야식증후군 진단질문지(NEDQ)의 타당화 연구. 한국심리학회지: 임상심리 연구와 실제, 5(1), 65-90. https://doi.org/10.15842/CPKJOURNAL.PUB.5.1.65

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual differences, 42*(5), 815-824. https://doi.org/10.1016/j.paid.2006.09.018

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186-3191. https://doi.org/10.1097/00007632-200012150-00014

Brislin, R. W. (1970). Back-translation for crosscultural research. *Journal of Cross-Cultural Psychology 1*(3), 185-216. https://doi.org/10.1177/135910457000100301

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33*, 267-334. https://doi.org/10.1007/BF02289327

Brown, T. A. (2015). Confirmatory factor analysis for applied research(2nd ed.). The Guilford Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Cattell, R. B. (1978). The Scientific Use of Factor Analysis in Behavioral and Life Sciences. New York: Plenum.

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A Propensity Score Method for Investigating Differential Item Functioning in Performance Assessment. *Educ Psychol Meas, 80*(3), 476-498. https://doi.org/10.1177/0013164419878861

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychol Assess, 31*(12), 1412-1427. https://doi.org/10.1037/pas0000626

De Winter, J. C., & Dodou, D. (2012). Factor

recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of applied statistics, 39*(4), 695-710. https://doi.org/10.1080/02664763.2011.610445

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry. 126*(3), 237-240. https://doi.org/10.1192/bjp.126.3.237

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4*(3), 272. https://doi.org/10.1037/1082-989X.4.3.272

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science-Revue Canadienne Des Sciences Du Comportement, 49*(2), 78-88. https://doi.org/10.1037/cbs0000069

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72-91. https://doi.org/10.1037/a0032138

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate behavioral research, 25*(1), 33-39. https://doi.org/10.1207/s15327906mbr2501_3

Gregoire, J. (2018). 검사 번역/번안을 위한 국제 지침: 한국어판 2017년 2판 [ITC guidelines for translating and adapting tests (2nd ed.)] (서동기와 이순묵 역). *International Journal of Testing, 18*(2), 101-134. https://doi.org/10.1080/15305058.2017.1398166

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N: q hypothesis. *Structural equation modeling, 10*(1), 128-141. https://doi.org/10.1207/S15328007SEM1001_6

Jenkins, M., & Griffith, R. (2004). Using Personality Constructs to Predict Performance: Narrow or Broad Bandwidth. *Journal of Business and Psychology, 19*(20), 255-269. https://doi.org/10.1007/s10869-004-0551-9

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatsschneider (Eds.), Applied quantitative analysis in the social sciences (pp. 171-207). New York, NY: Routledge.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-

Loevinger, J. (1957). Objective tests as instruments

of psychological theory. *Psychological reports, 3*(3), 635-694. https://doi.org/10.2466/PR0.3.7.635-694

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

McDonald, R. P. (1999). Test theory: A unified approach. Mahwah, NJ: Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychol Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

*Moon, C., & Lee, S. (2023). Validating Korean Narcissistic Admiration and Rivalry Questionnaire (NARQ): Its Relations with Big Five, Self-esteem, NPI and Benign and Malicious Envy. *Korean Journal of Clinical Psychology, 42*(3), 45-57. https://doi.org/10.15842/kjcp.2023.42.3.00

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior research methods, instruments, & computers, 32*(3), 396-402. https://doi.org/10.3758/BF03200807

Raykov, T., & Marcoulides, G. A. (2011). Introduction to psychometric theory. Routledge.

Thurstone, L. L. (1947). Multiple factor analysis. Chicago: University of Chicago Pres.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327. https://doi.org/10.1007/BF02293557

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health, 8*(2), 94-104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Whittaker, T. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *Journal of Experimental Education, 80*(1), 26-44. https://doi.org/10.1080/00220973.2010.531299

1차원고접수 : 2024. 05. 27
최종게재결정 : 2024. 07. 30

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 261-295
http://dx.doi.org/10.22257/kjp.2024.9.43.3.261

A Systematic Review of Translating and Adapting Psychological Test: Practices and Recommendations

Mirim Kim **Yeji Im**

Korea University

This current study conducted a systematic review of translating and adapting psychological tests based on the Standards for Educational and Psychological Testing and the ITC Guidelines. We reviewed a total of 107 KAPA articles published from 2017 to 2023 and examined validation and reporting practices. The current study examined whether sufficient information about the psychometric properties of the tests was reported and whether each study conducted adequate analyses to ensure validity. Specifically, we reviewed the implementation and reporting of factor analysis for validation. In short, we found that most studies reported reliability and used factor analysis for testing construct validity; therefore, we suggested recommendations in this regard. We also indicated that academic terminologies in psychology research should be unified as the studies often used different terms for the same concept. Although the study results are based on the translating and adapting tests, the findings and recommendations will be also useful for development and validation new tests.

Key words : translating and adapting tests, confirmation guidelines, systematic review, psychometric properties, factor analysis

연구자 윤리 서약 및 저작권 이양에 대한 동의서

제1조 저작물의 표시

논문 제목: _____

제2조 저작재산권의 양도

① 저자(들)는 본 논문에 대한 저작재산권 전부를 한국심리학회에게 양도한다.

제3조 저작재산권을 양도한 후의 저자의 권리 행사

① 저자(들)는 본 논문의 내용으로 특허권 출원, 실용신안권 출원, 디자인권 설정등록, 상표 설정등록을 할 수 있다. 저자는 이 경우를 제외하고는 본 논문을 상품화하기 위하여 논문에 대한 권리를 영리단체에 양도할 수 없다.

② 저자(들)는 교육 또는 개인의 연구 등 개인적인 목적으로 사용하기 위해 논문의 전부 또는 일부를 복제하고 배포할 수 있다.

③ 저자(들)는 논문의 전부 또는 일부를 본인의 개인 웹사이트, 저자가 소속된 기관 및 단체의 웹사이트, 연구비를 지원한 단체의 웹사이트에 게재하고 배포할 수 있다.

③ 위 사항에 대한 이용은 한국심리학회에서 학술지를 발행한 후에 가능하다.

제4조 보증 및 책임

① 본 동의서에 서명함으로써 저자는 다음 사항에 보증한 것으로 본다.

1) 저자(들)는 본 논문에 실질적이고 지적인 공헌을 하였으며 논문의 내용에 대하여 공적인 책임을 공유한다.
2) 논문이 기존에 다른 곳에 공표되지 않았으며 본 학술지에만 제출한 것이다.
3) 논문 내용에 타인을 비방하거나 불법적 문장이 없으며, 타인의 권리를 침해하거나, 피해를 입힐 수 있는 내용이 포함되어 있지 않다.
4) 만약 저작권이 있는 타인의 논문에서 발췌된 내용이 포함된 경우, ′갑′은 그 권리자에게 허락을 받거나 적법한 인용의 범위 내에서 출처를 표시하고 이용한다.

② 본 저작물의 내용이 제3자의 권리를 침해하여 학회 또는 제3자에 대하여 손해를 끼친 경우에는 저자가 그 책임을 진다.

	성 명	소 속	이메일
제 1 저자			
제 2 저자			
제 3 저자			
제 4 저자			
제 5 저자			

※ 논문에 기술된 순서대로 모든 저자의 성명, 소속, 이메일을 기재하여 주십시오.

※ 본 위원회에서 수신한 교신저자의 투고 이메일은 모든 저자들이 연구자 윤리서약 및 저작권 이양에 대한 동의서에 서명날인한 것으로 간주합니다.

※ 심사료와 게재료에 대한 규정을 모든 저자가 확인해주십시오. {관련규정 아래 붙임}

※ 교신저자에게는 다른 공동저자들께 이 저작권 동의서에 기술된 모든 사실을 투고 전에 반드시 알릴 책임이 있습니다.

2024. . .

한국심리학회 귀하

논문작성 양식

작성양식은 한국심리학회에서 기획하여 출판한 “학술논문 작성 및 출판 지침 2판(2012, 박영사)”에 따른다. 그 출판 지침의 일부를 아래에 소개한다. 영문 작성의 경우 미국심리학회에서 출판한 최근 지침에 따른다.

1. 기본 사항

제목 및 초록은 1단 편집, 본문은 2단 편집 (단 간격 5.0mm)
단, 심사용 논문에서의 본문은 1단 편집도 무방하나, 게재 확정 후에는 반드시 2단으로 제출

편집용지: A4
용지 여백: 위쪽 37mm, 아래쪽 38mm
왼쪽 35mm, 오른쪽 35mm
머리말 13mm, 꼬리말 12mm
용지 방향: 좁게
문단모양: 문단 시작은 두 칸(한 글자)만큼 띄고 시작.
줄간격 160%
마침표 다음: 한 칸 띄도록 (두 칸이 아님)
본문, 참고문헌: 휴먼명조, 10호, 보통모양
국문초록, 영문초록: 휴먼명조, 9호, 보통모양
쪽수 표시

2. 세부 형식

제목 *휴먼명조, 16호, 진하게, 가운데
[‘**제목**’ 다음에 두 줄 띄우십시오]
국문초록 시작 *휴먼명조, 9호, 보통모양, 양쪽 혼합, 첫 칸을 띄지 않고 시작
문단모양: 왼쪽 3, 오른쪽 3
[‘국문초록’ 다음에 한 줄 띄우십시오]
주요어 *맑은고딕, 9호, 보통모양
[‘주요어’ 다음에 두 줄 띄우십시오]
본문 시작 *휴먼명조, 10호, 보통모양, 양쪽 혼합, 문단 첫줄은 두 칸 띄고 시작
여기서부터 2단 시작 (좌우 양단으로 편집함. 단 간격은 5mm)
본문소제목 *맑은고딕, 10호, 진하게, 양쪽 혼합
[‘**본문소제목**’이 끝나면 한 줄 띄우십시오]

방 법 *휴먼명조, 11호, 진하게, 가운데
[**'방 법'** 다음에 한 줄 띄우십시오]
연구대상, 측정도구, 절차 등 *맑은고딕, 10호, 진하게, 양쪽 혼합, 좌측 첫째 칸에서 시작
[**'연구대상, 측정도구, 절차'** 다음에 한 줄 띄우십시오]
연구대상, 측정도구, 절차의 내용 *휴먼명조, 10호, 보통모양, 양쪽 혼합, 첫 칸을 띄우고 시작
[**'연구대상, 측정도구, 절차의 내용'** 다음에 한 줄 띄우십시오]
결 과 *휴먼명조, 11호, 진하게, 가운데
[**'결 과'** 다음에 한 줄 띄우십시오]
결과의 내용 *휴먼명조, 10호, 보통모양, 양쪽 혼합, 첫 칸을 띄우고 시작
표 1. 표 제목 *맑은고딕, 9호, 보통모양, 표 제목은 표의 위쪽 좌측에,
제목이 길어서 두 줄 이상을 차지하는 경우에는 들여쓰기나 내어쓰기를 하지 않고 그대로 표기
그림 1. 그림 제목 *맑은고딕, 9호, 보통모양, 그림 제목은 그림 아래쪽 좌측에
논 의 *휴먼명조, 11호, 진하게, 가운데
[**'논 의'** 다음에 한 줄 띄우십시오]
참고문헌 *휴먼명조, 11호, 진하게, 가운데
[**'참고문헌'** 다음에 한 줄 띄우십시오]
참고문헌의 내용 *휴먼명조, 10호, 보통모양, 양쪽 혼합, 문단 첫 줄부터
여백; 왼쪽 여백 0, 오른쪽 여백 0
첫째줄; 내어쓰기 4
정렬; 양쪽 혼합
[**영문초록**]
영문제목 *휴먼명조, 16호, 진하게, 가운데, 페이지를 바꾸어서 시작
[**'영문제목'** 다음에 한 줄 띄우십시요]
영문초록시작 *휴먼명조, 9호, 보통모양, 양쪽 혼합
문단모양: 왼쪽 3, 오른쪽 3
['영문초록시작' 다음에 한 줄 띄우십시요]
Keywords: *휴먼명조, 9호, 이탤릭체, 양쪽 혼합, 첫 칸을 띄지 않고 시작
문단모양: 왼쪽 3, 오른쪽 3
[부록이 있을 경우 페이지를 바꾸십시오]
부 록 *휴먼명조, 11호, 진하게, 가운데
[**'부 록'** 다음에 한 줄 띄우십시오]
부록의 제목 *휴먼명조, 10호, 진하게, 가운데
(부록이 여러 개인 경우 부록마다 일련번호를 붙임)
부록의 내용 *휴먼명조, 9호, 보통모양, 양쪽 혼합

3. ANOVA(Analysis of Variance) 결과에 대한 제시

평이한 다원설계(factorial design)까지는 본문에 풀어쓰고 유의한 경우 유의하지 아니한 경우 모두 *F,*

df, p, MSE, 및 효과크기(η^2, ω^2, *d*, *f* 등)를 제시한다. 그러나 설계가 복잡해질수록(예: 집단내/집단간, 위계적 설계 등) 분석의 전문성을 살리는 차원에서 ANOVA표를 제시한다. 이 때 *MSE*를 제외한 *SS*와 *MS*는 생략하되 효과크기는 반드시 제시한다. ANOVA표의 예시는 아래와 같다.

(ANOVA표의 예시)

변산원	*df*	*F*	η^2	*p*
집단간				
인지(A)	2	.80	.05	.52
감정(B)	1	5.57*	.14	.03
AxB	2	1.64	.18	.20
집단내 오차(S/AB)	30	(20.05)		
집단내				
시점(C)	4	1.52	.05	.20
CxA	6	2.52*	.22	.03
CxB	3	3.98**	.26	.01
CxAxB	6	0.30	.02	.70
집단내 오차(CxS/AB)	120	(1.40)		

주. 괄호안의 수치는 오차제곱평균(*MSE*)을 나타냄.

*$p < .05$, **$p < .01$

4. 편집다지인 적용 후 검토 시 주의사항

저자의 수정사항을 파란색 또는 붉은색 글씨로 표시한다. 단, 파일의 환경이나 서체 등은 그대로 두고 내용 수정만 한다.

5. 저자의 이름과 소속

투고하는 원고에 저자의 인적사항이 포함되지 않도록 주의한다. 투고 시 저자 정보, 사사표기 및 연구지원 정보, 학위논문의 출판에 대한 알림은 저작권 이양 동의서 양식 투고 신청서에 기록하며 투고하는 원고에서 생략한다. 게재 확정 후 편집단계에서 저자 이름과 소속 정보를 원고에 기록한다.

한국심리학회 임원진

운영위원

회 장	최훈석 (성균관대학교 심리학과)
부 회 장	최기홍 (고려대학교 심리학부)
부 회 장	한영석 (호서대학교 산업심리학과)
총무 이사	용정순 (성균관대학교 심리학과)
재무 이사	최혜만 (가천대학교 심리학과)
홍보 이사	김민정 (아주대학교 교육대학원)
대외 이사 1(국내)	서동기 (한림대학교 심리학과)
대외 이사 2(국외)	조이수 (성균관대학교 심리학과)
정보 이사	박준성 (중앙대학교 심리서비스대학원)
학외 이사	윤세리 (법무법인 율촌)

상임위원장

편집위원회	나진경 (서강대학교 심리학과)
윤리위원회	조영일 (동국대학교 경찰행정학부)
학술위원회	한영석 (호서대학교 산업심리학과)
심리검사심의위원회	박준호 (경상국립대학교 심리학과)
학회발전기획위원회	서경현 (삼육대학교 상담심리학과)
자격제도위원회	최윤경 (계명대학교 심리학과)
공공정책위원회	윤상연 (경상국립대학교 심리학과)
심리학회보편집위원회	곽세열 (부산대학교 심리학과)
재난심리위원회	최해연 (충북대학교 심리학과)
심리사법위원회	최기홍 (고려대학교 심리학부)
학문후속세대교류위원회	김현식 (서강대학교 심리학과)
홍보위원회	김민정 (아주대학교 교육대학원)
국제교류위원회	조이수 (성균관대학교 심리학과)

임시위원장

자살예방위원회	고선규 (임상심리전문가 그룹 마인드웍스)
심리지원정책위원회	정경미 (연세대학교 심리학과)
청년정책위원회	김향숙 (서울대학교 심리학과)
심리학대중화위원회	박준성 (중앙대학교 심리서비스대학원)
심리학R&D지원위원회	최준식 (고려대학교 심리학부)

당연직이사

전임학회장	최진영 (서울대학교 심리학과)

감사

운영감사	정우현 (충북대학교 심리학과)
재무감사	원성두 (대구가톨릭대학교 심리학과)

분과학회장

제 1 분과 임상심리학회	배대석 (영남대학교의료원 정신건강의학과)
제 2 분과 상담심리학회	박성현 (서울불교대학원대학교)
제 3 분과 산업및조직심리학회	한영석 (호서대학교 심리학과)
제 4 분과 사회및성격심리학회	허태균 (고려대학교 심리학과)
제 5 분과 발달심리학회	송현주 (연세대학교 심리학과)
제 6 분과 인지및생물심리학회	김채연 (고려대학교 심리학과)
제 7 분과 문화및사회문제심리학회	서경현 (삼육대학교 심리학과)
제 8 분과 건강심리학회	조성근 (충남대학교 심리학과)
제 9 분과 여성심리학회	한영주 (벤쿠버기독교세계관대학교)
제 10 분과 소비자·광고심리학회	강정석 (전북대학교 심리학과)
제 11 분과 학교심리학회	남숙경 (국민대학교 상담심리학과)
제 12 분과 법심리학회	최이문 (경찰대학교 행정학과)
제 13 분과 중독심리학회	서보경 (을지대학교 중독재활복지학과)
제 14 분과 코칭심리학회	정은경 (강원대학교 심리학과)
제 15 분과 심리측정평가학회	김수영 (이화여자대학교 심리학과)
제 16 분과 디지털심리학회	신민섭 (서울대병원 소아청소년정신과)