한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 1, 1-29
http://dx.doi.org/10.22257/kjp.2026.3.45.1.1

국내 정신건강 개입 프로그램 효과성 평가 연구의 방법론에 대한 체계적 문헌 고찰 및 제언*

신 소 연 전 은 기 김 한 조†
연세대학교 연세대학교 연세대학교

정신건강 증진 및 예방 개입의 중요성이 증가함에 따라 개입 효과를 검증한 연구 방법론에 대한 체계적인 검토가 연구자와 실무자 모두에게 중요한 과제이다. 본 연구는 국내 정신건강 개입 프로그램의 효과성 연구들을 대상으로 해당 연구의 방법론적 특성을 체계적으로 분석한 실태조사이다. 최근 10여 년간 한국심리학회 및 산하 학회지에 게재된 총 177개의 논문을 검토하여 연구 설계 및 통계 분석 방법 등 연구의 방법론 위주로 비교 분석했다. 조사 결과, 두 가지 주요 사실을 발견했다. 첫째, 90%가 넘는 연구에서 집단별 표본크기가 30을 초과하지 않았다. 둘째, 비동등 집단 설계를 사용한 연구가 거의 절반(48.5%)을 차지했다. 이와 같은 결과를 바탕으로, 프로그램 평가 시 적용할 수 있는 다양한 방법론적 제언과 가이드라인을 제시했다. 본 연구는 프로그램 효과 분석에 필요한 방법론에 대한 기초 자료를 제공하는 데 의의가 있다.

주요어 : 예방 및 진흥, 개입 프로그램, 프로그램 효과성 평가, 소규모 표본, 비동등 집단 설계

* 이 논문은 2023~2024년도 연세대학교 연구비의 지원을 받아 수행되었습니다(2023-22-0120; 2024-22-0126). 연구 자료 수집 및 분석 등에 도움을 주신 연세대학교 진흥 및 예방 프로그램 평가 연구실의 송현이, 김영제, 윤정원, 김류위, 김채원 선생님께 감사드립니다.

† 교신저자: 김한조, 연세대학교 심리학과 교수, E-mail: hanjoekim@yonsei.ac.kr

서 론

전 세계적으로 정신건강 문제가 중요한 사회적 문제로 떠오르고 있다. 세계보건기구(WHO)에 따르면 2019년 기준 세계 인구 8분의 1이 해당되는 약 9억 7천만명이 정신질환을 앓고 있었으며, 2020년 코로나 팬데믹 이후 불안 및 우울 장애를 겪는 사람의 수가 급증했다(World Health Organization, 2022). 국내 상황도 예외는 아니다. 보건복지부 국립정신건강센터가 발표한 2024년 국민 정신건강 지식 및 태도 조사는 정신건강 문제를 경험한 인구 비율이 꾸준히 증가하고 있음을 보여준다. 특히, 2024년 1년간 스트레스, 우울감, 기타중독의 경험률이 22년도 대비 모두 10% 이상 증가했다(국립정신건강센터, 2024). 정신질환 유병률은 높아지는 반면, 정신건강 서비스에 대한 접근성은 여전히 제한적이다. 저소득 및 중간 소득 국가(LMICs)의 경우 정신건강 서비스 부족으로 인해 정신질환으로 어려움을 겪는 사람의 75% 이상이 적절한 치료를 받지 못하고 있다(Singh et al., 2022). 한국의 경우에도 2021년 정신건강 실태조사에서 평생 정신장애 진단을 받은 사람 중 단 12.1%만이 정신건강 서비스를 이용한 것으로 나타났다(보건복지부, 2022). WHO는 정신질환이 단순히 개인의 문제로 국한되지 않고 국가 경제와 사회 전반에 미치는 영향을 고려할 때, 공공 보건 전략으로서 정신건강 예방과 증진이 필수적임을 강조한다(World Health Organization, 2002). 조기 개입을 통해 질병이 심각해지는 것을 늦추거나 예방할 수 있다면 장기적으로 정신질환의 유병률 감소와 사회경제적 비용 절감, 나아가 개인의 삶의 질 향상으로 이어질 수 있다.

정신건강 예방(Mental Health Prevention)은 정신질환의 발생과 위험요인을 사전에 차단하거나 완화하기 위한 체계적 개입을 의미한다(World Health Organization, 2022). 이는 발병 전 위험요인 식별 및 완화 뿐만 아니라, 질환이 발현된 후에도 증상과 지속기간을 최소화하여 개인의 회복과 사회적 활동으로의 복귀를 목표로 한다.

정신건강 증진(Mental Health Promotion)은 개인의 정서적 안정, 인지적 유연성과 사회적 관계 강화 등을 통해 삶의 질을 향상시키는 접근이다(World Health Organization, 2022). 정신건강 증진 개입의 목표는 개인의 잠재력과 대처 능력 강화 뿐만 아니라 정신건강의 중요성에 대한 인식 제고까지 포함한다. 예방이 위험요인의 조기 발견과 개입에 초점을 두는 반면, 증진은 긍정적 자원과 역량을 개발하고 촉진한다는 점에서 차이가 있다. 그러나 두 접근법 모두 위험요인 감소와 건강한 정신 상태 유지라는 공통의 목표를 가진다.

세계보건기구(WHO)는 정신건강 예방과 증진은 상호보완적이며 통합으로 시행될 때 시너지 효과를 창출할 수 있음을 강조한다. 이러한 통합적 접근의 실제 사례로, 허은혜 등(2018)은 청소년 대상 학교 기반 정신건강 프로그램에서 사회불안장애 예방을 위해 고위험군 학생들에게 정신건강 선별검사와 인지행동치료를 실시하였다. 이와 함께 전교생을 대상으로 정신건강의 중요성을 강조하는 캠페인과 생명존중 교육을 제공하여 정신건강 증진 전략도 병행하였다. 김재준 등(2017)은 소방공무원을 대상으로 한 정신건강 증진 프로그램에서 PTSD, 우울, 불안 고위험군에게 집중 개별 상담을 진행하고, 전체 소방공무원에게는 정신건강 교육을 실시했다. 이처럼 예방 및 증

진 개입을 통합적으로 실행함으로써 고위험군뿐만 아니라 일반 집단에도 긍정적인 영향을 미칠 수 있으며, 적은 비용으로 다수에게 개입을 제공하는 효율성을 가진다.

세계보건기구는 2012년부터 각국의 정신건강의 예방 및 증진 개입에 대한 체계적인 가이드라인을 만들고 있으며, 회원국들이 이를 통해 인식을 높이고 효과적인 개입 전략을 개발하도록 권장하고 있다(World Health Organization, 2002). 미국 질병통제예방센터(Center for Disease Control and Prevention; 이하 CDC)에서는 1986년부터 전국에 예방연구센터(Prevention Research Center)들을 설립 및 지원하고 있고(CDC, 2025), Society for Prevention Research(SPR; https://preventionresearch.org/)와 같은 국제 학회에서도 매년 5월 정신건강 예방과 증진을 주제로 학술대회를 개최하고 있다.

개입 프로그램의 궁극적인 목적은 한 번 실행 후 종료되는 것이 아니라 프로그램의 지속적인 효과이다. 그리고 정신건강 개입은 기대되는 효과가 정서 변화, 삶의 만족도, 회복탄력성 등과 같이 직접 관찰하기 어려운 잠재변수로 나타나기 때문에, 프로그램 활동과 기대 성과 간의 인과적 구조를 명확하게 설정하는 것이 중요하다.

이 때 프로그램 효과를 정확히 검증하기 위해서는 연구 설계와 통계적 분석 방법의 적절한 선택이 핵심이 된다. 처치 효과(treatment effect)를 추정하기 위해서는 내적 타당성이 확보된 연구 설계와 적절한 분석 방법이 필요하다. Royse 등(2015)에 따르면, 실험 설계는 크게 사전실험설계(Pre-experimental research designs), 준실험설계(Quasi-experimental designs), 실험설계(Experimental designs), 단일 사례 연구 설계(Single System Research Design)으로 구분된다. 나아가 처치 집단 외 통제(비교) 집단의 유무, 무선 할당 여부 등에 따라 단일군 사전사후 설계(One-group Pretest-Posttest Design), 비동등 집단 설계(Nonequivalent Control Group Design), 무작위 대조군 설계(Randomized Controlled Trials; 이하 RCT) 등으로 세분화된다(Royse et al., 2015). 가장 이상적인 설계는 RCT이지만, 실제 현장에서는 무작위 배정이 어려운 경우가 많다. 이로 인해 준실험설계, 단일군 사전-사후 설계 등 다양한 연구 설계가 활용되고 있다. 분석 단계에서는 RCT를 사용하지 못 하는 경우, 보완할 통계적 보정 방법이 함께 사용된다. 즉, 연구 설계는 처치 효과의 내적 타당도를 확보하고, 통계 분석은 해당 연구 설계에서 효과 추정의 불확실성을 줄여 정밀도를 높인다.

다양한 개입 프로그램이 개발되고 여러 분야에 적용됨에 따라 체계적인 평가의 중요성도 커지고 있다(Royse, Thyer, & Padgett, 2015). 개입 프로그램들이 학교, 병원, 지역사회 등 다양한 맥락에서 활발히 시행되고 있고, 정신건강 분야에서도 프로그램 효과성을 검증하기 위한 실증 연구들이 축적되고 있다. 예를 들어, 서울의료원의 한 연구는 전국 10개 기초정신건강센터에서의 중증정신질환자 대상 집중사례관리 서비스를 제공한 후 그 효과를 평가하여 재발률 및 재입원율의 감소라는 결과를 확인하였다(국립정신건강센터, 서울의료원, 2024). 이는 정부의 정신건강 혁신 방안 수립과 현장 적용을 위한 근거 자료로 활용되었다. 학교 기반 정신건강증진 프로그램 연구에서는 서울시 36개교 3,870명의 학생을 대상으로 실시한 사회성 증진 프로그램의 효과성을 입증했다(심서연 외, 2015). 이러한 평가는 현장에서 개입 프로그램을 개선하고 확산시키기 위

한 중요한 과정이다.

예방 및 진흥 개입 프로그램의 체계적인 평가를 위해서는 위에서 언급했듯이, 연구 설계 및 분석과 관련된 방법론에 대한 체계적인 검토가 선행되어야 한다. 이를 통해 개입 프로그램 효과성을 정확히 측정하고 현장에 개입 프로그램을 적용하고 지속하기 위한 방안을 고찰할 수 있다. 기존 연구들은 대부분 특정 대상이나 특정 문제 영역, 주제에 국한된 프로그램 체계적 검토 및 메타분석을 수행한 경우가 많았다. 예를 들어, Neil & Christensen (2009)은 아동 및 청소년을 대상 불안 예방 개입 프로그램 20개 독립적인 연구의 방법론을 검토하고 각 프로그램의 효과크기를 비교하였다. Choi & Hector(2012)는 노인 낙상 예방 개입 프로그램 17개의 무작위 대조군 실험을 낙상 감소율과 낙상 발생률을 같은 종속변수로 설정하여 효과를 비교하였다. Su & Reeve(2010)는 자율성 지지를 위한 19개 개입 프로그램의 구조와 실행 시간 등 조건에 따른 효과크기를 비교하였다. 국내 변증법적 행동치료(DBT) 기반 프로그램 8개의 특성과 효과크기를 비교하였다(김동일 외, 2024). 이처럼 기존 연구들은 특정 대상이나 프로그램 유형 내에서 효과성을 비교하고 있다.

이처럼 기존 연구들은 특정 대상이나 프로그램 유형 내에서 효과성을 비교하는 데 집중해 왔으며, 국내에서 수행된 다양한 정신건강 개입 연구들의 방법론 실태를 포괄적으로 진단한 연구는 부족한 실정이다. 이러한 방법론에 대한 중요성은 해외에서도 꾸준히 제기되어 왔으며, 이를 해결하기 위한 제도적 노력이 이어져 왔다. 대표적으로 미국 교육과학기관(Institute of Educational Sciences)의 What Works Clearinghouse(WWC)는 개입 연구의 구체적인 방법론적 절차와 보고 표준(Procedures and Standards)을 수립하여 제시한다(IES, 2022). 그러나 국내의 경우, 이와 같은 가이드라인이 존재하는지, 그리고 실제 연구 현장에서 어느 정도 준수되고 있는지, 혹은 국내 연구 환경의 특수성을 고려할 때 어떠한 보완이 필요한지에 대한 체계적인 검토가 여전히 미흡하다.

이에 본 연구는 체계적 문헌고찰을 통해 국내에서 시행된 다양한 정신건강 프로그램들의 방법론 실태를 종합적으로 조사하고자 한다. 구체적으로 최근 10여 년간 국내 주요 학술지에 게재된 총 177편의 논문을 대상으로 연구 설계와 분석 방법을 중심으로 프로그램 평가의 현황을 고찰한다. 특히 국내 연구 현장의 특수성으로 인해 발생할 수 있는 실질적인 방법론적 이슈들이 실제 연구에서 어떻게 다뤄지고 있는지 체계적으로 파악하고자 한다.

본 연구의 결과는 기존의 일반적인 원칙 제시 수준에 머물렀던 방법론적 제언을 177편의 실증적 데이터를 바탕으로 구체화하고 업데이트한다는 점에서 학술적 기여를 가진다. 비록 본 연구가 WWC와 같은 방대한 수준의 표준을 모두 포괄하지는 못하더라도, 국내 연구 현장에서 나타나는 구체적인 한계와 양상을 반영하여 연구자들이 즉각 활용할 수 있는 평가 체크리스트와 기록용 요약표를 제안하고자 한다. 이를 통해 프로그램 평가 과정에서 나타날 수 있는 한계점을 보완하고, 한국의 연구 환경에 최적화된 방법론적 가이드라인을 새롭게 제시하는 데 본 연구의 목적이 있다.

방 법

논문 선정

국내 정신건강 증진 및 예방 프로그램 평가에 대한 실태조사를 실시하기 위하여 한국심리학회와 각 분과학회 학술지(한국심리학회지: 상담 및 상담치료, 일반, 코칭, 소비자 광고, 법, 학교, 임상심리 연구와 실제, 사회 및 성격, 인지 및 생물, 건강, 발달, 문화 및 사회문제, 산업 및 조직, 중독, 여성, Korean Journal Of Clinical Psychology)에 게재된 논문을 검색할 수 있는 Access ON 데이터베이스를 활용했다. 한국심리학회 및 산하 학회지를 본 연구의 분석 대상으로 선정한 이유는 국내 사회과학 분야에서 이미 엄격하고 과학적인 연구 기준을 유지하고 있기 때문이다. 본 연구는 이러한 높은 학문적 기반 위에서 향후 더 정교한 연구 설계와 분석 방법을 확장할 수 있는 방향을 제안하고자 한다.

논문은 참가자의 정신건강을 예방하거나 증진시키는 심리개입을 실시하고, 심리와 관련된 구성개념을 종속변수로 측정한 연구를 대상으로 했다. 또한 최근 12년 (2013년 1월 ~ 2024년 7월) 동안 발행되었으며 한국에서 실시된 프로그램 연구를 포함하였다. 자세한 검색 과정은 다음과 같다. 먼저 여러 검색어를 각각 포함하기 위해서 검색 연산자 'OR'을 사용하여 '프로그램', '처치','개입', '정신건강', '증진', '예방', '진흥'을 검색하였다. 이와 동시에 한국심리학회지 및 산하 학회지로 한정하여 상세검색했을 때 총 2,015개의 논문이 검색되었다. 이 중 발행연도를 최근 12년으로 (2013년 ~ 2024년 7월) 제한하였을 때 1,279개의 논문이 확인 됐다. 검색된 1,279개의 국내 학술논문 중 논문 제목과 초록을 검토하여 정신건강 증진 및 예방을 목적으로 한 프로그램 개발 및 효과성 검증 연구만을 선정했다. 한편, 심리적 혹은 행동적 기제를 파악하기 위한 실험연구는 본 연구의 목적과 다르므로 제외하였다. 최종적으로 177개의 국내 학술논문을 본 연구의 분석 대상으로 포함했다. 그림 1은 전체 논문 선정 과정을 도식화한 것이다.

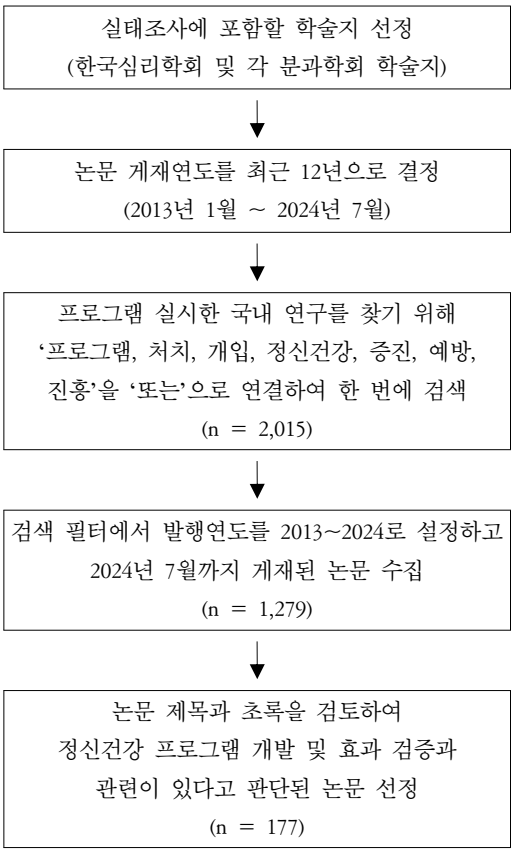

그림 1. 논문 선정 과정

선정한 논문 코딩 절차 및 변수

최종 177개의 논문을 코딩하기 위해 다음과 같은 절차를 거쳤다. 먼저 일관된 내용으로 논문을 분석하기 위해 변수를 정하고 코딩 매뉴얼을 만들었다. 코딩 메뉴얼 내용을 대학원

생과 학부생 연구조교들에게 교육하고 논문마다 한 명의 책임 검토자와 2명의 교차검증자를 배정하였다. 모든 검토자는 교차 검증한 내용을 매주 회의하고 진행사항을 공유하였다. 불일치가 있을 경우 토론을 통해 합의에 이르렀다.

각 논문은 심리개입 프로그램의 특성, 참가자 특성, 연구설계, 측정변수, 통계방법 등의 영역을 중심으로 정리하였다. 정리한 내용을 바탕으로 주요 변수들을 코딩하였다. 최종 코딩한 주요 변수는 (1) 집단별 표본 수, (2) 연구 설계, (3) 선택 편향 대처 방법, (4) 처치효과 추정 및 검정 방법, (5) 매개 또는 조절 분석 여부, (6) 신뢰구간 보고 여부, (7) 효과크기 보고 여부이다. 각 변수들에 대한 구체적인 설명은 다음과 같다.

집단별 표본 수

분석된 참가자 수가 집단별 30명 초과, 30명 이하일 때로 나누어 코딩하였다. 이는 소규모 표본을 사용한 연구를 구분하기 위함이다. 집단별 표본 수는 연구 내 각 설계 그룹의 표본 크기를 나타낸다. 연구 설계를 고려하여 처치 효과는 총 표본 크기가 아닌 해당 조건의 참가자 수로 분석하기 때문에 집단별 표본 수를 검토했다(Holmes, 2011). 충분히 큰 표본 크기에 대한 기준은 통계 모형과 모형의 복잡도에 따라 다르지만(McNeish, 2017) 통계 분석에서는 30명이 여러 분야에서 경험에 의한 법칙(Rules of Thumb)으로 사용된다. 일반적으로 중심극한정리에 따르면, 표본 크기가 충분히 클 경우(통상적으로 30 이상으로 간주되기도 함), 모집단의 분포와 상관없이 표본평균의 분포가 정규분포에 근사한다고 알려져 있다. 따라서 표본크기가 30명 이상이면 통계적 추론에 필요한 정규성 가정을 만족한다고 간주한다(Mascha & Vetter, 2018).

연구 설계

연구 설계는 크게 다섯 가지로 구분했다. RCT는 처치 집단과 통제 집단 (혹은 비교집단)을 무작위로 할당하여 집단의 선택적 차이를 제거하는 연구 설계이다. 비동등 집단 설계는 RCT와 마찬가지로 처치 집단과 통제 집단은 있으나 무작위로 집단을 할당하지 않은 연구설계이다. 단일군 사전-사후 설계는 처치 집단만을 대상으로 개입 전후를 비교하여 효과를 평가하는 설계이다. 단일 사례 설계는 특정 개인이나 소규모 집단에 대한 개입 효과를 종단 자료를 수집하여 평가하는 설계이다. 마지막으로 인터뷰나 후기와 같은 질적 자료를 수집하고 내용을 분석한 질적 연구로 구분했다.

선택 편향 대처 방법

비동등 집단 설계에서는 무선 할당을 하지 않았기 때문에 선택 편향(Selection Bias)으로 인한 내적 타당성 위협이 있다(Steiner et al., 2015). 선택 편향이란 집단 간 존재하는 체계적 오차이며 이는 개입 효과 추정치에 편향을 일으킬 수 있다(Wong & Steiner, 2018). 따라서 본 연구에서는 비동등 집단 설계를 사용한 연구들에서 내적 타당도를 확립하기 위해 어떤 방법을 사용했는지 확인하고 네 가지 방식으로 코딩하였다. (1) 집단 간 차이를 확인 및 통제하지 않은 연구, (2) 인구통계학적 정보 혹은 사전 점수로 집단 간 동질성 검정을 한 연구, (3) 처치효과 추정 모형에서 변수를 추가하여 통제하는 통계 분석방법(ANCOVA)을 사용한 연구, 또는 (4) 동질성 검정과 통계 분

석 방법 둘 다 사용한 연구로 구분하였다.

처치효과 추정 및 검정 방법

심리개입 프로그램 평가 연구에서 주로 어떤 분석 방법을 사용하여 프로그램의 효과성을 검증했는지 알고자 했다. 총 177개의 연구의 분석 방법은 반복측정 변량분석(Repeated Measures ANOVA), 대응표본 t-검정(Paired t-test), 혼합변량분석(Mixed ANOVA), 공분산분석(ANCOVA), 비모수 검정(Non-parametric Tests), 독립표본 t-검정(Independent Samples t-test), 질적 분석(Qualitative Analysis)으로 분류하였다. 선행 연구에 따르면, 표본 크기가 작은 경우에는 모수 통계보다 비모수 통계 방법을 선택하는 경향이 보고된 바 있다(신소연, 김한조, 전은기, 2024). 이러한 경향이 본 연구의 분석 대상에서도 일관되게 나타나는지 확인하기 위해 다음의 여섯 가지 범주로 분류하여 코딩했다. (1) 모수 통계를 활용한 평균 비교 분석, (2) 비모수 통계 분석, (3) 종속변수가 정규성 검증을 기각했을 경우 모수와 비모수 분석을 모두 실시, (4) 통계 분석 없이 단순 기술통계만을 활용한 수치 비교, (5) 변수 간 관계 분석(상관과 회귀 분석), (6) 질적평가만을 수행한 연구.

매개 또는 조절 분석 여부

매개변수는 개입 프로그램이 결과 변수에 영향을 미치는 인과적 메커니즘을 설명한다. 조절변수는 프로그램의 효과가 참가자들의 특성에 따라 다르게 나타나는지를 설명하는 변수로 프로그램이 어떤 참가자들에게 더 효과적인지 이해할 수 있다. 본 연구에서 매개 및 조절 분석을 실시했는지 여부를 나누어 코딩하였다.

효과크기 보고 여부

통계적 유의성과 함께 효과크기는 개입 프로그램 효과성을 해석할 때 효과의 실질적인 의미를 판단하는 데 사용할 수 있다. 본 연구에서 효과크기를 산출하고 기술했는지 여부를 코딩하였다.

신뢰구간 보고 여부

신뢰구간은 추정의 불확실성을 나타내는 지표로 사용할 수 있다. 본 연구에서는 개입 효과의 신뢰구간을 구하고 보고한 여부에 따라 코딩하였다.

결 과

개입 프로그램의 특성

표 1에는 전체 177개 논문에 대한 정보와 주요 변수들과 내용을 확인할 수 있다. 아래는 각 변수별 결과를 정리한다.

집단별 표본 수

집단별 표본 수 30명 이하로 표집한 연구는 총 177개 중 161개(90.96%)이었다. 집단별 표본수 30명 초과는 총 15개(8.47%)이었다. 이 외에 논문에 여러 개의 연구가 있고 두 경우를 모두 포함한 논문이 1개(0.56%) 있었다.

연구설계

심리개입 프로그램의 효과성을 분석한 177개의 논문에서 무작위 대조군 설계와 비동등 집단 설계가 가장 많이 사용되었다. 무작위 대조군 설계를 사용한 연구는 전체 177개 중

표 1. 각 변수별 결과표

카테고리	코드	내용	개수	백분율(%)
집단별 표본 수 (계 : 177)	1	집단별 30명 이하	161	90.96
	2	집단별 30명 초과	15	8.47
	3	둘 다 사용	1	0.56
연구설계 (계 ; 177)	1	무작위 대조군 설계	81	45.76
	2	비동등 집단 설계	86	48.59
	3	무선할당이 불확실한 설계	0	0.00
	4	단일군 사전-사후 설계	6	3.39
	5	단일 사례 설계	2	1.13
	6	질적 연구	1	0.56
	7	기타	1	0.56
비동등 집단 설계 중 선택 편향 대처 방법 (계 : 86)	0	확인 및 통제하지 않음	3	3.53
	1	동질성 검정으로 확인	75	87.06
	2	통계분석방법으로 통제	2	2.35
	3	1, 2 둘 다 실시	6	7.06
분석방법 (계 : 177)	1	단순 수치 비교	2	1.13
	2	모수통계 평균 비교	147	83.05
	3	비모수통계 방법	22	12.43
	4	종속변수별로 다른 분석 방법	5	2.82
	5	질적 평가	1	0.56
매개 또는 조절효과 분석 (계 : 177)	0	실시하지 않음	169	95.48
	1	조절효과 분석	2	1.13
	2	매개효과 분석	6	3.39
효과크기 보고 여부 (계 : 177)	1	보고함	64	36.16
	2	보고하지 않음	113	63.84
신뢰구간 보고 여부 (계 : 177)	1	보고함	3	1.69
	2	보고하지 않음	174	98.31

주. 데이터셋은 OSF(DOI: 10.17605/OSF.IO/H8DB6)에 공개되어 있음: https://osf.io/h8db6/

81개(45.76%)이고 비동등 집단 설계를 사용한 논문은 86개(48.59%)이었다. 비동등 집단 설계를 사용한 86개 중 1편의 연구는 양적 분석과 질적 분석을 둘 다 사용했다. 통제(비교)집단 없이 처치 집단 단일군 사전-사후 설계를 사용한 논문은 6개(3.39%)이었다. 그리고 소규모 표본을 반복 측정하여 개개인의 변화에 초점을 맞춘 단일 사례 설계를 사용한 논문은 2개(1.13%)로 조사되었다. 단일 사례 설계를 사용한 각 논문의 표본 수는 2명, 4명으로 소규모 연구에서 ABA 설계를 이용하여 변화를 비교했다. 그 외에 개별 인터뷰의 의미 구조를 도출하는 현상학적 연구 방법론을 사용한 질적 연구 1개(0.56%)가 있었다. 마지막으로 1개의 연구(0.56%)는 한 종속변수의 평균점수를 사용하여 군집 분석하였다. 군집 분석을 통해 세 집단에서 개입 프로그램이 미치는 효과를 검증하였다.

선택 편향 대처 방법

무작위 배정이 어려워 선택 편향 우려가 있는 경우 이를 통제하려는 방법이 대부분 연구에서 이루어졌다. 사전에 실험집단과 통제집단 간 동질성 검증(나이와 성별 등 인구통계학적 변인이나 사전 측정치 비교)을 실시하여 두 집단 간 차이가 없음을 확인하고 추가적인 통계적 조정이 이루어지지 않은 논문은 75개(87.21%)가 해당됐다. 두 개 논문(2.33%)은 사전 측정치를 공변인으로 설정하여 ANCOVA를 통해 프로그램의 효과를 분석했다. 여섯 개 논문(6.98%)은 동질성 검정 후 유의한 차이를 보인 변인을 공변인으로 투입하거나, 차이가 없더라도 사전 측정치를 공변인으로 처리하는 동질성 검증과 통계적 통제를 병행하였다. 한편, 비동등 집단 설계에서 선택적 차이에 대한 어떠한 통제나 확인도 이루어지지 않은 논문은 86개 중 3개(3.49%)가 해당됐다.

분석 방법

총 177개의 검토한 논문 중 147개 논문(83.05%)는 모수 통계 방법(RM-ANOVA, Mixed ANOVA, ANCOVA, Paired t-test, Independent t-test)을 사용하여 처치 집단과 통제(비교)집단의 평균 차이 검정을 실시했다.

총 22개 논문(12.43%)는 비모수통계 방법(Wilcoxon Signed-Ranks Test, Mann-Whitney U Test, Kruskal-Wallis Test, Friedman Test, non-parametric survival analysis)을 사용하였다. 비모수 통계를 사용한 22개 논문 중 1개는 생존분석을 사용하였다. 생존 분석한 연구는 일정 기간 동안 재범이 발생하는 비율을 비교하기 위해 Kaplan-Meier 분석법과 Generalized Wilcoxon Test를 사용하여 처치 집단과 비교집단의 재범률을 비교 분석했다. 이 연구 외에 비모수 통계를 사용한 22개의 연구 중 21개는 집단별 표본 수가 30명 이하인 연구였다. 한편, 5개 논문(2.82%)은 종속변수 별로 다른 분석방법을 사용하였다. 예를 들면, 정규성 및 등분산성 가정을 만족한 변인은 모수 통계 방법을, 만족하지 않은 변인은 비모수 통계방법을 사용했다. 이외에 2개 논문(1.13%)은 통계적 검정을 하지 않고 단순 평균 수치의 변화를 비교했다. 이 2개 연구는 모두 단일 사례 연구 설계(Single System Research Design)을 사용했다. 그 외에는 질적 분석을 사용한 논문 1개(0.56%)가 있었다.

매개 또는 조절효과 분석

매개 효과를 검증한 논문은 177개 중 6개(3.39%)였다. 이는 대부분의 연구가 개입 효과

의 직접 효과 검증에 집중하고 있다는 것을 시사한다. 조절변수를 검증한 논문은 2개(1.13%)였다. 이는 프로그램 효과성 분석에 있어서 조절효과 분석이 국내 연구에서 상대적으로 미비한 상황임을 보여준다.

효과크기

분석 대상으로 선정된 177개의 논문에서 처치 효과에 대한 효과크기를 보고한 연구는 64개(36.16%), 보고하지 않은 연구는 113개(63.84%)였다. 효과크기를 보고한 연구들 중에 평균 차이를 표준화한 지표인 $Cohen's\ d$ 보고한 연구는 25개(39.06%) 해당됐다. 분산분석에서 종속변수의 변동성 중 특정 변수의 효과크기를 계산하여 partial eta square를 보고한 연구는 29개(45.31%) 해당됐다. 그리고 둘 다 보고한 연구는 10개(15.63%)였다.

신뢰구간

처치 효과에 대한 신뢰구간을 보고한 연구는 3개(1.69%), 보고하지 않은 연구는 174개(98.31%)로 나타났다. 신뢰구간을 보고한 3개의 연구는 매개효과 유의성 검정을 위해 부트스트랩 신뢰구간을 해석했다.

다각적 분석 결과

다음으로, 개입 특성들 간 양상을 파악하기 위해 교차분석을 실시했다. 연구 설계, 표본크기, 프로그램 대상군, 분석 방법의 변수을 조합하고 어떤 특징을 보이는지 살펴본 결과를 항목별로 제시했다. 이러한 결과를 파악하여 더 정교한 방법론적 제언을 하고자 한다.

연구설계 유형별 표본 크기

표 2를 참고하면, 모든 설계 유형에서 집단별 30명 이하의 소규모 표본을 사용한 빈도가 90.96%로 컸다. 집단별 30명 넘는 표본을 활용한 경우는 무작위 대조군 연구에서 6편(3.4%), 비동등 집단 설계 연구에서 10편(5.6%)이 나타났다. 이 외에 모든 설계는 집단별 30명 이하의 소규모 표본을 사용했다. 이는 모든 설계 유형의 연구에서 큰 표본을 활용하기 어려운 현실을 보여준다. 이러한 양상은 연구 참여자 모집의 어려움과 현장의 제약에서 비롯된 것으로 보인다. 논문에서 직접적으로 드러나지는 않았지만, 임상군이나 위기군 대상으로 한 연구의 경우 개입 효과의 높이기 위해 대규모 모집이 제한되거나 개입의 대상자가 적어 많은 표본의 표집이 어려웠을 가능성

표 2. 연구설계 유형별 집단별 표본 크기 분할표

연구 설계	집단별 표본 크기 30명 이하	집단별 표본 크기 30명 초과	합계
무작위 대조군 설계	75	6	81
비동등 집단 설계	76	10	86
그 외 설계	10	0	10
합계	161	16	177

주. 한 논문에서 두 표본 규모를 모두 사용(연구1, 연구2)한 사례는 30명 초과인 경우 포함

이 있다. 일반 집단 대상이라 해도 시간과 비용, 협조기관 확보의 어려움 등으로 대규모 무작위 배정을 현실적으로 수행하기 어려울 수 있다.

연구설계 유형별 프로그램 대상군

대상군 분류 기준에 따라 프로그램 대상군을 임상군, 위험군, 비임상 집단으로 구분했다. 임상군은 의학적, 정신의학적 진단을 받은 집단(예를 들면, 조현병, 발달장애, ADHD, 암, 중독, 당뇨병 등)을 의미한다. 위험군은 심리적 증상이 선별척도 기준 이상으로 나타나는 집단(예를 들면, 우울, 불안, 신체화, 공격성, 사회공포, 문제음주 등)을 말한다. 비임상 집단은 이러한 임상군과 위험군 기준에 해당하지 않는 대상(예를 들면, 워킹맘, 대학생, 노인, 군인, 직장인, 상담사, 보육교사 등)으로 정의했다.

표 3에 따르면, 무작위 대조군 설계 연구에서는 위험군 대상 연구가 43편(24.3%)으로 가장 많았고, 비임상 28편(15.8%), 임상군 10편(5.6%)으로 나타났다. 비동등 집단 설계에서는 비임상 대상 연구가 43편(24.3%)으로 가장 많았고 위험군 29편(16.4%), 임상군 14편(7.9%)으로 나타났다. 이를 종합하면, 무작위 대조군 설계에서는 위험군의 비중(24.3%)이 특히 높았고, 비동등 집단 설계에서는 비임상 비중(24.3%)이 특히 높았다.

이는 각 대상군이 가진 특성과 현장 상황이 연구 설계 선택에 제약을 줬을 가능성을 시사한다. 예를 들어, 비임상 대상 연구는 학교나 지역사회 기반으로 실시한 예방 및 증진 프로그램이 많고 현장에 비교적 간편하게 적용할 수 있는 비동등 집단 설계가 많이 사용되었다. 반면, 위험군은 증상 수준이 잠재적인 경우가 많아 윤리적 부담이 덜하며 동시에 예방 및 조기 개입 프로그램의 효과를 과학적으로 입증하려는 필요성이 커 무작위 대조군 설계가 더 많이 사용된 것으로 볼 수 있다. 이와 달리 임상군 연구는 윤리적 제약이 크고 치료 개입의 긴급성이 높아 무선 배정이 어려운 경우가 많다. 이런 경우, 아무런 처치를 하지 않는 통제집단 대신 대기자 통제 집단(waitlist control)을 두어 추후 바로 개입을 제공하는 조건으로 윤리적 문제를 보완하기도 한다. 실제로 총 81편의 무작위 대조군 설계 연구 중 대기자 통제 집단을 사용한 비율은 대상군별로 차이를 보였다. 위험군 대상 연구는 43개 중 21개(48.8%), 임상군 대상 연구는 10개 중 4개(40%)가 대기자 통제 집단을 사용한 반면, 일반인 대상 연구는 28개 중 7개(25%)가 사용했다. 이는 치료나 개입의 필요성이 높은 임상군과 위험군 연구에서 연구 참여자의 개입을 보장하기 위한 윤리적 장치로서 대기자 통제

표 3. 연구설계 유형별 프로그램 대상군 분할표

연구 설계	임상군 대상	위험군 대상	일반인 대상	합계
무작위 대조군 설계	10	43	28	81
비동등 집단 설계	14	29	43	86
그 외 설계	4	3	3	10
합계	28	75	74	177

집단을 활용했음을 시사한다.

표본크기 별 분석 방법

표 4를 참고하면, 집단별 30명 이하인 연구와 이를 초과한 연구 모두에서 모수 통계 분석 사용이 절대적으로 많았다. 한편, 비모수 통계 분석을 사용한 22편의 연구 중 21편(95.5%)이 집단별 30명 이하의 소규모 표본 대상으로 했다. 모수와 비모수 통계 모두 사용한 5편의 연구에서도 모두 소규모 표본을 대상 이러한 결과는 표본 크기가 작을 경우 비모수 검정을 채택하는 경향이 있음을 시사한다. 이러한 경향에 대한 적절한 방법론적 제언은 이후 방법론적 제언에서 제시될 예정이다.

연구설계별 분석 방법

표 5를 참고하면, 무작위 대조군 설계 와 비동등 집단 설계에서 모수 통계 사용 비율이 절대적으로 높았다. 무작위 대조군 설계 연구의 경우 69편(85.2%)의 연구가 모수 통계 분석을 사용했고, 비모수 통계 8편(9.9%), 종속변수가 여러 개일 때 모수와 비모수 통계를 함께 사용한 연구는 4편(4.9%)였다. 비동등 집단 설계의 경우 74편(86%)의 연구가 모수 통계 분석을 사용했고, 비모수 통계 11편(12.8%), 함께 사용한 연구는 1편(1.2%)였다. 단일군 사전-사후 설계에서는 모수와 비모수 통계 각각 3편씩 사용된 반면, 단일군 사례 설계는 모두 단순 수치 비교만 수행했다. 단일 사례 설계는 소수의 대상에 초점을 맞추기에 시간에 따른 변화 추이 그래프나 수치 비교를 사

표 4. 표본 크기별 분석 방법 분할표

연구 설계	모수 통계	비모수 통계	모수와 비모수 둘다 사용	단순 수치 비교	질적 평가만 사용	합계
집단별 30명 이하	132	21	5	2	1	161
집단별 30명 초과	15	1	0	0	0	16
합계	147	22	5	2	1	177

주. 한 논문에서 두 표본 규모를 모두 사용(연구1, 연구2)한 사례는 30명 초과인 경우에 포함

표 5. 연구 설계별 분석 방법 분할표

연구 설계	모수 통계	비모수 통계	모수와 비모수 둘다 사용	단순 수치 비교	질적 평가만 사용	합계
무작위 대조군 설계	69	8	4	0	0	81
비동등 집단 설계	74	11	1	0	0	86
단일군 사전-사후 설계	3	3	0	0	0	6
단일군 사례 설계	0	0	0	2	0	2
그 외 설계	1	0	0	0	1	2
합계	147	22	5	2	1	177

용했다. 종합하면, 집단 차이 검증하는 연구에서 모수 통계가 주된 분석 도구로 사용되고 있다.

개입 프로그램 평가에 대한 방법론적 제언

증진 및 예방 프로그램을 실시하고 효과성을 검증한 학술지 논문 총 177개를 체계적으로 검토했다. 국내 연구는 학교, 지역사회, 임상 등 다양한 현장에서 개입을 적용하고 그에 대한 효과를 검증했다.

이러한 연구들은 현실을 잘 반영하기에 유용하고 많은 함의점을 갖고 있다. 그리고 현장의 제약 속에서도 동질성 검정, 공변인을 통제한 ANCOVA 와 같이 다양한 분석 방법을 활용하여 프로그램의 효과를 검정하려는 노력을 볼 수 있었다. 이러한 국내 연구의 축적은 정신건강 개입 연구의 기반을 다지는 데 중요한 역할을 할 것이다.

한국심리학회 및 산하 학회지에 실린 논문들은 이미 높은 수준의 연구 방법론을 사용하고 있지만, 본 연구는 177편의조사 결과에서 관찰된 방법론적 선택과 분석 방법을 근거로, 그 기반 위에서 더 정교한 연구 설계와 분석 방법을 제언하여 발전하고자 했다. 특히, 국내 연구 환경을 고려하면서도 과학적 정밀성을 강화할 수 있는 실용적 개선 방안을 포함했다. 이를 바탕으로 프로그램 효과성 검증 시 고려할 주요 항목을 다음과 같이 도출했다: 내적 타당도 확보, 소규모 표본 연구에서의 통계적 방법론, 매개와 조절 효과 분석의 중요성, 개입 프로그램 평가의 체계화

내적 타당도 확보를 위한 제언

개입의 효과를 평가할 때 무작위 대조군 연구(Randomized Controlled Trial; RCT)는 가장 높은 내적 타당도를 확보할 수 있는 설계로 간주된다(Shadish, Cook, & Campbell, 2002). RCT는 참가자를 실험 조건에 무선 할당함으로써 실험집단 간 개인차를 제거한다. 이는 사전의 개인적 특성 차이가 처치 외 효과에 미치는 영향을 최소화하여 결과 편향을 일으키는 혼입 변수를 통제하는 것이다. 이러한 과정은 집단 간 결과를 비교할 때 순수한 처치 효과로 인과 추론을 가능하게 하여 내적 타당도가 가장 높은 연구 설계로 간주된다(Tein et al., 2018).

하지만 현장에서는 RCT를 적용하기 어려운 여러 윤리적 및 실질적인 제약과 방법론적 문제점이 있다(Sanson-Fisher, Bonevski, Green, & D'Este, 2007). 특정 개입이 필요한 사람에게 통제 조건을 무선 할당하는 것은 윤리적으로 문제가 될 수 있다. 또한, 대기-통제 집단 설계(waitlist-control group design)의 경우 연구 기간 동안에 한정적으로 참가자 일부에게만 처치를 할당하는 경우가 있다. 이때, 참가자들의 시간적인 제약 혹은 처치를 빨리 받고자 하는 동기 때문에 일부 참가자들에게 먼저 처치조건을 할당하게 된다. 따라서 현실적으로 참가자가 스스로 처치를 선택(self selection)하거나 교사, 운영 관리자, 치료사, 의사 등에 의해 조건이 할당(administrator selection)되는 경우가 많다(Shadish, Cook, & Campbell, 2002). 그리고 철저한 무선할당이 시행되었더라도 통제되지 않은 실제 현장에서 발생하는 요인들로 인해 무선 할당이 깨질 수 있다.

이러한 한계점들로 인해 개입 연구에서는

무선 할당 없이 처치 효과를 추정하는 준실험 설계(quasi-experimental design)가 대안적으로 널리 사용된다. Reichardt(2019)에 따르면, 준실험 설계의 핵심 특징은 무선 할당이 이루어지지 않는 것이며 비동등 집단 설계(nonequivalent group design), 회귀 불연속 설계(regression discontinuity design), 단일 사전-사후 검사 설계(pretest-posttest design), 중단 시계열 설계(Interrupted Time-Series Design, ITS), 단일 사례 설계(Single-Case Design, SCD) 등 다양한 연구 설계를 포함하는 개념이다.

본 논문에서 검토한 177개의 연구 설계 중 준실험설계로는 비동등 집단 설계, 단일군 사전-사후 설계, 단일군 사례 설계가 있었다. 단일군 실험 설계는 통제 집단이 없기에 실행 가능성이 높지만, 처치 효과에 대한 인과 추론의 타당성이 떨어진다(Spurlock, 2018). 비동등 집단 설계는 현장의 집단을 반영하여 연구를 수행할 수 있다는 장점이 있으나 선택 편향(selection bias)으로 인한 내적 타당성 위협이 있다. 따라서 개입 연구에서 선택 편향을 인지하고 통제하기 위한 연구 설계와 분석 방법을 강구해야 한다(Larzelere, Kuhn, & Johnson, 2004). 비동등 집단 설계에서 적절한 통계적 방법을 사용하여 한계점을 보완하고 개입 프로그램의 실제적 효과를 평가할 수 있다(Larzelere, Kuhn, & Johnson, 2004).

검토한 177개의 논문에서 가장 많은 비율(85개, 48.02%)로 비동등 집단 설계를 사용하고 있고 비동등 집단 설계를 사용한 85개의 연구 중 약 74개(87%)는 집단 간 인구통계학적 변수 혹은 사전 측정치를 기준으로 동질성 검정만을 실시했다. 사전 동질성 검정은 연구에서 처치가 시작되기 전에 집단 간 특성이 유사한지 확인하는 통계적 절차이다. 보통 각 집단의 사전 점수나 인구통계학적 정보로 χ^2 검정이나 독립표본 t-검정 등을 수행하여 집단 간 차이가 통계적으로 유의한지 확인한다. 검토한 74개 연구들은 집단 간 유의한 차이가 없음을 확인하면 이후 처치효과 분석을 진행했다. 하지만, 통계적으로 집단 간 유의한 차이가 없다는 결과가 잠재적인 모든 변수의 집단 간 차이가 통제되었다는 것을 의미하지 않는다. 즉, 인구통계학적 변수와 사전 측정치 외 처치 효과에 영향을 줄 수 있는 잠재적인 차이가 집단 간 여전히 존재할 수 있다. 따라서 본 고찰 결과가 보여주듯, 비동등 집단 설계를 사용할 때는 동질성 검정에서 더 나아가 선택 편향을 줄이는 설계 및 분석 전략을 병행할 필요가 있다.

공분산분석(ANCOVA)는 분산분석(ANOVA)에서 종속변수에 영향을 줄 수 있는 공변인(covariate)을 포함하여 분석하는 방법이다. 공변인을 포함하여 분석함으로써 공변인의 영향을 통계적으로 통제하여 독립변수의 순수한 효과를 추정한다. 하지만 개입 효과의 추정치는 모형에 포함된 공변인들에 따라 달라진다는 한계가 있다(Tein et al., 2018). 이는 공변인 선택이 신중하게 이루어져야 함을 나타낸다. 그리고 사전 측정치, 인구통계학적 변인 외에도 측정되지 않은 다양한 혼입변수(confounder)가 존재할 수 있다(Tein et al., 2018). ANCOVA의 한계와 공변인 선택의 중요성은 Steiner, Cook, Shadish, and Clark (2010)의 연구에서도 확인된다. 이 연구는 동일한 개입을 RCT 집단과 비동등 집단 설계 집단에 적용한 뒤, 다양한 공변인 조합을 사용하여 처치 효과의 편향이 얼마나 감소하는지 비교했다. 분석 결과, 공변인을 추가하는 것이 항상 편향을 줄이는 것이 아니었다. 특히 결과 변수와 관련성이 낮은

공변인을 포함할 경우 오히려 편향이 증가하는 현상이 나타났다. 이처럼 어떤 공변인을 포함하느냐에 따라 개입 효과의 추정 결과가 달라질 수 있음을 잘 보여준다.

준실험설계에서 선택적 차이를 효과적으로 통제하기 위해 성향 점수 분석(propensity score analysis)이 대안으로 제시된다. 성향 점수는 처치 집단에 할당될 확률을 의미하며(Rosenbaum & Rubin, 1983), 개입 전에 수집한 공변수 정보들을 이용하여 추정할 수 있다. 추정된 성향 점수와 matching, weighting 또는 stratification 방법을 이용하면 관찰된 공변수에 대해서는 선택 편향의 영향을 통제하여 순수 개입의 효과를 추정할 수 있다. 이를 통해 비동등 집단 설계의 한계인 내적 타당성에 대한 위협을 완화할 수 있다. 성향 점수가 같은 처치집단과 통제집단 사람들은 같은 공변인 분포를 갖게 되므로, 사전 동질성 검정과 달리 성향 점수 분석을 이용하여 관찰된 공변인들이 처치 효과 추정에 주는 영향을 통제할 수 있다(Austin, 2011). 공분산분석도 통계적으로 관찰된 공변인의 영향을 통제할 수 있으나, 성향 점수 분석과 달리 많은 공변인을 모형에 포함하여 모형이 복잡해지면 과대 적합(Babyak, 2004) 또는 다중공선성 문제(Lee & Acharya, 2022)가 야기될 수 있다. 또한 공분산분석은 분석에 포함할 공변인을 사후에 정하기 때문에, 연구자가 원하는 결과가 나오도록 모형에 포함할 공변인을 취사선택하는 *p*-hacking 문제가 발생할 수 있다(Kim, 2019). 성향 점수 분석에서는 많은 공변인들의 정보를 하나로 압축한 성향 점수를 이용하기 때문에 과대 적합 또는 다중공선성 문제가 발생하지 않는다. 또한 성향 점수 분석은 연구 계획 단계에서 분석에 포함할 공변인을 고려하여 정하기 때문에 *p*-hacking 문제가 발생하지 않는다.

소규모 표본 연구에서의 방법론

본 조사 결과에 따르면, 개입 연구에서 많은 참가자를 모집하는 것이 어렵고 소규모 표본의 연구가 흔히 이루어진다는 점을 알 수 있다. 예방 연구에서 연구자가 통제할 수 없는 이유(예, 적은 참가자 모집, 초기 참가자의 중도 탈락 등)로 표본 크기가 작아지고 이는 낮은 통계적 검정력을 일으킨다. 낮은 통계적 검정력은 실제 효과가 있어도 발견하지 못할 확률로 제 2종 오류율(Type Ⅱ error rate, β)이 증가함을 나타낸다. 특히, 본 고찰에서는 집단별 표본 크기가 30을 넘지않는 연구가 관찰되었으며, 이러한 표본구성은 효과 추정의 불확실성을 함께 고려한 해석을 요구한다. 소규모 표본이 가진 한계에도 불구하고 Hopkin, Hoyle, and Gottfredson (2015)와 Bacchetti, Deeks, and McCune (2011)은 예방 연구의 결과가 통계적 유의성 여부를 넘어서 과학적 가치를 지닌다고 강조한다. 따라서 본 논의에서는 표본이 상대적으로 적은 경우 연구자들이 연구 결과의 활용도를 높일 수 있는 해석과 대안적인 분석 및 추정 방법을 제시하고자 한다.

첫 번째, 연구 결과 해석 시 유의성 검정 결과 외에 개입 효과의 추가적인 정보를 제공하는 효과크기(effect size)와 신뢰구간(confidence interval)의 보고 및 해석이 권장된다. 영가설 유의성 검정(Null Hypothesis Significance Testing, NHST)은 연구 결과가 통계적으로 유의한지 여부만을 제시하며 효과의 실질적인 크기나 추정치의 불확실성을 반영하지 못한다는 한계를 가지고 있다(Cumming, 2014). 특히 영가설 유의성 검정은 표본 크기에 민감하며 표본 크

기가 클수록 통계적 검정력이 증가한다. Téllez et al.(2015)의 연구에 따르면, 0부터 100까지의 삶의 질 척도를 사용하는 연구에서 각 집단의 표본 수가 50명일 경우 두 집단 간 평균 차이가 약 10점 이상이어야 $p < .05$ 수준의 유의성을 보일 수 있는 반면, 각 집단의 표본 수가 500명으로 늘어나면 3점 정도의 차이도 통계적 유의성이 나타날 수 있다고 보고한다. 따라서 연구 결과를 해석할 때 단순한 p값에 의존하기보다, 효과크기를 통해 개입의 실질적인 크기를 나타내고 신뢰구간을 통해 해당 추정치의 불확실성을 고려해야 한다. American Psychological Association(2010)에서 "NHST는 출발점에 불과하며 효과크기와 신뢰구간 등 추가적인 보고를 통해 연구 결과의 의미를 완전히 전달할 수 있다."고 명시하였다.

효과크기는 프로그램이나 개입의 실제적인 효과가 얼마나 큰지, 작은지 정량적으로 평가하는 지표이다. Téllez 등(2015)은 효과크기가 변수 간 관계의 강도를 의미하고 p값은 이러한 효과의 강도가 우연일 확률을 나타낸다고 설명한다. Rosnow and Rosenthal(1989)은 p값이 유의하지 않더라도 효과크기 산출이 가능하고 효과크기를 보고하는 것이 추후 연구에 필요한 표본 크기를 결정하는 데 도움을 준다고 강조한다. 흔히 사용되는 효과 크기 지표로는 두 집단 간 평균 차이를 표준화하여 나타낸 $Cohen's\ d$와 각 독립 변수가 종속 변수의 분산에 기여하는 정도를 나타내는 부분 에타 제곱(η_p^2)이 있다.

신뢰구간은 추정된 효과크기 또는 모수의 불확실성을 정량적으로 보여주는 범위이다 (Téllez et al., 2015). 95% 신뢰구간은 반복해서 실험했을 때 95% 확률로 그때마다 산출된 신뢰구간이 모수를 포함함을 의미한다. 신뢰구간으로 가설검정 시 구간 내에 0이 포함되면 영가설을 기각하지 않고, 0을 포함하지 않으면 영가설을 기각한다. 신뢰구간은 영가설의 기각 여부 뿐만 아니라 추정치의 불확실성에 대한 정보도 제공한다. 넓은 신뢰구간은 상대적으로 포괄적인 범위에서 모수가 포함된다는 것을 의미하며 추정치의 정밀도가 낮은 것을 나타낸다. 좁은 신뢰구간은 상대적으로 특정한 범위에서 모수가 포함된다는 것을 의미하며 추정치의 정밀도가 높은 것을 나타낸다.

두 번째, 비모수 검정의 적절한 활용과 해석에 대한 고찰을 하고자 한다. 연구자는 데이터의 특성과 연구 가설에 따라 모수 검정(parametric test)과 비모수 검정(nonparametric test)을 선택해야 한다(Politi, Ferreira, & Patino, 2021). 모수 검정은 모집단이 특정한 수학적 분포를 따른다는 가정 하에 모수 추정 및 가설 검정을 한다. 모수 검정은 평균 차이를 검증하는 independent t-test, paired t-test, ANOVA와 변수 간 관계를 검증하는 pearson 상관분석, 회귀분석 등을 포함한다. 비모수 검정은 모집단에 대한 특정한 분포를 가정하지 않고 데이터의 순위(rank)나 중앙값(median) 등을 기반으로 비교한다. 비모수 검정은 중앙값 차이를 검증하는 Mann-Whitney U test, Wilcoxon signed-rank test, Friedman test와 변수간 관계를 검증하는 Spearman 상관분석 등을 포함한다. 연구자들이 흔히 갖는 오해 중 하나는 표본 수가 적으면 비모수 검정을 선택해야 한다는 것이다. 그러나 표본 크기가 작다는 사실 자체는 비모수 검정을 선택할 이유가 되지 않는다. 소규모 표본을 사용하더라도 두 집단 간 평균의 차이가 유의하게 다른지 분석하는 것이 목적이라면 비모수 검정보다 모수 검정이 적절하다. 비모수 검정은 측정 도구가 서열

척도(ordinal scale) 인 경우 적절하며 두 집단 서열(순위) 차이를 분석할 수 있다. 따라서 소규모 표본 연구 설계에서 측정 도구의 특성과 연구 목적을 고려하여 평균 비교가 필요한 경우에 모수 검정(t-test, ANOVA, ANCOVA 등)을, 서열(순위) 비교가 목적일 경우에 비모수 검정을 선택해야 한다. 그리고 비모수 검정 후 결과 해석 시 평균 차이와 순위의 차이를 명확히 구분할 필요가 있다. 비모수 검정은 통계량으로 순위와 중앙값을 다루기 때문에 결과를 해석할 때 각 집단의 중앙값과 분포 형태를 보고하고 “두 집단 순위에 유의한 차이가 관찰되었다.”로 해석할 수 있다(de Winter & Dodou, 2010).

세 번째, 소규모 표본 연구의 한계를 보완하기 위해 부트스트랩(bootstrap) 방법을 적용하는 것은 추천하지 않는다. 부트스트랩은 관찰된 자료에서 복원 추출(resampling with replacement)을 반복하여 분포와 신뢰구간을 추정하는 기법이다. 하지만 부트스트랩 분포는 원래 표본을 반영하기에, 대표성이 부족한 소규모 표본의 근본적인 문제점을 보완할 수 없다. Zientek and Thompson(2007)은 다음과 같이 강조한다. “부트스트랩은 우리가 가진 데이터의 한계를 초월하게 해주는 마법이 아니다. 대표성이 부족한 표본을 대표성 있게 만들어 주는 것도 불가능하다.”즉, 표본이 작고 편향되어 있다면 부트스트랩을 통해 얻은 추정치 또한 왜곡될 가능성이 크다. 또한 Koopman 등(2015)은 심리학에서 매개 효과를 검증하기 위해 흔히 사용되는 부트스트랩 방법이 작은 표본 크기에서 신뢰하기 어렵다고 지적했다. 특히 20명에서 80명 사이의 표본에서 부트스트랩은 통계적 검정력 부족과 높은 1종 오류의 위험이 동시에 존재했다(Koopman et al., 2015). 따라서 소규모 표본에서 부트스트랩 사용에 대한 신중한 고려를 강조하고, 대안으로 베이지안 추정법을 제안한다(Koopman et al., 2015).

마지막으로, 소규모 표본에서 베이지안 추정을 사용하면 중심극한정리(Central Limit Theorem; 이하 CLT)가 깨지는 한계를 극복할 수 있다(Sawada, 2021). 빈도주의 접근법에서는 표본 크기가 작으면 CLT가 제대로 작동하지 않아 표집분포(표본평균의 분포)가 정규분포와 크게 다를 수 있다. 이로 인해 계산된 p값이나 신뢰 구간에 대한 통계적 추론의 신뢰성이 저하된다. 대안으로 베이지안 추정(Bayesian Estimation)을 사용할 수 있다(McNeish, 2017). 베이지안 추정은 자료와 사전 분포(prior distribution)를 결합하여 모수의 사후 분포(posterior distribution)를 추론한다. 표본이 적더라도 사전분포의 정보를 활용하여 정확하고 신뢰할 수 있는 추정을 할 수 있다. Schoot et al. (2015)의 시뮬레이션 연구에서는 매우 작은 표본(n=8, 14, 22 세 가지 다른 표본 크기)에서 생성한 데이터를 최대우도 추정법(ML)와 베이지안 추정법으로 분석하여 각 방법의 모수 추정 편향, 95% 신뢰구간에 포함될 확률, 검정력(Power), 평균 제곱 오차(MSE)를 비교 평가했다. 베이지안 분석이 최대우도 추정법보다 높은 검정력, 더 적은 모수 편향, 더 높은 신뢰구간 포함 확률을 보였다. 물론 베이지안 방법도 주의할 점이 있다. 사전분포의 적절한 설정이 중요하고 특히 작은 표본에서는 사전 분포의 영향이 더 크다. 따라서 분석 전에 가지고 있는 지식이나 정보를 활용하여 신중하게 사용하고 사전 분포에 대한 민감도 분석을 수행하여 사전 분포 설정이 결과에 미치는 영향을 확인해야 한다.

매개 또는 조절효과 분석

개입 연구에서 매개 및 조절효과 분석은 개입이 “어떻게”, 그리고 “누구에게” 작동하는지 알아보기 위한 중요한 연구 질문들을 탐구할 수 있게 해준다 (MacKinnon & Luecken, 2008). 하지만, 본 조사 결과, 전체 분석 대상 중 4% 이하의 매우 적은 비율의 연구에서만 이러한 매개 및 조절효과 분석에 관심을 갖는 것으로 나타났다. 이는 국내 개입 연구가 효과가 있었는가에 집중하는 반면, 효과가 나타난 경로와 효과가 달라지는 조건에 대한 보고는 상대적으로 제한적일 수 있음을 시사한다. 이러한 현상이 나타나는 이유는 다음과 같이 세 가지 원인으로 추론해볼 수 있다.

개입 연구에서 매개 및 조절효과 분석의 중요성 간과

현재 한국심리학회 및 산하 학회들에서 개최하는 학술대회 발표나 발간하는 학술지 논문들을 보면 매개 및 조절효과 분석의 비중이 높은 것을 볼 수 있다. 예를 들어, 2024년 제 78차 한국심리학회 연차학술대회 초록집 중 회원 포스터 섹션에서만 총 140건의 포스터 중 62건(44%)이 매개(20건), 조절(32건), 혹은 조절된 매개(10건) 연구를 진행했다. 이 중 대부분 상관 연구(예: 설문조사)로서 수집한 변수들을 가지고 매개, 조절, 또는 조절된 매개 효과를 탐구하고 있었다. 총 140건의 포스터 중 단 2개의 연구만이 개입/실험 연구였다. 본 국내 실태조사에서 나타난 것처럼 개입 연구에서 매개 및 조절 효과 분석은 상대적으로 덜 이루어지는 것 같다. 이는 전통적으로 실험설계 연구에서 독립변수와 종속변수를 제외한 제 3의 변수들을 통제해야 한다는 인식이 지배적이라서 그런 것 같다. 하지만, 현장에서 실시하는 개입 연구의 경우 제 3의 변인들을 통제하는 것은 현실적인 어려움이 있을 수 있다. 이때 연구자들이 사용할 수 있는 방법은 제 3의 변인들을 측정하여 연구 모형에 포함시키는 것이다. MacKinnon과 Luecken (2008)에 의하면, 제 3의 변수는 크게 네 가지로 나눌 수 있다: 혼입변수(confounder), 공변수(covariate), 조절변수(moderator), 그리고 매개변수(mediator). 혼입변수와 공변수는 개념적으로 비슷한데, 독립변수가 종속변수에 미치는 효과의 크기나 통계적 유의성에 영향을 미치는 변수는 혼입변수, 어느 정도 영향은 미치지만 그 크기가 작고 연구자가의 관심변수가 아니라면 공변수라고 부른다. 매개변수는 독립변수가 종속변수에 영향을 미칠 때 그 중간에서 역할을 하는 변수이다. 즉, 중간에서 독립변수가 어떻게 종속변수에 영향을 미치는지 그 기제를 설명하는 변수가 매개변수이다. 개입 연구에서 매개변수가 중요한 이유는 개입 처치가 어떻게 종속변수에 영향을 주는지 설명해줄 수 있는 변수이기 때문이다. 마지막으로, 독립변수가 종속변수에 미치는 영향이 제 3의 변수에 의해 달라지면 그때 제 3의 변수를 조절변수라고 부른다. 조절변수는 주로 인구통계학적 변수들 혹은 종속변수의 기저 측정치를 사용하는데, 어떤 특징을 가진 사람들에게 개입 처치가 생각한대로 영향을 미쳤는지(혹은 영향을 미치지 않았는지) 알려주므로 개입 연구에서 중요한 역할을 한다. 이렇게 제 3의 변수들을 고려하는 것은 중요한데, 특히나 매개나 조절변수들을 고려하는 다양한 연구 질문들이 가능하다는 점에서 개입 연구에서 핵심적인 역할을 한다. 이에 따라 앞으로 조절 또는 중재가 포함된 연구가 활발히 이루어지기를

기대한다.

매개 및 조절효과 분석 방법론에 대한 오해

많은 연구자 및 실무자들이 매개분석을 하기 위해서는 일단 독립변수가 종속변수에 미치는 영향이 유의해야하는 것으로 알고 있다. 이는 Baron과 Kenny(1986)가 제시한 매개효과를 검정하기 위한 4단계 중 첫 번째 단계로, 독립변수와 종속변수 간 유의한 관계를 보여야 한다는 것 때문에 생긴 오해이다. Baron과 Kenny(1986) 이후로 독립변수와 종속변수 간 유의한 관계가 없어도 매개효과가 유의할 수 있다는 것을 보여줬으며(Shrout & Bolger, 2002), 개입 연구에서 개입이 종속변수에 미치는 효과가 유의하지 않아도 매개효과를 검정해보는 것이 중요할 수 있다고 강조한다(O'Rourke & MacKinnon, 2018).

개입 연구에서 '누구에게 개입이 가장 효과적인가'에 대한 질문에 답하기 위해 조절효과 분석을 실시할 수 있다. 하지만 연구자들이 사전에 이러한 연구 질문을 하기보다는 사후에 개입 효과가 유의하지 않았을 때 이를 설명하려고 조절효과 분석을 실시하는 연구가 많은 것 같다. 이때 문제는 처음부터 조절효과를 고려하지 않았으면, 적절한 조절변수를 측정하지 않았을 가능성이 있다. 또한, 통계적 검정력이 부족하다고 생각하여 조절효과를 처음부터 회피하거나, 통계적으로 유의하지 않은 조절효과를 보고하지 않는 경향도 있는 것 같다. 하지만, 통계적 검정력이 부족하더라도 흥미로운 연구 질문을 탐색해보고 현실적 의미를 해석하는 과정은 필요하다.

이러한 방법론적 오해를 해소하고 분석의 신뢰성을 확보하기 위해서는 방법론 교육이 활발해지는 것이 필요하다. 실제로 이러한 필요성에 따라 2010년대 중반부터 국내 여러 대학의 심리학과에서는 계량심리학 또는 심리측정 분야로 교수 임용이 이루어지고 있다. 이에 따라 각 대학의 교과 및 비교과 과정으로 방법론 관련 강의 개설이 크게 늘고 있다. 이뿐만 아니라 한국심리측정평가학회 및 기타 기관에서도 통계 방법론 워크숍을 활발하게 진행하고 있다. 더불어 한국심리학회지: 일반과 같은 학술지에 매개 및 조절 효과 분석과 관련된 방법론 논문들이 게재되고 있다 (정선호, 서동기, 2016; 정선호, 양태석, 박중규, 2019). 이러한 방법론 논문들을 포함하여 독자들이 특정 방법론을 잘 학습할 수 있도록 도움을 줄 수 있는 방법론 논문들이 더 많이 나왔으면 하는 바람이다.

적은 표본수와 낮은 검정력 문제

본 연구에서의 조사 결과 개입 연구들 중 90%가 넘는 연구들의 표본 크기가 집단별로 30을 넘지 않는 것으로 나타났다. 표본 크기가 작으면 통계적 검증력이 낮아지기 때문에 실제 효과가 있어도 통계적으로 유의한 효과라고 결론 내리기가 어려워진다. 특히, 조절효과를 검정할 때 필요한 표본 크기는 일반적으로 같은 효과크기를 가진 주효과를 발견할 때 필요한 표본 크기의 4배 까지도 필요한 것으로 알려져 있다(Aiken & West, 1991). 이러한 현실적인 어려움 때문에 매개 및 조절효과 연구를 실행하기 어려운 것일 수 있다. 하지만, 위에서도 언급한 것과 같이, 매개 및 조절효과 분석은 개입 연구에서 '어떻게'와 '누구에게' 개입의 효과가 있었는지를 규명할 수 있는 중요한 부분을 차지하기 때문에 개입 연구에서 확대되어야 할 연구 분야이다.

논 의

지금까지 국내 정신건강 개입 연구 177편의 분석 결과에서 확인된 실증적 현황을 핵심 방법론적 쟁점과 연계하여 제언하였다. 본 논의에서는 이를 바탕으로 연구자와 실무자가 현장에서 즉각적으로 적용할 수 있는 구체적인 가이드라인을 도출하고자 하였다.

개입 프로그램 평가의 체계화

WHO는 전 세계적으로 프로그램 평가의 체계화와 표준화를 강조한다. WHO evaluation practice handbook(2013)에 따르면 평가는 정책 또는 개입에 대한 체계적이고 공정한 검토로 정의되며 프로그램의 과정, 결과와 인과관계를 분석하며 목표 달성도를 판단한다. 더불어, 개입의 적절성(relevance), 효과성(effectiveness), 효율성(efficiency), 영향(impact) 및 지속가능성(sustainability) 등을 평가하도록 권고한다. 특히, WHO는 이러한 평가를 돕기 위한 단계별 체크리스트와 템플릿 등을 제공한다. 예를 들어, 수행되는 모든 평가가 준수해야하는 체크리스트(Checklist for compliance with the WHO evaluation policy), 최종 평가 결과를 보고하기 전에 확인해야하는 체크리스트(Checklist for evaluation reports), 평가 과정을 계획하는 템플릿(Evaluation workplan template) 등의 자료를 통해 프로그램이 성공적으로 실행되기 위한 평가 기준을 제공하고 있다.

한편, 로직모델(Logic Model)이 정신 건강 개입의 설계와 평가에서 널리 활용된다. 로직모델은 프로그램이 어떻게 작동하는지 시각적으로 보여주는 도식도이다. 프로그램의 투입(input)-활동(activity)-산출(output)-성과(outcome)-영향(impact)의 인과 흐름을 구조화하여 제시함으로써 프로그램이 왜, 어떻게, 무엇을 변화시키는지 한눈에 체계적으로 설명한다. 그리고 로직모델은 프로그램 기획-실행-평가-보고 단계에서 모두 활용될 수 있고 프로그램 개발자, 실행자, 평가자 간 공통된 합의와 이해를 돕는다. 구체적으로, 프로그램의 목표와 기대되는 효과, 투입하는 자원 등을 사전에 정리하면서 프로그램 기획 단계를 체계화한다. 그리고 프로그램이 진행될 때는 계획대로 수행되고 있는지 점검한다. 평가 단계에서는 로직모델로 분석한 프로그램의 기대 효과의 지표를 측정하고 분석한다. 마지막으로 프로그램의 효과를 보고할 때는 활동, 결과, 영향의 흐름을 명확히 설명할 수 있다. 로직모델의 템플릿은 여러 연구소, 기관에서 찾아볼 수 있고 강조점에 따라 다르지만 근본적인 도식도는 비슷하다. 이처럼 정신건강 증진 및 예방처럼 복잡한 변화 과정을 다루는 개입에서 프로그램의 타당성을 확보하고, 평가 지표 설정과 연구 설계를 구체화하는 데 핵심적인 역할을 한다.

본 연구는 국내 학술지에 게재된 177편의 논문 분석 내용과 로직 모델의 개념을 참고하여 실제 평가 현장에서 연구자들이 활용할 수 있는 평가 체계화 방안을 제시하고자 한다. 구체적으로는 프로그램 평가 체크리스트와 평가 기록용 요약표를 제안한다. 이 도구들은 연구자가 설계와 분석의 전 과정을 스스로 사전에 검토하고 정리할 수 있게 돕는다. 이는 향후 연구에서 선택 편향의 통제나 신뢰구간 보고 등 방법론적 정밀성을 강화할 수 있는 실질적인 지침이 될 것이다. 특히 체크리스트의 10번 항목은 통계적 유의성을 넘어 효과크기와 신뢰구간을 함께 확인하도록 권장함으로

써, 개입 효과의 실질적 의미를 전달하도록 돕는다. 이러한 체계적인 평가 도구의 활용은 국내 정신건강 개입 연구의 과학적 엄밀성을 상향 평준화하고, 나아가 연구 결과가 실제 정책과 현장에 효과적으로 반영될 수 있는 근거 기반을 공고히 하는 데 기여할 것이다.

프로그램 평가 체크리스트

프로그램 평가 체크리스트는 표 6에 제시했다. 프로그램 기획 영역에는 프로그램의 이론적 근거 제시 여부와 프로그램의 기대되는 목표 명확성을 점검한다. 아직 프로그램의 목표가 명확하지 않다면 서론에서 언급한 로직 모델(logic model)이나 변화 이론(Theory Of Change; TOC)을 통해 개입 프로그램의 기대되는 목표를 설정할 수 있다. 프로그램 운영 영역에는 프로그램 진행의 충실도(fidelity)와 연구 윤리 준수 여부를 확인한다. 실행 단계의 충실도는 프로그램 진행자의 충실도와 프로그램 참여자의 성실도로 구분할 수 있다. 더불어 인간 대상 개입 연구는 연구 시작 전에 반드시 기관생명윤리위원회(IRB)의 심의를 받아야 한다. 연구 설계 영역의 항목들은 사용된 연구 설계의 타당성과 충분한 표본 크기 확보를 다룬다. 만약 적절한 표본 크기에 대한 확신이 부족하다면, G*Power(Faul et al., 2007) 또는 시뮬레이션 연구를 통해 효과 크기(effect size), 유의수준(α), 검정력(power, $1-\beta$)을 고려한 사전(a priori) 검정력을 분석하여 적정 표본수를 산출할 것이 권장된다(Erdfelder et al., 1996). 측정 도구 영역에서는 프로그램 목표에 적합한 측정 도구의 선택과 해당 도구의 신뢰도 및 타당도를 점검한다. 평가하고자 하는 프로그램의 목표와 연관된 척도를 선정하고 필요하다면 척도를 개발한다. 척도 개발 시 사전에 전문가 검토를 거쳐 내용타당성을 확보해야 한다. 마지막으로 통계 분석 영역에서는 연구 설계와 자료 특성에 적절한 분석 방법 선택과 분석 결과 여부를 점검한다. 결과를 보고할 시 해석을 풍부하게 하기 위해 효과크

표 6. 프로그램 평가 체크리스트

평가 영역	세부 항목
프로그램 기획	1. 프로그램의 이론적 근거가 제시되었는가?
	2. 프로그램의 기대되는 목표를 명확하게 설정했는가?
프로그램 운영	3. 프로그램이 계획된 대로 충실히 실행되고 있는가?
	4. 연구 윤리를 잘 지켰는가?
연구 설계	5. 평가 연구의 설계를 명확히 수립하였는가?
	6. 충분한 표본 크기를 확보하여 통계적 검정력이 충족되었는가?
측정 도구	7. 프로그램 목표에 맞는 측정 도구를 사용하였는가?
	8. 측정도구의 신뢰도와 타당도를 확인하였는가?
통계 분석	9. 사용된 통계 분석 방법이 연구 설계와 자료 특성에 부합하는가?
	10. 주요 결과에 대한 통계적 유의성뿐 아니라 효과크기, 신뢰구간을 함께 확인하였는가?

표 7. 프로그램 기록용 요약표

프로그램 기본정보		
프로그램명:		
담당 평가자:	평가일: 년 월 일	
대상 집단:	연령:	인원: 총 명 ☐ 실험집단 - 명 ☐ 비교집단 - 명 ☐ 통제집단 - 명
개입 기간 : ( )주	주 ( )회	회기당 시간: ( )분
프로그램 개요 및 실행		
프로그램 목표 :		
이론적 기반:		
예상되는 효과:		
프로그램 운영자:		
운영자의 가이드라인 준수 여부:		
윤리 승인 : (승인 번호:)		
연구 설계 및 분석		
연구 설계 : ☐ 무작위 대조군 ☐ 준실험 ☐ 단일군 ☐ 사례연구 (기타:)		
측정 시점 : ☐ 사전 ☐ 사후 ☐ 추적 (간격:)		
사용된 측정 도구 : 1) (신뢰도: 타당도:) 2) (신뢰도: 타당도:) 3) (신뢰도: 타당도:) 4) (신뢰도: 타당도:)		
독립변수:	종속변수:	
제 3의 변수: ☐ 조절변수 ☐ 매개변수 ☐ 공변수 ☐ 혼입변수 변수명 :		
집단 간 사전 동질성 테스트 및 분석 방법:		
개입 프로그램 효과 분석 방법:		
분석에 사용한 소프트웨어 프로그램 :		
결과 및 향후 계획		
결과 요약:		
향후 계획:		
한계점:		
개선점:		

기와 신뢰구간 등을 함께 보고하는 추세이다. 체크리스트를 활용함으로써 평가 과정의 체계화가 가능하고 더 나아가 프로그램 평가 연구를 설계하거나 논문을 작성할 때 가이드라인으로 사용될 수 있다.

프로그램 평가 기록용 요약표

프로그램 평가 기록용 요약표는 표 7에 제시하였다. 요약표는 개입 프로그램의 핵심 정보를 일목요연하게 정리하는 도구로서 연구 논문의 결과 요약이나 실무 현장 보고서 등에 활용될 수 있다. 프로그램 기본 정보는 프로그램이 언제, 누구에게, 얼마나 시행되었는지 포함한다. 프로그램 개요 및 실행은 해당 프로그램이 무엇을 목표로, 어떤 근거 기반으로 설계되었는지, 누구에 의해 실시되었는지, 충실하게 실행되었는지를 기록한다. 연구 설계 및 분석은 프로그램 효과성을 평가하기 위한 연구 설계와 분석 방법의 핵심 사항들을 정리한다. 마지막으로 결과 및 향후 계획은 프로그램 평가의 주요 결과를 요약하고 이를 바탕으로 향후 계획이나 제언을 기술한다. 이는 단순히 계획이라기보다, 해당 평가 결과를 해석하고 함의점을 정리하는 것이다. 예를 들면, “본 프로그램의 효과를 확인하였으니 광역시 범위로 확대 실시할 것”과 같은 제언을 적을 수 있다. 한계점에서는 개입 프로그램을 진행하면서 부딪힌 어려운 점을 적을 수 있다. 예를 들면, “자기보고식 설문에 의존하여 객관적 지표 부족” 혹은 “중도 탈락자 다수 발생”과 같은 한계를 적을 수 있다. 모든 연구에는 한계가 있고 이를 투명하게 공개하는 것이 발전의 시작이다. 이러한 내용은 향후 유사한 프로그램을 개발 및 평가하려는 연구자들에게 유용한 정보가 된다. 프로그램 평가 요약표는 개입 프로그램의 요약을 가능하게 하며 이러한 표화된 요약이 있다면 추후 프로그램들 간 특성과 효과를 비교할 수 있다.

연구의 한계와 후속 연구 방향

본 연구는 다음과 같은 한계를 가진다. 첫째, 한국심리학회 및 산하 학회지에 최근 10여년간 발표된 논문 177편만을 검토하였기에, 국내 수행되는 모든 예방 및 진흥 프로그램을 대표한다고 보기 어렵다. 특히, 학술지에 보고되지 않았거나 실제 현장에서 이루어진 프로그램의 특성과 평가 방법이 반영되지 못했을 가능성이 있다. 둘째, 본 연구는 측정도구에 대한 정밀한 심리측정 관련 정보를 충분히 다루지 못하였다. 개입 프로그램 연구에서 심리측정은 매우 중요한 위치를 차지한다. 개입 프로그램의 효과성을 탐구하기 위해서는 심리측정도구의 신뢰도 및 타당도에 대한 검증이 앞서야 한다. 본 조사에서는 연구 설계 및 통계 방법론에 집중하여 심리측정에 대한 조사가 미흡했다. 셋째, 분석 대상이 심리학 분야 학술지로 한정되었다는 제한이 있다. 이러한 학술지 범위 설정은 심리학 분야 개입의 방법론적 특징을 먼저 분석하기 위한 선별이었다. 하지만 전체적인 개입 프로그램을 파악하거나 심리학 분야 개입 프로그램만의 고유한 특징을 파악하기엔 어려웠다.

추후 연구에서는 분석 범위와 분야 및 내용의 확장이 필요하다. 국내 학술지를 포함하여 학술지에 실리지 않는 정부 연구 보고서, 기관의 평가 보고서, 석박사 학위논문 등 자료를 포괄적으로 수집 및 분석함으로써 전체적인 개입 프로그램의 평가 방법론을 파악할 수 있다. 이 뿐만 아니라 사회복지학, 교육학 등

타 학문 분야의 개입 연구들과의 비교 분석이 필요하다. 이를 통해 심리학 기반 개입의 강점과 보완점을 명확히 알 수 있다. 더 나아가 한국 뿐만 아니라 해외 개입 연구와 비교하여 국내 연구의 방법론적 특징을 선별하고 국내 연구와 해외 연구의 체계적 평가를 비교 및 표준화해볼 수 있다. 그리고 향후 연구는 시간적 범위를 넓혀 지난 수십년 간 축적된 개입 프로그램 연구를 종합적으로 분석할 수 있다. 마지막으로 후속 연구에서는 개입 프로그램 평가 연구에서 심리측정의 관행들에 대한 심층적인 분석이 필요하다. 개별 논문에서 사용된 측정 도구의 신뢰도와 타당도, 종속변수의 특질 변수와 상태 변수 구분, 신뢰도와 타당도 지수에 기반한 가정 등 고려할 수 있다.

본 연구는 국내 증진 및 예방 프로그램의 평가 연구에 대한 체계적인 현황 파악과 방법론적 통찰을 제공함으로써 학문적인 의의가 있다. 최근 10년 간 프로그램의 효과성 연구를 종합하여 효과적인 프로그램 평가를 위한 기초 자료를 구축했다. 특히, 내적 타당성 확보를 위한 방법, 소규모 표본 연구에서의 통계 분석 활용, 조절 및 매개효과의 중요성 등 향후 프로그램 평가 방법론을 개선하고 발전시키는 데 토대를 제공한다. 학문적 의의뿐만 아니라 실제 현장에서의 프로그램 평가 개선에도 중요한 함의를 가진다. 177개의 연구 분석을 통해 얻은 방법론적 제언들은 프로그램 기획 및 운영 시 직면하는 문제를 해결하는 데 도움을 줄 수 있다. 예를 들어, RCT가 어려운 경우 준실험 설계의 활용과 선택 편향을 줄일 수 있는 통계 분석 방법, 소규모 집단에서도 효과크기와 신뢰구간 보고를 통한 결과 해석 등은 과학적인 평가를 수행하도록 돕는다. 그리고 실제 프로그램 평가 시 이용할 수 있는 가이드라인을 제시함으로써 프로그램 평가의 체계화를 구축한다. 이러한 적용을 통해 프로그램의 평가 및 지속적 개선, 나아가 효과가 증명된 개입은 정책에 반영되어 정신건강 증진에 기여할 수 있을 것이다.

참고문헌

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, *66*(3), 411-421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bacchetti, P., Deeks, S. G., & McCune, J. M. (2011). Breaking free of sample size dogma to perform innovative translational research. *Science Translational Medicine, 3*(87), 87ps24. https://doi.org/10.1126/scitranslmed.3001628

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182. https://doi.org/10.1037//0022-3514.51.6.1173

Choi, K.S., Nam, J.A., Ko, B.S., Kim, J.E., Heo, E.H., & Lee, C.H. (2018). Mental Health Intervention for Adolescents: A School-Based Program to Address Social Anxiety. *Anxiety and Mood, 14*(2), 88-98. https://doi.org/10.24986/anxmod.2018.14.2.005

Choi, M., & Hector, M. (2012). Effectiveness of Intervention Programs In Preventing Falls: A Systematic Review of Recent 10 Years and Meta-Analysis [Review]. *Journal of the American Medical Directors Association, 13*(2), 9, Article 188.e13. https://doi.org/10.1016/j.jamda.2011.04.022

Cumming, G. (2014). The new statistics: why and how. P*sychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*(1), 1-11. https://doi.org/10.3758/BF03203630

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175-191. https://doi.org/10.3758/BF03193146

Hopkin, C. R., Hoyle, R. H., & Gottfredson, N. C. (2015). Maximizing the Yield of Small Samples in Prevention Research: A Review of General Strategies and Best Practices. *Prevention Science, 16*(7), 950-955. https://doi.org/10.1007/s11121-014-0542-7

Institute of Education Sciences. (2022). What Works Clearinghouse procedures and standards handbook (Version 5.0). U.S. Department of Education. https://ies.ed.gov/ncee/wwc/handbooks

Jung, S., & Seo, D. G. (2016). Assessing Mediated Moderation and Moderated Mediation: Guidelines and Empirical Illustration. *Korean Journal of Psychology: General, 35*(1), 257-282. https://doi.org/10.22257/kjp.2016.03.35.1.257

Jung, S., Yang, T. S., & Park, J. (2019). Testing Mediated Moderation Using Moderated Multiple Regression: Conceptual and Methodological Considerations. *Korean Journal of Psychology: General, 38*(3), 323-346. https://doi.org/10.22257/kjp.2019.09.38.3.323

Kim, D., Lee, Y.S., & Son, S.(2024). A Systematic Literature Review and Meta-analysis of DBT Programs in Korea. *Korea Journal of Counseling, 25*(5), 1-22.

Kim, J. J., Choi, T. Y., Park, J. H., & Kim, S. Y. (2017). The Effect of Mental Health Improvement Programs for Firefighters. *Anxiety and Mood, 13*(1), 17-24.

Koopman, J., Howe, M., Hollenbeck, J. R., & Sin, H. P. (2015). Small sample mediation testing: misplaced confidence in bootstrapped confidence intervals. *Journal of Applied Psychology 100*(1), 194-202. https://doi.org/10.1037/a0036635

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: an underrecognized confound in intervention research. *Psychological Bulletin, 130*(2), 289-303. https://doi.org/10.1037/0033-2909.130.2.289

Lee, S. W., & Acharya, K. P. (2022). Propensity score matching for causal inference and

reducing the confounding effects: statistical standard and guideline of Life Cycle Committee. *Life Cycle, 2*, e18. https://doi.org/10.54724/lc.2022.e18

MacKinnon, D. P., & Luecken, L. J. (2008). How and for whom? Mediation and moderation in health psychology. *Health Psychology, 27*(2s), S99-s100. https://doi.org/10.1037/0278-6133.27.2(Suppl.).S99

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348. https://doi.org/10.2466/03.11.Pms.112.2.331-348

Mascha, E. J., & Vetter, T. R. (2018). Significance, Errors, Power, and Sample Size: The Blocking and Tackling of Statistics. *Anesthesia & Analgesia, 126*(2), 691-698. https://doi.org/10.1213/ane.0000000000002741

McNeish, D. (2017). Challenging Conventional Wisdom for Multivariate Statistical Models With Small Samples. *Review of Educational Research, 87*(6), 1117-1151. https://doi.org/10.3102/0034654317727727

Ministry of Health and Welfare. (2021). 2021 mental health survey results: Presentation materials. https://www.ncmh.go.kr/kor/Cms/Edu/Edu_Gallery_View.aspx?cid=6491

National Center for Mental Health(2024). 2024 national survey of public knowledge and attitudes toward mental health in Korea. National Center for Mental Health. https://www.ncmh.go.kr/

National Center for Mental Health. (2024). A study on the effectiveness of community mental health programs: Final report on intensive case management. National Center for Mental Health. (ISBN 979-11-93630-51-8). https://www.ncmh.go.kr/ncmh/board/boardView.do?no=9637&fno=106&gubun_no=&menu_cd=04_02_02_04&bn=newsView&search_item=&search_content=&pageIndex=1

Neil, A. L., & Christensen, H. (2009). Efficacy and effectiveness of school-based prevention and early intervention programs for anxiety. *Clinical Psychology Review, 29*(3), 208-215. https://doi.org/10.1016/j.cpr.2009.01.002

O'Rourke, H. P., & MacKinnon, D. P. (2018). Reasons for Testing Mediation in the Absence of an Intervention Effect: A Research Imperative in Prevention and Intervention Research. *The Journal of Studies on Alcohol and Drugs, 79*(2), 171-181. https://doi.org/10.15288/jsad.2018.79.171

Organization, W. H. (2002). Prevention and Promotion in Mental Health.

Organization, W. H. (2022). Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Politi, M. T., Ferreira, J. C., & Patino, C. M. (2021). Nonparametric statistical tests: friend or foe? *Jornal Brasileiro de Pneumologia, 47*(4), e20210292. https://doi.org/10.36416/1806-3756/e20210292

Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Reichardt, C. S. (2019). Quasi-experimentation: A guide to design and analysis.

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

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276-1284. https://doi.org/10.1037/0003-066X.44.10.1276

Royse, D., Thyer, B. A., & Padgett, D. K. (2015). Program Evaluation: An Introduction to an Evidence-Based Approach.

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American College of Preventive Medicine, 33*(2), 155-161. https://doi.org/10.1016/j.amepre.2007.04.007

Sawada, T. (2021). Conditions of the Central-Limit Theorem Are Rarely Satisfied in Empirical Psychological Studies. *Frontiers in Psychology, 12*, 762418. https://doi.org/10.3389/fpsyg.2021.762418

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

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods, 7*(4), 422-445.

Sim, S.Y., Park, S.H., Lee, M.H., & Lee, M.S. (2015). A study on application of school-based mental health improvement programs: Focusing on sociality improvement programs.

Singh, V., Kumar, A., & Gupta, S. (2022). Mental Health Prevention and Promotion-A Narrative Review. *Frontiers in Psychiatry, 13*, 898009. https://doi.org/10.3389/fpsyt.2022.898009

Spurlock, D. R., Jr. (2018). The Single-Group, Pre- and Posttest Design in Nursing Education Research: It's Time to Move on. *Journal of Nursing Education, 57*(2), 69-71. https://doi.org/10.3928/01484834-20180123-02

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267. https://doi.org/10.1037/a0018719

Steiner, P. M., D., C. T., Wei, L., & and Clark, M. H. (2015). Bias Reduction in Quasi-Experiments With Little Selection Theory but Many Covariates. *Journal of Research on Educational Effectiveness, 8*(4), 552-576. https://doi.org/10.1080/19345747.2014.978058

Su, Y. L., & Reeve, J. (2011). A Meta-analysis of the Effectiveness of Intervention Programs Designed to Support Autonomy [Article]. *Educational Psychology Review, 23*(1), 159-188. https://doi.org/10.1007/s10648-010-9142-7

Téllez A., García C.H., & Corral-Verdugo V. (2015). Effect size, confidence intervals and statistical power in psychological research. *Psychology in Russia: State of the Art, 8*(3), 27-47.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N.

E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology 6*, 25216. https://doi.org/10.3402/ejpt.v6.25216

WHO evaluation practice handbook. (2013). World Health Organization.

Winter, J. F. C. d., & Dodou, D. (2010). Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012). *Practical Assessment, Research, and Evaluation, 15*(1). https://doi.org/10.7275/bj1p-ts64

Wong, V. C., & Steiner, P. M. (2018). Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings. *Evaluation Review, 42*(2), 176-213. https://doi.org/10.1177/0193841x18778918

Zientek, L. R., & Thompson, B. (2007). Applying the bootstrap to the multivariate case: bootstrap component/factor analysis. *Behavior Research Methods, 39*(2), 318-325. https://doi.org/10.3758/bf03193163

1차원고접수 : 2025. 06. 06
2차원고접수 : 2025. 12. 16
최종게재결정 : 2026. 02. 02

한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 1, 1-29
http://dx.doi.org/10.22257/kjp.2026.3.45.1.1

A Systematic Literature Review and Methodological Recommendations for Studies Evaluating the Effectiveness of Mental Health Intervention Programs in Korea

SoYeon Shin Yonsei University
EunGi Jeon Yonsei University
HanJoe Kim† Yonsei University

With growing attention to mental health promotion and prevention interventions, the systematic examination of research methodologies used to evaluate their effectiveness has become a fundamental task for both researchers and practitioners. This study conducted a methodological review of effectiveness studies on mental health promotion and prevention programs implemented in South Korea. A total of 177 peer-reviewed articles published over the past decade in journals affiliated with the Korean Psychological Association were analyzed, focusing on research design and statistical analysis methods. The review revealed two notable findings: first, over 90% of the studies had fewer than 30 participants per group; second, nearly half (48.5%) of the studies employed nonequivalent group designs. Based on these findings, this study proposes several methodological recommendations and practical guidelines to enhance program evaluation practices. The results provide foundational insights for improving the scientific rigor of future evaluations of mental health promotion and prevention programs.

Key words : mental health prevention and promotion, intervention program, program effectiveness evaluation, small sample size, nonequivalent group design