# 재정사업 사전검증체계 강화를 위한 연구

## - RCT 도입방안을 중심으로

2015. 12.

오영민 · 박노욱 · 강희우

한국조세재정연구원

# 서 언

새로운 재정사업이, 충분한 사전점검 없이 정치적인 요구에 의해 급격하게 도입되는 사례가 증가하면서, 시행착오로 인한 사회적 비용이 증가하고 다. 이러한 부작용을 막기 위해 500억원 이상의 대형사업의 타당성을 조사하는 예비타당성조사가 운영되고 있으나 정책적 효과를 정확히 계량화할 수 없는 교육인 복지사업에 대한 사전평가는 만족할 만한 수준이 아니다. 예비타당성조사가 나름대로 재정사업의 무분별한 시행을 막는 완충작용을 해왔을지라도 사업의 정확한 편익을 추정할 수 없는 사업에는 적용하기 어려운 한계가 존재하기 때문이다.

이런 의미에서 우리나라에서 최근 급속하게 확대되고 있는 복지나 고용 관련 사업을 사회실험 방식의 RCT 평가 가능성을 검토한 본 연구는 매우 시의 적절한 보고서라 판단된다. 특히 외국에서도 최근 교육이나 복지사업에 대해서는 RCT 평가방식이 적용되고 있기 때문에 최근 재정의 어려움을 겪고 있는 우리나라에 도입하는 데 이 보고서가 활용되기를 기대해 본다. 본 보고서는 RCT 평가기법을 이론적으로 기술하고 해외에서 RCT가 어떻게 활용되고 있는지를 소개하였다. 또한 아직 RCT가 크게 활용되고 있지 않은 우리나라의 적용가능성을 사전사후평가와 정책분야별로 탐색하였으며 실험의 적용과정에서 발생할 수 있는 장애요인과 대처방안을 제시하였다. 특히, 우리나라와 해외의 실제 사회실험 사례를 소개함으로써 독자들의 이해도를 높였으며 마지막 장에는 실험방식의 RCT 평가를 우리나라 실정에서 도입할 수 있는 방안을 구체적으로 제시하였다. 이러한 본고의 내용이 실제 정책평가를 담당하는 실무자에게 많은 도움이 될 것으로 기대한다.

본 보고서는 재정성과평가센터의 오영민 박사가 연구책임을 맡고 박노욱 박사와 강희우 박사가 공동으로 집필하였다. 공동 연구진 외에 정덕재, 전예원 연구원과 윤혜순 행정원도 자료수집과 정리에 많은 도움을 주었다. 또한

중간 및 최종평가에서 주옥같은 의견을 주신 전문가 및 공무원들에게도 깊은 감사를 표한다. 끝으로 본 보고서는 한국조세재정연구원의 공식 의견이 아니라 저자들의 개인 의견임을 밝힌다.

2015년 12월

한국조세재정연구원

원장 박 형 수

# 요약 및 정책적 시사점

## Ⅰ. 서 론

새로운 재정사업이 충분한 사전점검 없이 정치적인 요구에 의해 급격하게 도입되는 사례가 증가하면서 시행착오로 인한 사회적 비용이 증가하고 있다. 이러한 부작용을 줄이기 위해 예비타당성조사가 시행되고 있으나 교육인 복지사업 등 사업의 정확한 편익을 추정할 수 없는 사업에는 적용하기 어려운 한계가 있다. 이런 의미에서 본 보고서는 최근 해외를 중심으로 확대되고 있는 RCT(Randomized Control Trial) 방식의 실험평가 방식의 도입 가능성을 소개한다.

## Ⅱ. RCT(Randomized Control Trial) 방법론 소개

RCT는 진실험 방식의 실험평가로서 사회실험, 현장실험 등 다양한 형태로 불리운다. 이러한 RCT는 정책평가의 내생성(Endogeneity)를 제거하여 위해 사용된다. RCT에서는 내생성을 제거하기 위해 동질적인 처치집단과 통제집단을 구성하고 처지집단에만 정책을 시행하여 정책처치 후에 정책효과의 발생여부를 통제집단과 비교함으로써 추정한다. RCT에서는 두 집단 간 동질성을 확보하기 위해 무작위 추출(Randomization)을 시행한다. 그러나 현실 정책평가에서 무작위 추출이 어렵기 때문에 실제평가에서는 준실험 방식의 평가가 많이 사용된다. 준실험 방식의 평가에는 사전사후비교법, 단순평균비법, 매칭법, 회귀불연속법, 이중차분법의 평가방법이 있다.

## Ⅲ. RCT의 국제적 동향

실제 정책평가에서 실험방식의 평가가 많이 보급되지 않은 우리나라와는 달리 해외에서는 RCT에 의한 정책평가가 활성화되었다. 대표적인 예가 영국의 Behavioral Insight Team과 미국 오바마 행정부에서 RCT를 활성화하기 위해 시행된 증거기반 정책평가(Evidence base Evaluation)이다. 특히 개발도상국의 빈곤문제를 해결하기 위한 정책의 효과성을 과학적으로 평가하기 위해 MIT에 J-PAL이 설치되었다. 특히 J-PAL은 많은 RCT 기반 정책평가를 실시하고 RCT 관련 교육을 실무자에게 실시하여 과학적인 정책평가의 확산과 보급에 크게 기여하였다.

## Ⅳ. 우리나라의 RCT 현황과 적용 가능성 탐색

RCT는 우리나라의 여러 정책평가에서 활용될 수 있다. 우선적으로 RCT는 사전평가에서 활용될 수 있으며 현행 예비타당성 제도를 보완할 수 있다. 특히 정확한 사업의 편익을 추정할 수 없는 복지 및 교육사업에 시범사업 형태로 적용할 수 있다. 또한 사업시행 전에 실험방식의 정책평가설계가 준비된 경우 사후평가에서도 적용가능하며, 정책평가 후 사업의 폐지나 변경의 증거로 활용될 수 있다. RCT가 적용가능한 구체적인 정책분야로는 교육, 복지, 문화, 노동, 해외원조 사업 등이 있으며 SOC, 환경, 국방 등의 분야는 적용이 불가능하다. RCT가 여러 정책분야의 사전사후평가에서 활용가능할지라도 제도적, 윤리적 장애요인과 실험과정의 오류, 외적인 타당성의 부족 등으로 인하여 실제 정책평가에서 크게 활용되지 않고 있다.

## Ⅴ. RCT를 통한 실제 정책평가 설계 사례

RCT 또는 그와 유사한 준실험 평가방식으로 평가를 진행한 국내외 사례로는 희망리본사업과 미국 IES의 보충수업사례가 있다. 희망리본사업은 RCT

가 완벽히 적용된 사업은 아니지만 처지집단과 통계적으로 인구통계학적 차이가 없는 비교집단을 구성함으로써 비교한 준실험적 방식의 평가사례이다. 해외사례로서 엄격한 RCT 평가방법론을 준수하여 진행된 Institute of Education Science의 보충수업 정책실험이 있다.

## Ⅵ. RCT를 통한 재정사업 사전검증체계 도입방안

RCT를 실제 정책평가에 도입하기 위해서는 여러 장애요인을 극복해야 한다. 이러한 장애요인으로서 제도적인 요인이 큰데 이를 해결하기 위해서는 RCT에 대한 법적 근거를 마련하고 행정자료공유시스템과 전담조직을 설치해야 한다. 이런 제도적 장애요인이 해결되면 세밀한 RCT 수행체계에 대한 절차를 평가대상사업의 선정, 평가방법의 결정과 실행에 걸쳐 명확하게 규정해야 한다. 또한 RCT의 보급과 확산을 위해서는 담당공무원의 평가 관련 역량향상, 대국민 홍보를 통한 실험평가의 거부감 축소, 평가 전문인력 확보 등의 환경적 요인을 해결해야 한다.

CONTENTS

# 목 차

## CONTENTS

CONTENTS

# 표목차

# 그림목차

# Ⅰ. 서론: 연구의 배경과 목적

새로운 재정사업이 충분한 사전점검 없이 정치적인 요구에 의해 급격하게 도입되는 사례가 증가하면서 시행착오로 인한 사회적 비용이 증가하고 있다. 재원조달 가능성이나 소요액의 부정확한 추정에 근거한 문제도 부각되고 있으며, 사업계획의 부적정성으로 인해 예기치 않은 부작용이 노정되는 사례도 발생하고 있다. 대형 국책사업을 중심으로 대형 신규사업의 적정성을 심사하는 예비타당성조사 제도가 운영되고 있으나, 노동이나 복지 등 새롭게 확대되고 있는 사업영역에 대한 점검 기능은 만족할 만한 수준이 아니라고 판단된다. 고용이나 복지 관련 사업까지 예비타당성조사 제도를 확대하는 시도가 이루어지고 있으나, 기존의 예비타당성조사 제도의 한계 및 부작용이 부각되는 상황에서, 예비타당성조사 제도의 확대 시행의 실효성에 대한 의문이 존재한다.

현행 예비타당성조사 제도는 정형화된 사업이나 정책에 대해서는 비교적 의미 있게 점검이 가능하지만, 새롭게 추진하는 사업이나 정책의 쟁점이나 효과성을 점검하기는 어려운 방법이다. 비교적 정형화된 사업과 사업의 효과성을 추정하는 주요 파라미터에 대한 신뢰성 있는 추정치가 존재할 때만 적용 가능한 방법이라고 볼 수 있다. 현재 복지사업에 대해서도 예비타당성조사가 확대되고 있는 바, 사업 계획의 적정성 점검 수단으로는 의미가 있으나 실제 사업의 효과성이나 사업 운영의 주요 쟁점에 대한 점검 수단으로서는 한계가 있을 수밖에 없다.

예를 들어, 최근 시행예정인 저소득층 영아 기저귀 · 분유값 지원 사업[1]의 경우 예비타당성조사로는 정책의 효과를 정확히 예측하기 어렵다. 이 사업은 저소득층의 경제적 부담이 출산율을 낮추는 원인이기 때문에 양육에 필

1) 미국의 경우 'WIC(Woman, Infant and Child) Program'이란 명칭으로 활성화되어 있다.

요한 기저귀와 분유값을 지원하여 출산율을 높이려는 국정과제이다. 보건복지부는 사업의 정책효과를 시범사업을 통하여 검증하기 위하여 50억원의 시범사업예산을 요청했지만 중기재정소요 500억원 이상인 복지사업의 경우 예비타당성조사를 받아야 하는「국가재정법」규정으로 인하여 시범사업은 보류되고, 복지사업에 대한 예비타당성조사 후 사업시행이 결정되었다.

그러나 위 사업에 대한 예비타당성조사의 결과로 얼마나 신뢰성 있는 정책효과를 추정할 수 있을지 의문이다. 편익에 대한 파라미터가 존재하는 SOC 사업과 달리 사업시행의 결과로 얻어지는 출산율 증가의 편익을 예측하기가 불가능하기 때문이다. 사업 시행의 결과로 저소득층 가정에서 출산율이 증가하는지를 정확히 예측할 수 있는 방법은 소규모라도 실제 정책을 시행하는 것이 유일하다. 따라서 Randomized Control Trial(RCT)[2]를 통한 진실험적 평가방법은 정책효과를 예측하는 유용한 평가방법이 될 수 있다. 시범사업에서 일정 소득기준 이하의 정책대상자를 신청을 통해 받은 후 추첨을 통하여 처치집단과 비교집단으로 구분하고 처치집단에 기저귀와 분유를 지원한 후 실제 처치집단에서 출산율이 비교집단에 비해 얼마나 증가하는지를 비교함으로 정확한 정책효과를 추정할 수 있기 때문이다.

위 사업의 경우 예비타당성조사에서 통과되면 향후 장기간에 걸쳐 막대한 재원이 소요되는 현실을 감안할 때, 정확한 평가방법론에 의한 재정사업 사전검증체계의 제도화가 시급한 정책과제라 할 수 있다. 또한 과거 진행하였던 청년층 취업지원 사업, 대학 연구지원 사업, 중소기업 지원 사업 등 수많은 복지, 고용 관련 사업도 사업의 효과에 대한 정확한 사전 검증 없이 시작하여 큰 정책효과를 거두지 못하고 재원만 낭비하고 있는 실정이다.

따라서, 우리나라에서 최근 급속하게 확대되고 있는 복지나 고용 관련 사업의 경우, 사회실험 방식의 RCT 정책평가를 통해서 사업의 쟁점과 효과성을 파악하여 정책을 추진하는 것이 바람직하다. 외국에서도 고용이나 복지

2) 무작위 추출을 통해 동질적인 처치(실험)집단과 비교(통제)집단을 구성하여 정책의 효과를 평가하는 방법은 Randomized Trial Control(RCT) 외에, 처치-비교 집단 무작위 추출 실험, 사회실험(Social Experiment), 현장실험(Field Experiment), 정책실험(Policy Experiment) 등 다양한 용어로 사용되고 있다.

사업에 대해서는 RCT 평가방식을 적용하고 있으며, 정형화된 예비타당성조사 방식을 적용하는 경우는 찾아보기 어렵다. RCT를 적절하게 적용할 수 있는 재정사업에 대해서는 재정사업의 도입 이전에 사회실험을 거치도록 하는 방안을 마련할 필요가 있다. 본 보고서에서는 우리나라에서 RCT를 제도화하기 위한 기초적인 방안을 마련할 목적으로 조사 분석을 시도하고자 한다.

# Ⅱ. RCT(Randomized Control Trial) 방법론 소개

## 1. RCT의 개념

프로그램, 특히 정부가 수행하는 재정사업 프로그램은 평가(evaluation)를 통해 그 성과를 측정하고 더 나아가 프로그램 설계를 개선할 수 있다. 본 보고서에서 의미하는 평가란 프로그램(정책 혹은 프로젝트)이 효과적이었는지, 나아가 그러한 효과의 크기가 어느 정도인지 파악하는 일련의 과정을 의미한다. 프로그램은 그 도입단계에서부터 달성해야 할 목적을 지니고 있기 마련인데, 효과성 평가는 이러한 목적을 달성했는지 여부를 점검하는 것으로 프로그램 평가의 주요 목적이다. 더 나아가 그 효과의 크기를 추정하는 것은 추후 효율성 평가 등에 중요한 근거를 제시한다.

예를 들어, 국가가 수학과목 학업성취도가 낮은 학생들을 대상으로 운영하는 보충수업 프로그램을 생각해보자. 실제로 교육부가 2015년 2월에 발표한 '제2차 수학교육 종합계획(2015~2019)'과 2014년 12월에 발표한 '사교육 경감 및 공교육 정상화 대책'에 따르면 학습부진 학생들의 효율적 자기주도 학습을 지원하기 위해 맞춤형 수학멘토링을 운영할 계획이며, 이 프로그램을 통해 수학에 대한 흥미와 자신감을 부여하고 학습결손을 보정할 수 있기를 기대하고 있다. 이러한 보충수업 프로그램을 담당하고 있는 공무원은 해당 프로그램이 학업부진 학생들의 학업성취도(또는 시험성적) 향상이라는 목표를 달성했는지, 그렇다면 그 효과의 크기는(투입된 비용 대비) 어느 정도인지 알고 싶을 것이다. 이러한 질문이 중요한 이유는 그 결과에 따라 프로그램을 개선할 수 있으며 프로그램 연장 등으로 인해 추가 예산이 필요한 경우 중요한 근거 자료가 되기 때문이다.

재정사업 프로그램을 운영하는 공무원 입장에서 이렇게 당연한 질문에 답

하는 것은 일반적으로 쉬운 작업이 아니다. 위의 예에서 보충수업 프로그램의 효과성을 판단하기 위해서는 해당 프로그램의 수혜를 받은 집단과 그렇지 않은 집단의 성과를 비교해야만 한다. 하지만 보충수업에 참여하는 것은 전적으로 개인의 선택에 달려 있기 때문에 프로그램의 수혜를 받은 집단과 그렇지 않은 집단 사이의 비교는 주의를 필요로 한다. 만약 자기주도 학습 능력이 상대적으로 뛰어난 학생들이 주로 이 프로그램에 지원했다면, 우리가 관찰할 수 있는 두 집단 간 성과의 차이는 프로그램 본연의 성과와 프로그램에 지원하기 이전의 개인의 역량의 차이가 결합된 것이기 때문이다. 만약 후자의 크기가 크다면 이러한 집단 간의 단순 비교는 프로그램의 진정한 성과를 왜곡시킬 수 있다.

위의 예에서 살펴본 바와 같이 프로그램의 효과성을 측정하는 일은 프로그램의 성과를 이해하는 데 중요한 작업이지만 그 해답을 찾는 과정은 쉽지 않다. 프로그램의 수혜를 받은 집단과 그렇지 않은 집단을 단순하게 비교한 결과는 프로그램 성과와 두 집단 사이의 본연적인 성과의 차이가 포함되었기 때문이다. 더군다나 프로그램 설계를 위해 효과성을 프로그램 시행 이전에 측정하는 일은 더욱 어려운 작업이다.

이러한 측면에서 RCT(randomized controlled trials)는 프로그램 시행 이전에 프로그램의 성과를 엄밀히 평가할 수 있는 방안 중 하나로서 최근 해외 여러 국가에서 각광받고 있다. 본 보고서에서 의미하는 RCT란 프로그램의 처치집단(수혜집단)과 비교집단을 무작위 방식을 통해 구성함으로써 두 집단 간의 성과를 비교할 때에 집단 사이의 본연적인 성과의 차이가 없도록 설계한 프로그램 평가방법이다. 아래에서는 이러한 RCT가 어떤 이유로 최근 주목을 받고 있는지, 또 그 구체적인 방법론은 무엇인지 살펴본다.

## 2. 프로그램 평가와 RCT[3)]

### 가. 내생성(endogeneity)과 무작위처치(randomization)

RCT의 특징을 설명하기 위해서는 전반적인 프로그램 평가론에 대한 이해가 필요하다. 이를 위해 위에서 설명한 예를 활용하여 프로그램 평가론의 기본적인 구조를 살펴보자.

보충수업 프로그램 참여를 고려하고 있는 개인을 영어 소문자 $i$로 표현하고 그가 프로그램에 참여했을 경우의 성과(예를 들어, 시험성적)를 $Y_i^T$로, 참여하지 않았을 경우의 성과를 $Y_i^C$로 표현하자. $(Y_i^T, Y_i^C)$는 개인 $i$의 잠재적 성과(potential outcomes)로서 사전에 정해져 있으며 프로그램 참여 여부에 따라 둘 중 하나의 값이 개인 $i$의 실현된 성과(realized outcome) $Y_i$로 관측된다고 가정한다.

보충수업 프로그램이 개인 $i$에게 미치는 영향은 프로그램에 참여했을 경우의 성과 $Y_i^T$에서 참여하지 않았을 경우의 성과 $Y_i^C$를 차감해야 한다. 하지만 개인 $i$는 프로그램에 참여하거나 하지 않는, 두 가지 중 하나의 선택만을 할 수 있기 때문에 개인 $i$의 잠재적 성과 $(Y_i^T, Y_i^C)$를 동시에 관측할 수 없으며, 따라서 프로그램이 개인에게 미치는 영향은 물리적으로 계산할 수 없다.

이러한 문제는 프로그램 효과의 평균을 구할 때에도 똑같이 발생한다. 학습능력, 기존의 학습량 등 프로그램 참여 이전의 개인 간의 차이에 따라 보충수업 프로그램의 효과가 다르게 나타날 수 있기 때문에 사실 프로그램 담당자 입장에서 더욱 중요한 숫자는 프로그램 효과의 평균값 $E[Y_i^T - Y_i^C] = E[Y_i^T] - E[Y_i^C]$이다. 하지만 위와 같은 이유로 관측할 수 있는 데이터만을 가지고 이 평균값을 계산하는 것은 불가능하다. 그 이유를 살펴보기 위해 변수 $D_i$를 프로그램에 참여할 경우 1, 참여하지 않는 경우에는 0의 값을 갖는 더미변수로 정의하자. 그렇다면 프로그램에 참여한 사람

3) 본문의 내용은 Duflo, Glennerster and Kremer(2007)의 내용을 위주로 저자가 편집하였다.

들을 대상으로 계산한 평균값은 조건부 평균 $E[Y_i^T|D_i=1]$으로 모든 개인을 대상으로 프로그램 참여성과를 평균한 $E[Y_i^T]$와는 일반적으로 다른 값을 가진다. 마찬가지로 프로그램에 참여하지 않은 사람을 대상으로 계산한 평균값은 조건부 평균 $E[Y_i^C|D_i=0]$으로 $E[Y_i^C]$와는 일반적으로 다르다. 따라서 관측가능한 데이터만을 가지고 프로그램 효과의 평균값을 계산 또는 추정할 수 있는 방법은 없다.

$$E[Y_i^T|D_i=1]-E[Y_i^C|D_i=0]=(E[Y_i^T|D_i=1]-E[Y_i^C|D_i=1]) \\ +(E[Y_i^C|D_i=1]-E[Y_i^C|D_i=0])$$

그렇다면 관측가능한 데이터를 이용해 프로그램에 참여한 사람의 성과 평균에서 참여하지 않은 사람의 성과 평균을 차감한 값은 무엇일까? 먼저 그 차이를 아래와 같이 다시 써보자.

$$E[Y_i^T|D_i=1]-E[Y_i^C|D_i=1]=E[Y_i^T-Y_i^C|D_i=1]$$

이 표현에 따르면 우리가 계산할 수 있는 평균의 차이는 다음 두 가지 항목의 합과 같다. 첫 번째 항목은 프로그램에 참여한 사람만을 대상으로 측정한 프로그램의 효과로서 처치집단 대상 평균처치효과(Average treatment effects for the treated, ATT)라고 부른다.

두 번째 항목은 실제 프로그램에 참여한 사람들이 만약 참여하지 않았을 가상의 경우의 성과평균 $E[Y_i^C|D_i=1]$에서 실제 프로그램에 참여하지 않은 사람들의 성과평균 $E[Y_i^C|D_i=0]$을 차감한 값이다. 다시 말해 이 값은 프로그램에 참여한 집단과 그렇지 않은 집단 사이에 관찰되는, 프로그램 참여 이전의 본연적인 성과의 차이이다. 따라서 만약 실제 보충수업 프로그램에 참여한 학생들이 자기주도 학습능력이 상대적으로 뛰어나다면 이 두 번째 항목은 양의 값을 갖게 된다. 두 번째 항목은 경제학에서 선택편향(selection bias)라고 부르며 일반적으로 경우에 따라 양수, 음수 또는 0의 값을 가질 수 있다.

위의 결과로부터 우리는 선택편향이 존재하지 않는다면 프로그램에 참여한 사람의 성과평균에서 참여하지 않은 사람의 성과평균을 차감한 값은 프로그램에 참여한 사람들을 대상으로 한 프로그램 효과의 평균값이라는 사실을 알 수 있다. 더불어 만약 한 개인의 잠재적 성과와 다른 사람의 프로그램 참여 여부가 무관(unrelated)할 때(Stable Unit Treatment Value Assumption), 이 값은 우리가 알고자 하는 프로그램 효과의 평균값

$$E[Y_i^T - Y_i^C] = E[Y_i^T] - E[Y_i^C]$$

과 같다는 사실이 알려져 있다(Angrist, Imbens and Rubin, 1996).

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

또한, 회귀식을 최소자승법(ordinary least squares) 방법을 이용해 추정한 $\hat{\beta_1}$값은 위에서 설명한 프로그램에 참여한 사람의 성과평균에서 참여하지 않은 사람의 성과평균을 차감한 값과 같다.

$$\hat{\beta} = \hat{E}[Y_i^T | D_i = 1] - \hat{E}[Y_i^C | D_i = 0]$$

하지만 이 값이 ATT 또는 프로그램 효과의 평균값과 같다는 결과는 선택편향이 존재하지 않는다는 가정하에서만 성립한다. 따라서 프로그램의 성과를 정확히 측정하기 위해서는 프로그램에 실제 참여한 사람과 그렇지 않은 사람 사이의 본연의 성과 차이인 선택편향이 왜 발생하는지 알아야 한다. 기본적으로 그 이유는 두 집단 구성원 간의 특성이 균형을 이루지 못했기 때문이며, 이는 두 집단의 구분이 무작위적(random)으로 이루어지지 않았기 때문이다. 이를 경제학에서는 원인변수(cause, 위의 예에서는 프로그램 참여 여부)가 내생성(endogeneity)을 갖고 있다고, 또는 원인변수의 내생성을 통제하지 못했다고 이야기한다.[4)]

4) 원인변수 내생성의 대표적인 원인에 대해 강창희 외(2013)는 누락변수(omitted variables), 역의 인과관계(reverse causality), 자기선택(self-selection), 그리고 측정오차(measurement error)를 들어 설명하고 있다.

따라서 선택편향을 없애기 위한 방법은 프로그램 참여여부를 무작위적으로 선택해 프로그램에 참여한 처치집단과 그렇지 않은 비교집단 사이의 특성의 차이가(통계학적으로) 균형을 이루게 하는 것이다. 이러한 사실은 오랫동안 사회과학 연구자들 사이에 잘 알려져 있었지만 여러 가지 이유로 그동안 널리 사용되지 않았다. 그중 대표적인 이유는 연구목적으로 재정사업 프로그램 참여 여부를 무작위로 결정하는 것이 비윤리적일 수 있기 때문이다. 보충수업 프로그램을 통해 혜택을 받고자 하는 개인이 무작위처치의 결과에 따라 혜택을 받지 못한다면, 나아가 그 결과 개인의 학업 비용과 시험 또는 진학 결과에 영향을 줄 수 있다면 연구방법론에 대해 비판적인 의문이 제기될 수 있다.

이러한 이유로 사회과학, 특히 경제학에서는 그동안 준실험적(quasi- experiment) 방법이 널리 사용되어 왔다. 준실험적 방법이란 윤리적인 이유 등으로 무작위로 처치대상을 선별할 수 없어 실험데이터(experimental data)를 사용할 수 없는 경우에 계량적 기법을 바탕으로 관찰데이터(observational data)를 분석해 프로그램의 성과를 파악하는 방법이다.[5] 다음 절에서는 가장 널리 사용되고 있는 준실험적 방법을 소개한다.

### 나. 준실험적 방법

준실험적 방법은 현실 속에서 얻을 수 있는 관찰데이터를 사용하기 때문에 대부분의 경우 처치집단과 비교집단의 배정이 무작위로 이루어지지 않을 뿐 아니라, 비교집단이 존재하지 않는 경우도 있다. 이러한 관찰데이터상의 제약 속에서도 사회과학에서는 계량적 기법을 통해 적절한 비교집단을 설계해 프로그램의 효과를 분석하고 있다. 이러한 계량적 기법을 사용할 때에는 처치 여부가 무작위로 이루어지지 않았기 때문에 필연적으로 몇 가지 가정을 수반하게 되는데, 프로그램이나 그 배경, 제도 등이 이러한 가정을 만족시키는지 여부를 분석모형 내에서 판단할 수 없다. 따라서 계량적 기법의

5) 혹자는 준실험적 방법을 비실험적(non-experimental) 방법이라고 부르기도 한다.

분석을 가능하게 하는 가정이 적절한지에 대해 주관적으로 판단할 수밖에 없으며 가정이 적절하지 않은 경우에는 분석결과의 신뢰도를 크게 떨어트릴 수 있다.

(1) **전후비교**(Pre-post comparison)**과 단순평균비교**(Simple difference of means)

프로그램의 효과를 알기 위해 생각할 수 있는 가장 간단한 방법은 프로그램에 참여한 그룹을 대상으로 프로그램 참여 전후로 관심 있는 성과변수의 변화의 평균을 측정하는 것이다. 이 방법을 사용하기 위해서는 처치 전후로 측정한 프로그램 참여자의 성과변수에 관한 자료가 필요하다. 이 방법이 프로그램의 효과를 정확히 측정하기 위해서는 처치 전후로 해당 프로그램만이 성과변수에 영향을 줄 수 있다는 가정이 성립해야 한다. 하지만 이러한 가정이 성립하는 경우는 일반적으로 찾기 힘들며, 따라서 전후비교 방법은 제한적으로 사용해야 한다.

다른 간단한 효과성 측정방법으로 처치집단과 비교집단의 성과변수 평균을 비교하는 방법이다. 이 방법을 사용하기 위해서는 처치 이후에 처치집단과 비교집단의 성과변수를 측정한 자료가 필요하다. 하지만 앞서 설명하였듯이 이 방법으로 효과성을 정확히 측정하기 위해서는 선택편향이 존재하지 않아야 한다는 가정이 필요하다. 현대 프로그램 평가론에서는 일반적으로 내생성에 의한 선택편향이 존재한다고 믿기 때문에 이 방법 역시 제한적으로 사용해야 한다.

(2) **관찰가능변수 통제**(Controlling for observables) **방법과 성향점수 매칭기법**(Propensity score matching methods)

관찰가능변수 통제 방법은 데이터에서 대상의 특징을 설명하는 변수를 통제해 비교집단을 설계한다. 일반적으로 관찰데이터에는 개별대상이 처치를 받았는지 여부뿐만 아니라 그 대상의 특징을 설명하는 변수들이 존재한다.

예를 들어 보충수업 프로그램에 참여한 사람들과 그렇지 않은 사람들의 데이터에는 각 개인의 성별, 나이, 이전 시험성적 등이 함께 나와 있다. 이러한 변수들은 연구자가 데이터를 통해 확인할 수 있기 때문에 관찰가능변수(observables 또는 covariates)라고 부른다.[6] 관찰가능변수 통제 방법은 관찰가능변수 중에 일부 또는 전체의 값이 같은 사람들 사이의 처치 여부는 무작위로 이루어진다고 가정하는 방법이다. 이 가정을 바탕으로 적절한 관찰가능변수를 기준으로 삼아 처치집단 내의 대상과 비교집단 내의 대상을 매칭(matching)시키게 된다. 예를 들어, 보충수업 프로그램 데이터에서 프로그램에 참여한 15세 남성을 참여하지 않은 15세 남성과 매칭시키고 이 두 집단 내에서는 프로그램 참여 여부가 무작위로 이루어졌다고 가정하는 것이다. 이를 위에서 사용한 수식을 이용해 표현하면 다음과 같다. 어떤 관찰가능변수 벡터 $X$에 대해

$$E[Y_i^C|X, D_i = 1] = E[Y_i^C|X, D_i = 0]$$

다시 말해, 이는 관찰가능변수 $X$를 통제하였을 때 처치집단과 비교집단 사이에 선택편향은 존재하지 않는다는 가정이다. 그리고 전체 프로그램의 효과는 관찰가능변수 $X$를 각각 통제하여 구한 효과를 가중평균(weighted average)하여 구하게 된다.

이는 가장 간단한 준실험적 방법인 만큼 가장 강한 가정을 내포하고 있는 방법이다. 주어진 관찰가능변수에 대해 그 값이 같은 대상들 사이에 처치가 무작위로 이루어졌다고 주장하기 위해서는 이를 뒷받침할 수 있는 제도적 근거를 제시해야 한다. 하지만 연구자가 직접 설계해서 수행하는 RCT가 아닌 현실에서 이러한 조건이 성립한다고 주장하기는 일반적으로 쉽지 않다. 그 이유는 데이터상에 나와 있지 않은 관찰불가능변수가 처치 여부를 결정할 수도 있고, 연구자나 프로그램 담당자가 특별히 무작위로 프로그램을 설계하지 않는 이상, 처치 여부는 일반적으로 내생성을 띠고 있기 때문이다.

---

6) 반면, 개인의 성격 등 데이터상으로 표현할 수 없는 변수나, 표현가능하나 주어진 데이터에 나와 있지 않은 변수들은 관찰불가능변수(unobservables)라고 부른다.

이와 다른 가정을 전제하지만 비슷한 방법으로 성향점수 매칭기법이 있다. 관찰가능변수 통제 방법에서 관찰가능변수 벡터 $X$의 차원, 다시 말해 관찰가능변수의 수가 많거나 변수가 연속적인 값을 갖게 되면 많은 경우 처치집단과 비교집단에서 같은 $X$값을 갖는 대상을 찾는 것은 불가능하다. 이러한 경우 각 대상의 관찰가능변수를 바탕으로 처치를 받을 확률을 계산하고 이를 기준으로 처치집단과 비교집단 내의 대상을 서로 매칭시켜 프로그램의 효과를 계산할 수 있다. 여기서 관찰가능변수를 바탕으로 계산한 처치를 받을 확률을 성향점수(propensity score)라고 부르고 이를 기준으로 매칭시켜 프로그램의 효과를 분석하는 방법을 성향점수 매칭기법이라고 한다.[7)]

이 방법은 관찰가능변수 통제 방법과 비슷하지만 다른 가정을 전제로 하고 있다. 성향점수 매칭기법에서는 성향점수를 계산할 때 쓰이는 관찰가능변수를 통제하였을 때 처치를 받은 대상과 그렇지 않은 대상 사이에 잠재적 성과의 차이가 없어야 한다고 가정한다. 이를 만족하기 위해서는 처치집단과 비교집단 사이의 차이를 관찰가능변수가 모두 설명할 수 있어야 한다. 이러한 가정의 타당성은 분석모형 내에서 점검할 수 없으며 프로그램이나 제도를 바탕으로 주관적으로 판단해야 한다. 하지만 일반적으로 최근에 경제학에서는 누락변수 등을 이유로 이러한 가정에 근본적인 의문을 제기하고 있다. 다시 말해, 데이터상에 나타나 있지 않은 변수가 처치 여부를 결정하게 된다면 성향점수 매칭기법으로 추정한 프로그램의 효과는 그 신뢰성을 잃게 된다.

### (3) **회귀단절법**(Regression discontinuity)

회귀단절법의 기본 아이디어를 이해하기 위해 위에서 언급한 보충수업 프로그램의 예를 다음과 같이 조금 변형해 보자. 프로그램 담당자는 지원자가 너무 많은 관계로 간단한 테스트를 거쳐 70점 이상의 점수를 받은 지원자에게만 프로그램에 참여할 수 있게 했고 합격자 중에서 참여를 포기한 사람은

7) 강창희 외(2013)에 의하면, 성향점수를 어떤 기준으로 매칭시키느냐에 따라 최근거리 매칭법, 구간 매칭, 칼리퍼 매칭, 커널 매칭 등이 있다.

없다고 가정하자. 이 경우, 예를 들어, 95점을 받은 합격자와 30점을 받은 불합격자 사이에는 잠재적 성과가 기본적으로 다르다고 생각하는 것이 자연스럽다. 하지만 71점을 받은 합격자와 69점을 받은 불합격자 사이의 잠재적 성과의 차이는 그보다 작을 것이며 이 두 지원자 사이의 처치 여부는 무작위에 가깝다고 생각할 수 있다. 회귀단절법은 이와 같이 특정한 경계를 기준으로 처치 여부가 결정되는 경우 그 경계점 근방의 처치집단과 비교집단의 성과를 비교하는 방법이다. 수식으로 설명하자면, 특정한 관찰가능변수 $X$에 대해 경계점 $\overline{X}$를 기준으로 처치 여부가 결정된다면, 회귀단절법은 다음을 만족시키는 (작은) 양의 숫자 $\epsilon$이 있다고 가정한다.

$$E[Y_i^C|\overline{X}-\epsilon < X < \overline{X}+\epsilon, D_i = 1] = E[Y_i^C|\overline{X}-\epsilon < X < \overline{X}+\epsilon, D_i = 0]$$

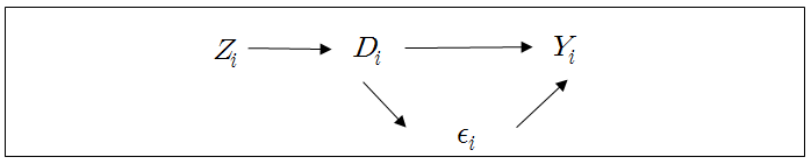
다시 말해, 회귀단절법은 경계점 $\overline{X}$를 근방으로 처치 여부는 무작위로 이루어졌다고 가정한다. 이러한 가정이 사실일 경우, 경계점 근방에서는 선택편향이 존재하지 않게 되며, 따라서 연구자는 경계점 근방의 처치집단과 비교집단의 성과를 비교하여 프로그램의 효과를 추정할 수 있다.

이는 앞서 설명한 관찰가능변수 통제 방법이나 성향점수 매칭법보다는 수긍할 수 있는 가정을 기반으로 하고 있어 최근 여러 연구에서 사용되고 있는 방법이지만, 그럼에도 불구하고 가정에 대해 몇 가지 문제점을 제기할 수 있다. 먼저 현실에서는 경계점을 기준으로 처치 여부를 엄격하게 결정하지 않을 수 있다는 것이다. 위의 예에서 프로그램 담당자가 정해진 규칙에도 불구하고 테스트 결과 69점을 받은 일부 지원자에게 참여를 허락한다면 회귀단절법으로 추정한 결과의 신뢰도는 떨어지게 된다. 또한 만약 프로그램 담당자가 테스트를 채점하는데 있어 어느 정도 주관을 반영할 수 있어서 일부 지원자에게 점수를 더 주거나 덜 줄 수 있다면 69점의 지원자와 71점의 지원자 사이에 내생성이 남아 있을 수 있다. 이러한 의문에 대해서는 연구자가 프로그램 설계나 제도 등을 바탕으로 설명해야 한다.

(4) **도구변수 추정법**(Instrument variables estimation)

도구변수 추정법은 처치여부 변수와 잠재적 성과 사이에 내생성이 존재할 때 처치여부 변수와는 관계가 있지만 성과변수와는 관계가 없는 도구변수(instrumental variables)를 이용해 효과성을 추정하는 방법이다. 여기서 도구변수의 의미를 이해하기 위해 다음 그림을 살펴보자.

[그림 II-1] 도구변수 추정법




출처: 저자 작성

처치변수 $D_i$가 내생적이라고 함은 처치변수 그 자체로 성과변수 $Y_i$에 영향을 주기도 하지만 다른 변수를 통해 간접적으로 영향을 주기도 함을 의미한다. 앞선 보충수업 프로그램의 예에서, 보충수업 프로그램은 학업성취도에 직접 영향을 준다. 하지만 만약 보충수업 프로그램에 참여하는 사람이 상대적으로 자기주도 학습능력이 뛰어나다면, 이는 보충수업 프로그램의 처치변수가 개인의 자기주도 학습능력(위의 그림에서는 $\epsilon_i$을 의미한다)을 통해 간접적으로 학업성취도에 영향을 주는 것으로 생각할 수 있다. 예를 들어, 보충수업 프로그램은 다른 모든 조건이 고정된 상태에서 학업성취도를 10만큼 늘려준다고 가정해 보자. 그리고 보충수업 프로그램에 실제 참여한 사람의 자기주도 학습능력은 상대적으로 뛰어나 학업성취도를 2만큼 높여준다고 가정하자. 그렇다면 실제 보충수업 프로그램의 효과는 10임에도 불구하고 처치집단과 비교집단 사이의 단순비교는 보충수업 프로그램이 학업성취도를 12만큼 늘려준다고 과대평가하게 된다.

이와 같이 처치변수에 내생성이 존재할 때 도구변수는 처치변수에만 직접 영향을 주고 성과변수에는 직접 영향을 주지 않는 변수이다(위 그림에서

$Z_i$). 도구변수는 처치변수에만 영향을 미치기 때문에 도구변수를 이용해 다른 모든 조건은 고정시킨 채 처치변수만을 조작할 수 있다. 따라서 도구변수가 존재한다면 우리가 찾고자 하는 처치효과를 추정할 수 있다.

이와 같은 내용을 정리하자면, 도구변수는 다음 세 가지 조건을 만족시켜야 한다. 먼저 도구변수는 성과변수의 오차항과 상관관계가 없어야 하며(외생성, exogeneity), 처치변수와는 상관관계가 있어야 한다(inclusion restriction). 마지막으로, 도구변수는 처치변수를 통하지 않고서는 성과변수에 영향을 미칠 수 없어야 한다(exclusion restriction).

하지만 현실에서 이와 같은 조건을 만족시키는 도구변수를 찾는다는 것은 쉽지 않다. 이상적인 도구변수가 있다고 하더라도 그것이 데이터에서 확인 불가능한 경우도 많다. 더군다나 위에서 언급한 도구변수의 세 가지 조건 중 첫 번째와 마지막 조건은 통계적으로 테스트하는 것이 불가능하기 때문에 도구변수가 필요조건을 만족시켰는지 입증하는 것도 쉬운 일이 아니다.

(5) **이중차분법**(Difference-in-differences methods)

이중차분법은 처치 이전에 처치집단과 비교집단 사이에 근본적인 차이가 있는 경우 이를 고려해 프로그램의 효과성을 분석하는 방법이다. 이해를 돕기 위해 위의 보충수업 프로그램의 예를 조금 변형해 살펴보자. 프로그램 담당자는 보충수업의 효과를 극대화하기 위해 평균성적이 70점대인 학생을 대상으로 할지, 아니면 평균성적이 60점대 이하인 학생을 대상으로 할지 고민 중이라고 가정하자. 어떤 집단을 대상으로 프로그램을 운영할지 결정하기 위해 평균성적에 따라 두 집단을 구성한 후 집단 사이의 특성이 균형을 이루도록 동전 던지기와 같은 무작위 방법으로 평균성적이 70점대인 집단을 처치집단으로 결정했다. 그리고 프로그램 운영 1년 후 두 집단의 성과의 차이를 측정했다. 이 경우 1년 후 성과의 차이는 프로그램의 효과를 보여주는 것일까?

만약 처치 이전에 두 집단 사이에 성과의 차이가 존재했다면 처치 이후에

관찰되는 성과의 차이는 프로그램의 효과와 처치 이전의 성과의 차이가 결합된 것이다. 따라서 이러한 경우 무작위처치 방법을 거쳤다고 하더라도 결과 해석에 주의를 필요로 한다.

이중차분법은 이렇게 처치 이전의 성과의 차이와 프로그램의 효과가 혼재해 있는 경우 프로그램의 효과를 분리해 내기 위해 한 가지 가정을 적용한다. 위의 예에서, 이중차분법은 비교집단인 평균성적이 60점대 이하인 학생들이 1년 후 보여주는 성과(학업성취도)의 변화량은 처치집단이 처치를 받지 않았다는 가상의 상황에서 1년 후 보여주는 성과의 변화량과 같다고 가정한다. 처치집단과 비교집단의 처치 이전과 이후의 성과의 차이는 관찰가능하기 때문에 처치집단에서 나타나는 추가적인 변화량은 프로그램의 효과로 해석가능하다.

이를 수식으로 나타내면 다음과 같다. 먼저 각 대상의 처치 이전의 잠재적 성과를 $Y_{i0}^{C}$라고 하고, 처치 이후 시점의 잠재적 성과를 $(Y_{i1}^{T}, Y_{i1}^{C})$라고 하자. 그렇다면 이중차분법의 가정은 아래와 같다.

$$E[Y_{i1}^{C} - Y_{i0}^{C} | D_i = 1] = E[Y_{i1}^{C} - Y_{i0}^{C} | D_i = 0]$$

다시 말해, 비교집단의 성과 변화량과 처치집단이 처치를 받지 않았을 가상의 경우 보이는 성과 변화량이 같다고 가정하는 것이다. 이 가정을 바탕으로 했을 때 이중차분법에 의한 프로그램의 효과는 다음과 같다.

$$DD = E[Y_{i1}^{T} - Y_{i0}^{C} | D_i = 1] - E[Y_{i1}^{C} - Y_{i0}^{C} | D_i = 0]$$

이중차분법은 비교적 자연스러운 가정을 바탕으로 하고 있으나 처치 이전의 데이터를 보유하고 있어야 사용할 수 있는 방법이다. 또한 만약 처치집단과 비교집단 사이의 처치 전후 성과 변화량이 같다는 가정이 타당하지 않은 경우라면 이중차분법을 적용해 얻은 결과는 정확성이 떨어지게 된다.

지금까지 살펴보았듯이 준실험적 방법은 관찰데이터의 한계를 극복하기 위해 각각 서로 다른 가정을 기반으로 하고 있다. 이러한 가정은 모형 내에서 그 타당성을 검증할 수 없기 때문에 연구자의 주관적인 판단을 필요로

하며, 타당성이 일반적으로 받아들여지지 않는 경우 준실험적 방법에 의한 추정 결과는 설득력을 잃게 된다. 또한 데이터가 충분치 않아 준실험적 방법도 사용하지 못하는 경우에는 더 엄격한 가정을 적용해 효과성을 측정해야 하는데 이러한 가정이 실제 충족되는지 입증하는 일은 쉽지 않은 작업이다. 이러한 한계점을 근본적으로 해결할 수 있는 방법은 처치 여부를 무작위로 결정하는 것이며, 따라서 최근 무작위처치를 핵심으로 하는 RCT 방법이 주목을 받고 있다.

## 다. RCT 방법론과 그 특징

이제 보충수업 프로그램 담당자는 RCT만이 프로그램의 효과성을 가장 정확하게 측정하는 방법이라는 사실을 알고 이를 시행하려고 한다. 하지만 구체적으로 어떻게 무작위처치를 수행하는지, 또 그 결과를 어떻게 해석해야 하는지 알아야 프로그램의 효과성을 정확하게 판단할 수 있다. 이 절에서는 이에 대한 내용을 다룬다.

### (1) 통계학적 가설 검정의 개요

본격적으로 RCT의 방법론을 살펴보기 위해서는 통계학적 가설 검정의 내용을 이해할 필요가 있다. 통계학적 가설 검정이란 모집단의 특성에 대한 서술인 통계적 가설을 모집단에서 추출한 표본을 이용하여 가설의 타당성을 추정하는 작업이다. 여기서는 기본적인 통계학적 가설 검정의 내용은 생략하고 RCT에 필요한 부분만을 간단히 정리한다.

보충수업 프로그램의 예를 계속해서 살펴보자. 프로그램 담당자는 해당 프로그램이 시험성적에 영향을 미치는지, 그렇다면 그 크기는 어느 정도인지 알고자 한다. 이를 위해 먼저 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)을 설정해야 한다. 이 예에서 귀무가설은 '$H_0$ :보충수업 프로그램이 시험성적에 미치는 영향은 없다'로 설정했고, 대립가설은 '$H_1$ :보충수업 프로그램은 시험성적을 향상시킨다'로 설정했다고 하자.

RCT 방식에 따르면 프로그램 담당자는 귀무가설을 채택 또는 기각하기 위해서 무작위처치를 바탕으로 처치집단과 비교집단을 구성해야 한다. 이를 위해 먼저 프로그램 담당자는 보충수업 프로그램의 잠재적 수혜 대상자 중에 $N$명의 학생을 무작위로 선정했다고 하자. 논의의 편의상 학생 전체의 시험성적 확률변수는 $\sigma^2$를 분산으로 갖는 분포를 따른다고 가정한다. 각 개인의 프로그램 참여 여부는 무작위처치를 통해 이루어지는데 전체 $N$명의 학생 중에 무작위처치를 거쳐 프로그램에 참여한 학생 수의 비율을 $P$라고 하자.[8] 계획된 보충수업을 처치집단을 대상으로 진행한 후 프로그램 종료 직후 처치집단과 비교집단을 대상으로 공통의 시험 문항을 출제한 후 그 성적 결과를 기록한다.

이러한 과정을 거친 후에 프로그램 담당자는 각 학생의 프로그램 참여 여부($D_i$)와 시험성적($Y_i$)에 대한 자료를 얻을 수 있다.[9] 그렇다면 앞서 설명하였듯이 프로그램의 효과성은 최소자승법을 이용해 다음의 회귀식에서 처치변수의 계수 $\beta_1$을 추정함으로써 얻을 수 있다.[10]

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

여기서 $\beta_1$의 추정값으로 $\widehat{\beta_1} = 1$을 얻었다고 가정해보자. 이 결과는 프로그램의 효과가 있음을 보여주는 것일까? 아니면 프로그램의 실제 효과는 없음에도 불구하고 표본 추출 등의 원인에 의해 양의 추정값이 나온 것일까? 이 질문에 답을 할 수 있다면 우리는 프로그램의 효과가 없다는 귀무가설을 채택 또는 기각할 수 있게 된다.

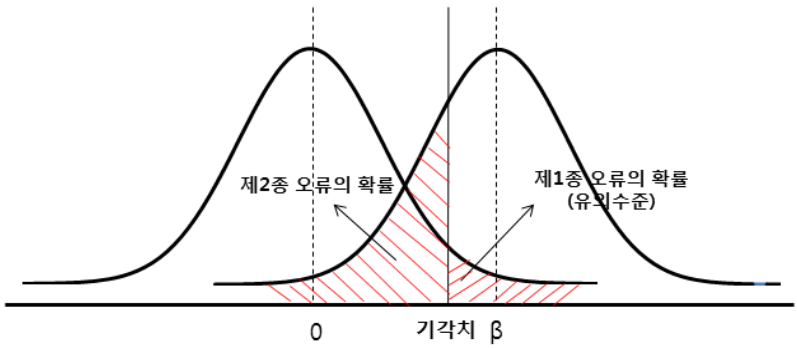
---

8) 구체적인 무작위처치 방법에 관한 내용은 다음 절에서 다루도록 한다.

9) 또한, 학생의 성별, 나이, 이전 시험성적 등에 관한 자료를 얻을 수 있으며 이를 회귀식에 포함시킬 수 있다.

10) 따라서 '보충수업 프로그램이 시험성적에 미치는 영향은 없다'는 귀무가설은 $H_0 : \beta_1 = 0$로, '보충수업 프로그램은 시험성적을 향상시킨다'는 대립가설은 $H_1 : \beta_1 > 0$로 표현가능하다.

[그림 Ⅱ-2] 통계학적 가설 검정




출처: 저자 작성

일반적으로 귀무가설이 참이면 [그림 Ⅱ-2]와 같이 $\beta_1$의 추정값인 $\hat{\beta_1}$의 분포는 0을 중심으로 종모양을 이루게 되고, 반대로 귀무가설이 거짓이고 프로그램이 양의 효과성을 갖는다는 대립가설이 참이라면 이보다는 오른쪽으로 평행이동해 종모양의 분포를 이루게 된다. 여기서 분포 그래프는 $\beta_1$의 추정값 $\hat{\beta_1}$이 가로축의 값을 갖는 확률을 세로축에 표현한 것으로 종모양이 중심축을 기준으로 더 뾰족한 모양을 갖게 되면 $\beta_1$의 참값에 가까운 추정치를 얻는 확률이 높아진다. 이렇게 $\hat{\beta_1}$의 종모양의 분포가 중심축으로부터 어느 정도 넓게 분포해 있는가를 표현하기 위한 개념이 $\hat{\beta_1}$의 표준오차(standard error)로, 최소자승법을 이용해 얻게 되는 $\hat{\beta_1}$의 표준오차는 다음과 같다.

$$SE(\hat{\beta_1}) = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

표준오차가 클수록 평균으로부터 넓게 분포하여 있으므로, 보다 정확한 $\beta_1$의 추정값을 얻기 위해서는 일반적으로 표본 수($N$)가 커야 하고, 처치집단과 비교집단의 구성원 수가 같아야 하며($P=1/2$), 모집단의 분산($\sigma^2$)이 작아야 한다. 이는 직관적으로도 자명하다. 표본 수가 클수록 실험결과로

나온 추정값을 더욱 신뢰할 수 있을 것이다. 또한, 이해를 돕기 위해, 극단적으로 모집단의 분산이 0이라면, 이는 모집단의 모든 학생이 처치 이전에 동일한 시험성적을 갖고 있다는 의미이므로 실험을 통해 프로그램의 효과를 바로 확인할 수 있다. 그리고 처치집단이나 비교집단의 구성원 수가 0이라면 프로그램의 효과를 확인할 수 없다.

[그림 Ⅱ-2]에서 $\beta_1$의 추정값 $\hat{\beta_1}=1$이 왼쪽 종모양의 그래프에서 나왔다면(다시 말해, $\hat{\beta_1}$이 충분히 작다면 또는 귀무가설이 참이라면), 프로그램 담당자는 귀무가설을 채택하고 프로그램의 효과가 없다고 결론지어야 한다. 반대로 $\beta_1$의 추정값 $\hat{\beta_1}=1$이 오른쪽 종모양의 그래프에서 나왔다면(다시 말해, $\hat{\beta_1}$이 충분히 크다면 또는 귀무가설이 거짓이라면), 귀무가설을 기각하고 프로그램의 효과가 있다는 대립가설을 채택해야 한다. 이러한 경우 일반적으로 통계학적 가설 검정에서는 기각치(critical value)라고 하는 기준값을 설정해 추정치가 기각치보다 작으면 귀무가설을 채택하고, 그렇지 않은 경우 귀무가설을 기각하게 된다. 따라서 기각치의 크기에 따라 통계학적 가설 검정에서 생길 수 있는 경우의 수는 다음 〈표 Ⅱ-1〉과 같다.

〈표 Ⅱ-1〉 통계학적 가설 검정에서 경우의 수

| | 귀무가설이 참일 때 | 귀무가설이 거짓일 때 |
|---|---|---|
| 귀무가설 채택 | 문제 없음 | 제2종 오류<br>(Type Ⅱ error) |
| 귀무가설 기각 | 제1종 오류(Type Ⅰ error) | 문제 없음 |

출처: 저자 작성

귀무가설이 참일 때 채택하거나, 거짓일 때 기각하는 것은 올바른 결론이다. 하지만 기각치의 값에 따라 귀무가설이 참임에도 기각하거나, 귀무가설이 거짓임에도 채택하는 오류를 범할 수도 있다. 통계학적 가설 검정에서 귀무가설이 참임에도 불구하고 기각하는 오류를 제1종 오류(Type I error)라고 하고 그 확률을 유의수준(significance level)이라고 한다. 우리의 보충수

업 프로그램의 예에서 유의수준은 위의 [그림 Ⅱ-2]에 표현한 것과 같이 왼쪽 종모양 분포에서 기각치보다 큰 영역의 넓이이다. 반면 귀무가설이 거짓임에도 채택하는 경우를 제2종 오류(Type II error)라고 하며, 그 확률은 오른쪽 종모양 분포 그래프에서 기각치보다 작은 쪽 영역의 넓이이다.

위의 [그림 Ⅱ-2]에서 확인할 수 있듯이 귀무가설과 대립가설이 주어진 상태에서 제1종 오류와 제2종 오류는 서로 대립되는 관계에 있다. 즉 기각치가 커진다면 제1종 오류를 범할 확률은 줄어들지만, 제2종 오류를 범할 확률은 높아진다. 반대로, 기각치가 작아진다면 제2종 오류를 범할 확률은 낮아지지만 제1종 오류를 범할 확률은 높아진다. 따라서 통계학적 가설 검정에서 적절한 기각치 설정은 중요한 문제이다.

먼저 제1종 오류의 확률, 즉 유의수준만을 고려해 기각치를 설정할 수 있다. 유의수준이 $\alpha$로 주어진다면 $t$분포로부터 $t_\alpha$값을 얻어 기각치를 다음과 같이 찾을 수 있다.

$$CV = t_\alpha \bullet SE(\hat{\beta})$$

또한 제1종 오류와 제2종 오류의 확률을 동시에 고려해 기각치를 설정할 수도 있다. 이를 위해 먼저 통계학적 가설 검정에서 중요한 개념 중 하나인 검정력(power of test)에 대한 이해가 필요하다. 검정력이란 귀무가설이 거짓일 때 귀무가설을 기각할 확률로서, 프로그램 평가에서는 프로그램의 효과가 있을 때 이를 잡아낼 수 있는 확률로 해석할 수 있다. 그러므로 검정력은 1에서 제2종 오류의 확률을 뺀 값과 같다.

만약 대립가설하에서 $\beta_1$의 값이 $\beta_1^* > 0$이고 유의수준 $\alpha$가 주어진다면, 대립가설하에서의 $\hat{\beta_1}$의 분포와 기각치를 알 수 있기 때문에 그 실험의 검정력을 계산할 수 있게 된다. 그리고 유의수준이 주어져 기각치가 설정되어 있는 상태에서는 대립가설에서 $\beta$의 값 $\beta_1^*$이 커질수록 실험의 검정력은 커지게 된다. 그러므로 유의수준 $\alpha$가 주어진 상태에서 검정력이 적어도 $\kappa$값을 갖기 위해서는 $\beta_1^*$은 다음 조건을 만족해야 한다.

$$\beta_1^* > (t_{1-\kappa} + t_\alpha) SE(\hat{\beta})$$

$$\begin{aligned} MDE &= (t_{1-\kappa} + t_\alpha) SE(\hat{\beta}) \\ &= (t_{1-\kappa} + t_\alpha) \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}} \end{aligned}$$

이는 곧 유의수준 $\alpha$와 검정력 $\kappa$를 만족시키는 최소한의 프로그램 효과의 크기는 $(t_{1-\kappa} + t_\alpha) SE(\hat{\beta})$라는 의미이며, 이를 최소확인가능효과(minimum detectable effect size, MDE)라고 한다.

최소확인가능효과는 주어진 유의수준 $\alpha$와 검정력 $\kappa$를 만족시키는 최소한의 기각치로 해석할 수 있다. 따라서 최소확인가능효과가 상당히 큰 값을 갖는다면, 프로그램의 실제 효과가 양수임에도 불구하고 우리는 프로그램의 효과가 없다는 귀무가설을 채택할 확률이 높아진다(즉, 주어진 검정력을 만족시키지 못할 가능성이 높다). 반면 최소확인가능효과가 상당히 작은 값을 갖는다면, 우리는 추정값이 작더라도 귀무가설을 기각하고 실험 결과로 나온 추정값을 프로그램의 효과로 받아들일 수 있게 된다. 따라서 RCT에서 최소확인가능효과를 최대한 작게 잡는 것이 중요하다. 이를 위해서는 앞서 설명했듯이 일반적으로 표본의 크기 $N$이 커야 하며 처치집단과 비교집단의 크기가 같아야 한다.

이제 이 결과를 바탕으로 RCT 설계에 대한 내용을 살펴보자.

### (2) 무작위처치 방법

RCT에서는 처치집단과 비교집단을 무작위로 결정하는 몇 가지 서로 다른 방법이 있다. 먼저 지원자에 비해 프로그램 운영에 필요한 자원의 양이 적을 때에는 가장 자연스러운 무작위 방법인 초과신청방법(oversubscription method)을 사용할 수 있다. 공급에 비해 수요가 많기 때문에 지원자를 대상으로 무작위 제비를 뽑게 해 처치집단을 결정할 수 있다. 이 방법의 단점은 실험의 대상으로 참여하는 사람들이 비교집단에 속해 있을 때 프로그램 수

혜를 받지 못해 실험에서 이탈할 가능성이 있다는 것이다.

이러한 단점을 극복할 수 있는 방법으로 무작위단계적도입(randomized order of phase-in)이 있다. 이 방법에 따르면 모든 실험 참가자가 처치를 받게 되지만 그 순서는 무작위로 정해진다. 따라서 한 시점에 비교집단에 있는 사람도 미래에 프로그램 수혜를 받을 수 있기 때문에 연구자와 꾸준히 협조관계를 유지할 유인이 있다. 하지만 이 방법은 프로그램의 장기효과를 고려할 때나, 프로그램 효과가 나타나는 시간이 길 것으로 예상되는 경우에는 사용이 제한된다. 또한 비교집단에 있는 사람들이 미래 처치집단에 속할 것을 기대해 행동에 변화가 생긴다면 정확한 프로그램의 효과 추정이 어려울 수 있다.

또 다른 무작위처치 방법으로 집단내무작위처치(within-group randomization)가 있다. 이 방법은 표본을 몇 개의 그룹으로 나눈 후 각 그룹 내에서 무작위처치를 시행하는 것이다. 인도의 발사키(Balsakhi)라고 불리우는 개인교사 지원 프로그램은 이 방법을 통해 무작위처치를 시행했다. 이 프로그램은 평가를 위해 한 지역 내의 학교를 대상으로 처치집단과 비교집단을 구성하는 방법을 사용하지 않고, 각 학교 내에서 처치집단과 비교집단을 구성했다. 예를 들어 한 학교에서는 3학년 학생에게만 개인교사를 지원하고 다른 학교에서는 4학년 학생에게만 지원했다. 그 결과 연구진은 학년별로 처치집단과 비교집단을 구성할 수 있었다. 이는 모든 학교가 지원을 받을 수 있기 때문에 학교 운영자들로부터의 저항도 심하지 않았다. 반면 이 예에서 집단내무작위처치 방법은 개인교사 배정에 따라 학교가 학년별로 학교 내 자원을 이용한 교육 지원을 다르게 할 유인이 생길 수 있어 설계에 있어 주의를 필요로 한다. 예를 들어 한 학교에서 3학년 학생들을 대상으로 개인교사 지원이 결정되었다고 한다면, 이 학교는 3학년보다는 다른 학년을 대상으로 교육지원을 늘릴 유인이 생긴다.

마지막 무작위처치 방법으로 참여권장방법(encouragement designs)이 있다. 이 방법에 따르면 연구자는 각 실험 참여자에게 처치 또는 비처치를 결정하지 않고, 대신 무작위로 선정된 실험 참여자에게 프로그램에 지원할 것

을 권유하게 된다. 즉, 모든 실험 참여자는 프로그램에 지원할 수 있지만 무작위로 선정된 참여자만이 프로그램 참여를 권유받는다. 이 방법은 프로그램 지원이 모든 사람에게 가능하지만 무작위처치가 윤리적 이유 등으로 적절하지 않을 때 사용가능하다. 단지 각 실험참여자는 프로그램 수혜를 받을 확률이 0과 1 사이의 수이기 때문에 실험결과를 분석할 때 추가적인 고려를 필요로 한다.

### (3) RCT 설계 시 주의할 사항

#### 가) 무작위처치 대상 레벨 설정

RCT를 설계할 때 연구자 또는 프로그램 담당자가 직면하는 문제 중 하나는 무작위처치의 대상레벨을 선정하는 것이다. 무작위처치의 대상은 개인이 될 수도 있고, 가족, 학교, 지역 등 특정 그룹이 될 수도 있다. 가장 자연스러운 선택은 프로그램의 수혜 대상을 처치 대상레벨로 선정하는 것이다.

하지만 개인 또는 그룹이 처치 대상레벨로 선정가능하다면 연구자는 여러 가지 사항을 고려해 결정을 내려야 한다. 가장 우선 고려해야 하는 것은 실험의 검정력이다. 만약 처치 대상레벨을 그룹으로 설정할 경우 그룹 내 실험참여자는 프로그램 처치 여부 이외에 다른 이벤트로 인해 성과변수와 같은 영향을 받을 가능성이 있다. 이러한 경우 프로그램 효과성을 측정하기 위해 다음의 회귀식을 생각해보자.

$$Y_{ij} = \alpha + \beta D_i + v_j + w_{ij}$$

여기서 아래첨자 $j$는 개인 $i$가 속한 그룹을 의미하고 $v_j$는 각 그룹 내의 공통충격항(common shock), $w_{ij}$는 개별충격항이다. 논의의 편의상 $J$개의 그룹이 있고 각 그룹 내 구성원 수는 $n$이라고 가정하며, $v_j$의 분산은 $\tau^2$, $w_{ij}$의 분산은 $\sigma^2$라고 하자. 이 경우 그룹을 대상으로 무작위처치를 했을 때 추정값 $\widehat{\beta_1}$의 표준오차는 $\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2+\sigma^2}{nJ}}$ 이다. 반면, 개인을

대상으로 무작위처치를 한 경우 $\widehat{\beta_1}$의 표준오차는 $\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}$ 이다. 따라서 개인을 대상으로 무작위처치를 했을 경우 $\widehat{\beta_1}$의 표준오차 대비 그룹 대상 무작위처치의 경우 $\widehat{\beta_1}$의 표준오차의 비는 $\sqrt{1+(n-1)\rho}$와 같다($\rho=\tau^2/(\tau^2+\sigma^2)$). 이로부터 그룹 내 구성원의 숫자가 많을수록 그룹을 대상으로 무작위처치를 했을 때 $\widehat{\beta_1}$의 표준오차가 증가하는 것을 알 수 있다.

또한 그룹을 대상으로 무작위처치를 실시했을 때 최소확인가능효과는

$$MDE_g = \frac{t_\alpha + t_{1-\kappa}}{\sqrt{P(1-P)J}}\sqrt{\rho+\frac{1-\rho}{n}\sigma}$$

이다. 이 식에서 알 수 있듯이 그룹을 대상으로 했을 때 최소확인가능효과는 그룹의 수 $J$나 그룹 내 구성원의 수 $n$이 증가할수록 작아진다. 하지만 $n$보다는 $J$를 늘리는 것이 최소확인가능효과를 더 빠르게 감소시킨다. 따라서 효율적인 실험 검정을 위해서는 그룹 내 구성원의 숫자보다는 그룹의 숫자를 증가시키는 것이 적절하다.

무작위처치 대상레벨을 결정할 때에는 검정력뿐만 아니라 전이 또는 오염(spillover or contamination)의 가능성도 고려해야 한다. 예를 들어 보충수업 프로그램의 예에서 프로그램 참여자가 비참여자와 교육 내용을 공유할 수 있다면, 지역 또는 학교 단위로 무작위처치를 시행하는 것이 적절하다.

#### 나) 비협조적 참여(Imperfect compliance)

RCT는 사람들을 대상으로 하는 실험이기 때문에 처치를 시행했다고 하더라도 개인의 선택에 따라 처치 적용이 완벽하지 않을 수 있다. 보충수업 프로그램의 예에서 보충수업을 받기로 한 학생이 개인의 선택에 의해 보충수업에 참여하지 않을 수도 있고, 반면 보충수업을 받지 않기로 한 학생이 보충수업에 참여할 수도 있다. 프로그램 담당자가 이를 강력하게 규제하지 못하는 상황이라면 기존의 RCT 설계와는 다른 방식으로 평가가 진행될 수 있다. 이와 같은 비협조적 참여는 실험의 검정력에 영향을 미치기 때문에 실

험 설계 단계에서 이를 고려해 표본 크기 등을 설정해야 할 필요가 있다.

비협조적 참여가 검정력에 미치는 영향을 알아보기 위해, 처치를 받기로 한 참여자가 실제 처치를 받을 확률을 $c$, 처치를 받지 않기로 한 개인이 처치를 받을 확률을 $s$라고 하자. 그렇다면 이 경우 최소확인가능효과는 다음과 같다

$$MDE_{ic} = (t_{1-\kappa} + t_{\alpha}) \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}} \frac{1}{c-s}$$

여기서 $c-s$는 0과 1 사이의 숫자이기 때문에 비협조적 참여는 최소확인가능효과를 크게 만든다. 따라서 비협조적 참여가 우려되는 상황이라면 이를 고려해 표본 수를 늘려야 한다. 또한 연구자는 실험참여자의 협조 정도가 서로 다른 설계 방법이 있을 경우에는 참여가 더욱 협조적일 것으로 기대되는 설계 방법을 택하는 것이 현명하다.

다) 관찰가능변수 통제(Control variables)와 표본 계층화(Stratification)

무작위처치를 기반으로 하는 RCT에서는 관심 있는 성과변수에 영향을 줄 수 있는 변수들을 회귀식에 포함시키는 것이 $\beta_1$의 추정값을 변화시키지 않는다. 하지만 이러한 변수들을 포함시키게 되면 $\beta_1$의 표준오차를 줄일 수 있게 되고, 따라서 상대적으로 작은 표본으로도 원하는 효과성 추정을 실행할 수 있다.

여기서 주의할 점은 처치 이후에 처치로 인해 영향을 받은 변수를 포함시키게 되면 이러한 변수가 추정하고자 하는 효과성의 일부를 흡수하게 되므로 $\beta_1$의 추정값에 영향을 줄 수 있다. 또한 성과변수와 관계가 없거나 영향을 거의 주지 않는 변수를 포함시키게 되면 오히려 표준오차를 증가시킬 수 있다. 따라서 관찰가능변수 통제는 처치 이전에 충분한 사전검토를 바탕으로 이루어져야 한다.

RCT를 수행할 때 관찰가능변수 통제와 함께 고려해봐야 할 것이 표본 계층화이다. 표본 계층화란 RCT를 위해 수집한 표본을 중요한 관찰가능변수

에 따라 소그룹으로 구분하는 것이다. 보충수업 프로그램의 예에서 프로그램 담당자는 학생의 성별, 나이, 거주 지역, 이전 성적 등에 따라 비슷한 학생들의 소그룹으로 구분해 RCT를 수행할 수 있다.

이러한 표본 계층화는 관찰가능변수 통제와 같이, 계층화에 쓰인 관찰가능변수가 성과변수에 영향을 주는 한 $\beta_1$의 표준오차를 줄일 수 있다. 또한 표본 계층화는 실제로 처치집단과 비교집단의 차이를 줄일 수 있다. 무작위 처치는 이론적으로 처치집단과 비교집단의 차이를 줄일 수 있지만, 극단적인 경우 효과성 추정을 어렵게 할 수도 있다. 예를 들어, 보충수업 프로그램에서 무작위처치 결과 특정 거주 지역의 학생들만 프로그램의 수혜를 받는다면 이로부터 추정된 효과성은 그 지역의 교육기반에 의한 것인지, 아니면 보충수업에 의한 것인지 판단하기 어려워진다. 따라서 처치 이전에 중요한 관찰변수에 따라 RCT 대상자를 소그룹으로 분류하고 각 소그룹에서 무작위 처치를 시행한다면 실제로 처치집단과 비교집단의 차이를 줄일 수 있고, 더욱 정확한 결과를 얻을 수 있다. 이 경우 전체 효과성은 각 소그룹에서 추정한 효과성을 가중평균해서 구할 수 있다.

# Ⅲ. RCT의 국제적 동향

대규모 재정이 투입되는 사업에 있어서 정확한 성과평가의 중요성이 부각됨에 따라 주요국들은 보다 과학적인 평가를 통해 재정사업의 효과를 예측하기 위한 노력을 해왔다. 이러한 무작위 추출 통제 실험평가(Randomized Control Trial; RCT)를 통하여 정책의 효과를 정확하게 평가하고 여기서 생성된 평가정보를 사업종료(Sunset)나 확대(Scale-up)를 위한 의사결정에 활용하기 위한 시도로 구체화되었다. 특히 영국과 미국은 RCT를 통하여 정책평가를 실시하는 전담조직을 설치하였다. 영국의 Behavioral Insight Team (BIT; 일명 Nudge Unit)과 미국의 오바마 행정부의 Evidence Based Policy Initiative를 근거로 최근 설치된 Social Behavioral Science Team(SBST)가 대표적인 예이다.

## 1. 영국의 Behavioral Insight Team[11)]

### 가. 설립연혁

영국의 Behavioral Insight Team은 세계 최초로 무작위 추출 통제 실험평가 방법(RCT)을 통해 정부정책을 평가하기 위해 2010년 설치되었다. Behavioral Insight Team은 정책을 평가하기 위한 방법으로 실험적 방법론을 활용하고 있다. BIT에서 RCT와 같은 실험적 방법론을 사용하는 이유는 팀명에서 보듯이 정책을 집행함으로써 인간의 행동이 변화하고 이러한 행동변화가 정책의 성과를 결정짓는다는 믿음에 근거하고 있다. 예를 들어, 학생의 학력

11) Behavioral Insight Team홈페이지 내용 정리,
http://www.behaviouralinsights.co.uk, 검색일자 2015. 9. 15.

향상을 위한 학습행동을 변화시키는 교육정책처치(Intervention)의 효과를 검증하는 수단으로 정책처치를 받은 학생그룹과 처치를 받지 않는 그룹 간의 학습행동변화 차이를 비교함으로써 정책의 효과성을 분석한다. BIT는 보수당과 노동당의 연립내각인 캐머런 정부의 내각사무처 산하의 직속 정부조직으로 2010년 설치되었으나 2014년 부분적으로 민영화되어 사회적 기업이 되었다. 현재, BIT는 사회적 기업으로서 내각사무처, 우리사주, 사회적 기업인 'Nesta'가 각각 1/3씩 소유지분을 나누어 갖고 있다.

### 나. 인력 및 조직현황

BIT에는 현재 56명의 정책분석가가 근무하고 있으며 이중 7명은 박사급 분석가이다. 정책분석가의 전공은 경제학, 정치학, 심리학 등 다양한 분야의 사회과학 학문영역을 포괄하고 있다. BIT의 분석가들은 행동과학 및 RCT에 대한 전문가 집단이며 내각사무처로부터 의뢰받은 다양한 정부사업에 대한 사회실험을 수행하고 있다. BIT는 현재 런던에 본부를 두고 있으며 뉴욕과 시드니에 지사를 두고 있다. BIT는 대부분의 정책영역에서 실험평가를 진행하고 있으며 주요평가분야는 〈표 Ⅲ-1〉과 같다. BIT는 30개 이상의 정부기관, 공공기관, 국제기구 등과 RCT 정책평가를 위한 파트너십을 구축하고, 매년 이틀 일정의 국제학술회의(International Behavioral Conference)를 개최하고 있다. 학술회의에는 행동과학 및 RCT 관련 대학, 연구기관, 정책분석가, 평가관리자들이 참여하여 실험평가 방법론과 사례에 대해 발표하고 토론한다.

〈표 Ⅲ-1〉 BIT의 주요 정책평가 분야

| |
|---|
| 소비자 및 금융서비스(Consumer & Financial Services) |
| 교육 및 개발(Education & Skills) |
| 에너지 및 지속성(Energy & Sustainability) |
| 보건(Health) |
| 내무 및 안전(Home Affairs & Security) |
| 국제사업(International Programmes) |
| 노동 및 경제개발(Labour Markets & Economic Growth) |
| 운영(Operations) |
| 연구 및 평가(Research & Evaluation) |
| 세제(Tax) |

출처: BIT 홈페이지, http://www.behaviouralinsights.co.uk/people, 검색일자 2015. 9. 15.

### 다. 운영성과

BIT는 2010년 이후 현재까지 150개의 정부사업을 RCT를 이용하여 평가하여 왔다. 영국의 공영방송인 BBC의 보도에 의하면 BIT에 의한 평가로 약 30억 파운드의 예산을 절감한 것으로 파악되었다.[12] BIT팀에 의해 가장 성공적으로 수행된 것으로 평가받는 정책평가는 장기기증 신청 홈페이지 메시지 변화에 따른 기증자의 신청률 변화 실험과 납부고지방식의 변화에 따른 체납납부 실적 변화 실험이다.

먼저, 장기기증 신청 홈페이지 메시지 변화에 따른 기증자의 신청률 변화 실험은 영국정부의 National Health Service(NHS)의 장기기증 신청서비스를 대상으로 행해졌다. 여론조사에 의하면 영국인 10중 9명이 장기기증의사가 있지만 실제 장기기증에 등록한 비율은 전체인구의 1/3에 불과하다. 이에 NHS는 장기기증 등록비율을 높이기 위하여 BIT에 의뢰하여 RCT를 수행하였다. 영국에서 장기기증은 NHS 홈페이지를 통하여 이루어진다. BIT는 기존의 단순 등록화면을 비교그룹으로 설정하고, 화면 하단에 장기기증의 가치와 중요성을 설명하는 문구[13]를 삽입한 등록화면을 처치그룹으로 설정하여, 각각 다른 등록화면을 통하여 얼마나 많은 사람들이 장기기증 등록을

12) BBC, http://www.bbc.com/news/uk-politics-26030205, 검색일자 2015. 9. 15.

13) 장기기증을 권장하는 아홉 가지 다른 문구가 장기기증 등록 홈페이지 하단에 표시되었다.

하였는지 비교하였다. 실험결과, 기존의 단순등록화면보다 장기기증의 가치와 중요성에 대한 문구를 등록화면 하단에 삽입한 홈페이지에서 평균 2.5% 장기기증 등록을 더 많이 한 것으로 나타났다.

BIT에 의한 성공적인 정책평가의 다른 예는 체납자에 대한 납부고지방식의 변화에 따른 납세실적의 변화 실험이다. BIT는 영국국세청과 함께, 장기체납자에 대한 납부 독촉고지서를 발송하였다. 이 실험에서 단순 납세촉구문구만을 표시한 고지서를 받은 체납자들을 통제그룹으로 설정하였고, 사회적 통념(Social Norm)을 통해 납무 의무를 강조하는 다양한 메시지[14]를 담은 고지서를 받은 체납자들을 처치그룹으로 설정하였다. 그 후 다른 형태의 메시지를 받은 체납자 그룹 간의 체납세액 납부실적을 비교하였는데, 납부의무를 강조하는 메시지를 받은 처치그룹의 납세자들이 일반적인 고지서를 받은 비교그룹의 체납자보다 2~5% 정도 체납세금을 더 납부한 것을 발견하였다.

〈표 Ⅲ-2〉 BIT 주요 실험평가의 정책개입의 예

| 그룹 | 장기기증 실험 | 체납고지 실험 |
|---|---|---|
| 통제 | No Message | No Message |
| 실험 | • Please join the NHS Organ Donor Register.<br>• Every day thousands of people who see this page decide to register.<br>• Three People die every day because there are not enough organ donors.<br>• You could save or transform up to 9 lives as an organ donor.<br>• If you needed an organ transplant would you have one? If so please help others. | • Nine out of then people pay their tax on time.<br>• Nine out of ten people in the UK pay their tax on time.<br>• Nine out of ten people in the UK pay their tax on time. You are currently in the very small minority of people who have not paid us yet.<br>• Paying tax menas we all gain from vital public services like the NHS, roads, and schools.<br>• Not paying tax means we all lose out on vital public services like the NHS, roads, and schools. |

출처: BIT, "Applying behavioural insights to reduce fraud, error and debt," 2012, p. 7.
BIT, "Applying Behavioural Insights to Organ Donation: preliminary results from a randomise controlled trial," 2015, p. 6.

14) 사회적 통념에 의해 체납에 대한 부정적인 인식을 제고하는 다섯 가지 다른 메시지가 독촉고지서에 포함되었다.

## 2. 미국 연방정부의 Evidence Based Policy Initiatives

미국의 사회실험을 통한 정책평가의 보급은 오바마 정부의 증거기반 정책 시행 계획(Evidence Based Policy Initiative)을 통해 이루어졌다. 오바마 정부는 부시 행정부의 PART가 평가자의 주관적인 인지에 의해 사업을 평가했기 때문에 정책의 정확한 효과를 측정하는데 한계가 있다는 비판을 수용하여 PART를 폐지하였다. 오바마 정부는 PART가 폐지된 후 성과정책을(Performance Initiative)을 발표하고 새로운 성과관리정책을 시행하였다. 'GPRA Modernization Act'와 'Evidence-Based Policy Initiatives' 등은 오바마 정부가 시행한 새로운 성과관리정책이다.

특히, 오바마 정부는 증거기반 정책시행계획을 구현하기 위해 추진전략을 마련하고 증거기반 평가를 활성화하는 노력을 기울였으며, 증거기반 평가 전담조직을 신설하는 접근방식을 택하였다. 요약하면 오바마 정부의 실증기반 정책평가계획은 가. 증거기반 정책평가 전략 수립 나. 증거기반 평가 활성화 사업과 전담조직 신설로 구체화되었다.

### 가. 증거기반 정책평가 전략 수립[15)]

오바마 행정부는 PART를 폐지하는 대신 성과계획서에 정책사업의 효과를 검증하는 것을 강조하였다. 이를 위해 신규 및 기존 사업 활동에 대한 예산 변경 및 사업조정을 위해서는 정책효과의 증거를 성과계획서에 제시하도록 하였다. 이러한 증거기반 성과정보는 진실험(Experiment)이나 준실험(Quasi-Experiment)을 통해 생성할 것을 권장하였다. 또한 과학적인 평가를 통해 생성된 성과정보는 전략적 리뷰(Annual Strategic Review)에서 활용될 수 있도록 하였다. 오바마 행정부는 증거기반 정책평가 확산을 위해 (1) 행정데이터 활용 (2) 저비용 실험평가 활성화 (3) 증거기반 공모제도 (4) 부처 평가역량강화와 같은 추진전략을 마련하였다.

---

15) OMB(2013), "Memorandum to the Heads of Department and Agencies,"의 내용을 참조하여 구성

(1) **행정데이터 활용**(High Quality Administrative Data)

재정사업을 정확하게 평가하고 비용을 절약하기 위해 개별부서에서 사용하는 있는 행정데이터를 확인하고 공유하도록 유도하였다. 원활한 공유를 위하여 개인 프라이버시와 보안에 대한 규정을 지키는 범위 내에서 각 개별부서가 보유하고 있는 데이터를 공유하여 정책평가에서 사용할 수 있도록 하였다. 구체적으로 'Provider Scorecards'를 만들어 각 개별부처가 보유하고 있는 자료를 공개하고 타 부서의 요청 시 자료를 공유할 수 있도록 하였다. 이러한 자료 공개를 통하여 평가를 진행하고자 하는 개별부처로 하여금 자료의 이용가능성을 확인하고 적은 비용으로 보다 다양한 과학적인 평가를 실시할 수 있도록 유도하였다.

(2) **저비용 실험평가**(Experimentation and low-cost evaluation)

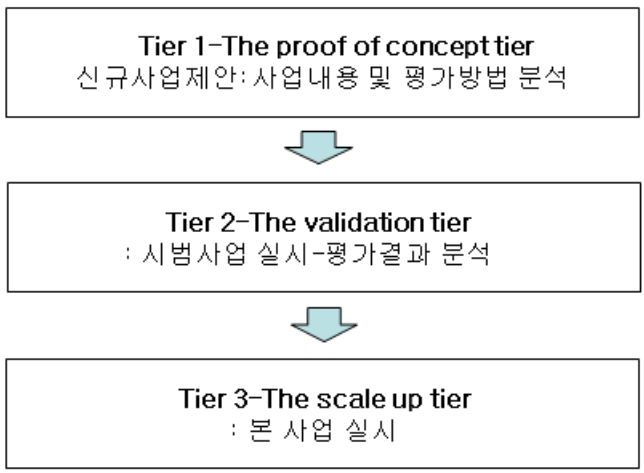
사업성과를 입증하는 평가수단으로 RCT를 적극적으로 권장하였다. 정책의 효과가 인간의 행동변화를 통해 나타나기 때문에 행동변화를 유도하는 정책을 시행하고 정책효과를 받은 집단과 받지 않은 집단의 행동변화를 비교하도록 했다. 그러나 무작위 추출(Randomized Sampling) 평가에는 비용과 윤리적인 측면에서 여러 장애요인이 많기 때문에 장애요인에 대한 대처방법을 교육하였다. 또한 평가를 위한 자료구축에는 비용이 많이 들기 때문에 앞서 소개한 자료공유시스템을 통하여 가용한 자료를 확인하고 이를 각 부처로 하여금 활용하게 하여 평가비용을 줄이려 했다. 윤리적인 문제를 줄이기 위해서는 단계별(Phase-in) 실험 또는 교대(Rotation) 실험을 통하여 모든 정책대상자가 시차를 두고 정책수혜자가 되도록 하였다.

(3) **증거기반 공모제도**(Evidence-based grant making)

오바마의 새로운 성과계획에서는 정책의 효과를 검증할 수 있는 정책공모를 적극적으로 활용하였다. 공모 제도를 통하여 각 부처나 지방 정부로부터 새로운 정책에 대한 제안을 받았다. 신규 정책 제안서를 제출하는 부서에는

정책의 효과에 대한 명확한 실증 증거를 제출하도록 하였다. 특히 정책공모를 통한 대상사업 선정은 단계별 실증평가 디자인(Tiered-evidence grant designs)방식으로 이루어진다. [그림 Ⅲ-1]과 같이 1단계에서는 신규 사업의 제안을 받는데, 신규 사업은 사업의 내용 및 평가방법을 기준으로 선정하고 평가준비를 위한 예산을 지원받는다. 2단계에서는 시범사업을 실시하고 평가결과를 분석하기 위한 지원을 받는다. 마지막 3단계에서는 시범사업 결과 효과가 있었던 사업을 전체 정책대상자로 확대하여 본 사업을 진행하고 이에 대한 예산을 지원한다. 현재까지 5개 부처에서 총 13개의 단계별 실증평가[16]를 진행해 왔다.

[그림 Ⅲ-1] Tiered-Evidence Grant Design의 구조




### (4) 부처 평가역량 강화

증거기반 정책평가를 정착시키기 위하여 오바마 정부는 부처의 평가역량을 강화하는 노력을 하였다. 이를 위해 각 부처로 하여금 우선순위가 높은

16) 평가분야는 education, teenage pregnancy prevention, home visitation programs, workforce, international assistance, 기타프로그램이다(OMB, 2013).

핵심 사업에 평가 관련 예산과 인력을 배정하는 부처 평가계획(Agency- wide Evaluation Plan)을 수립하도록 하였다. 또한 정부의 모든 평가결과를 저장한 'What Works Clearinghouse'를 만들어, 정책결정 및 중요사업결정에 활용하도록 하였다. 이러한 제도적 장치 외에 부처 간 평가 학습 네트워크(Cross-agency Learning Network)를 구축하여 각 부처의 평가자료, 평가기법, 평가 전략을 부처 간에 공유하도록 하였다.

### 나. 증거기반 평가 활성화 사업

증거기반 평가 추진전략을 세운 후에 이를 활성화하기 위해 노력하였다. 이러한 활성화 사업은 OMB의 주도로 이루어 졌다. 이를 위해 OMB는 실증기반 정책평가를 각 부처에 보급하기 위한 워크숍을 실시하였으며 RCT 활성화를 위한 'Low Cost RCT Competition을 실시하였다. 먼저, 각 부처에 증거기반 정책평가를 보급하기 위한 준비 작업으로 워크숍을 추진하였다. 다섯 차례에 걸쳐 워크숍[17]을 개최하였는데, 증거기반 정책평가를 위한 평가자원 배분, 평가자료 활용, 평가실시방법, 평가공모제도, 행동과학 평가의 적용 방법 등을 토론하기 위해 실시하였다.

Workshop Ⅰ: How can agencies focus evaluation resources on the most important program and policy question?

Workshop Ⅱ: How can agencies use administrative data sets from multiple programs and levels of government to answer important questions while protecting privacy?

Workshop Ⅲ: How can agencies conduct rigorous program evaluation and data analytics on a tight budget?

Workshop Ⅳ: How can agencies use their existing authority to turn a traditional competitive grant program into an innovative evidenced-based one?

17) OMB, "Memorandum to the Heads of Department and Agencies," 2013, p. 4.

Workship V: How can agencies harness research findings from the social and behavioral sciences to implement low-cost approaches to improving program results?

워크숍과 더불어, OMB는 RCT 실험평가를 확산시키기 위해 노력하였다. 이를 위해 정책평가 비영리기관인 'Coalition for Evidence-based Policy'를 통해 'Low cost RCT Competition'을 실시하여 저렴한 비용으로 RCT 실험평가를 실시하는 연구 과제를 지원하였다. 이는 RCT의 경우 자료측정에 많은 비용이 들기 때문에 기존 행정자료를 활용한 평가를 실시하여 RCT 실험평가를 확산시키고자 하기 위함이었다. 이 RCT 실험평가 지원 사업으로 20만 달러 이하 저비용 RCT 정책평가과제 세 개를 선정하여 평가를 진행 중이다. 평가과제는 다음과 같다[18].

(1) A large multi-site RCT of Bottom line(Evaluation Cost: $159,000)

저소득층 고등학생들에게 전문상담가가 개인, 학업, 진로선택에 대한 일대일 상담을 해주는 프로그램이 고등학생들의 대학진학률과 졸업률에 미치는 영향을 평가하는 과제이다. 이 과제는 지속사업에 대한 평가로서 과거 회귀불연속 방법(A regression-discontinuity study)에 의한 평가에서 상당한 효과가 있는 것으로 분석되었다. RCT 실험평가에선 선착순 방식(First Come First Served)으로 대상자를 선정했던 과거와는 달리 추첨방식(Lottery)으로 대상자를 선정하여, 추첨되어 처치집단으로서 상담을 받은 학생과 추첨이 되지 않아 상담을 받지 않은 비교집단의 학생들의 정책효과를 비교하는 자연실험 방식으로 진행되고 있다.

18) 실험내용은 Coalition for Evidence-based Policy 홈페이지 참조하여 구성. http://coalition4evidence.org/low-cost-rct-competition/, 검색일자. 2015.9.15

(2) A large RCT of Durham Connects(Evaluation Cost: $183,000)

이 실험평가는 노스캐롤라이나 주 Durham 카운티의 1,100개의 가구를 대상으로 실시되고 있으며 프로그램명은 'Nurse Home Visiting Program'이다. 이 프로그램은 간호사가 3~12주의 신생아가 있는 가정을 방문하여 신생아의 건강을 관리해 주는 프로그램으로서, 과거부터 존재해왔다. 기존에는 정책대상자를 선착순으로 선정하여 선정된 가정에 간호사가 방문하는 방식을 사용하였다. 그러나 현재는 서비스 신청 가정 중 간호사가 방문하는 가정을 무작위 추첨으로 선발한다. 따라서 추첨된 가정은 자연스레 처치집단이 되고 추첨되지 않은 집단은 비교집단이 된다. 실험 후 추첨된 가정의 신생아 건강상태와 추첨되지 않은 가정의 신생아 건강상태를 비교하여 정책의 효과를 평가하는 방식으로 진행되고 있다.

(3) A large multi-site RCT of workplace health and safety inspection (Evalution Cost: $153,000)

이 실험평가는 미국의 연방정부 OSHA(Occupational Safety and Health Administration)가 실시하는 프로그램이다. OSHA는 1970년대 이래로 직장 내 안전과 건강을 위한 조사를 실시하여 오고 있으나 이러한 조사의 효과는 정확하게 분석되지 않았다. 정확한 분석을 위해 OSHA는 2만 9,000개의 회사를 선발하여 처치집단과 비교집단으로 나누고 이 중 처치집단에 속한 회사에만 안전 및 건강조사를 실시하여 조사 실시 후에 처치집단에 속한 회사와 비교집단에 속한 회사 간에 장단기 안전사고 발생률(단기: 1~2년 후 안전사고 발생률; 장기: 3~4년 후의 안전사고 발생률)의 차이를 비교하는 평가를 진행 중이다.

### 다. 전담조직 설치[19)]

오바마 정부는 증거기반 추진전략과 활성화 방안 마련과 더불어 Behavioral Insight Team을 벤치마킹하여 2014년에 Social and Behavioral Science Team(SBST)를 설치하였다. RCT와 같은 과학적인 실험정책평가를 통하여 정책의 정확한 효과를 측정하는 영국의 Behavioral Insight Team의 성공에서 영향을 받았기 때문이다. SBST는 현재 미국 Office of Science and Technology에 속해 있다. SBST는 최근 1년 동안 미국 교육부(Department of Education)와 국방부(Department of Defense)의 RCT 평가를 지원하였다. SBST팀이 실시한 대표적인 RCT 평가는 현재 미국에서 사회문제가 되고 있는 대학학자금 대출의 미납을 줄이기 위한 실험이다. 학자금 대출 상환내용을 설명한 이메일을 받은 학생들이 받지 못한 학생들보다 상환계획 프로그램에 더 많이 참가하고 있음을 밝혀낸 평가실험을 진행하였다.

## 3. MIT의 J-PAL[20)]

### 가. 설립연혁

MIT 경제학과의 The Abdul Latif Jameel Poverty Action Lab(일명 J-PAL)은 주요 개발도상국의 빈곤문제해결을 위한 정책의 효과성을 무작위 추출평가(Randomized Evaluation) 방식으로 평가하기 위한 글로벌 전문가 네트워크로 2003년 설립되었다. 2003년 설립 이후 J-PAL은 아프리카, 유럽, 남아메리카, 북아메리카, 남아시아 지역 오피스를 설립하고 RCT에 기반하여 개발도상국의 빈곤을 극복하는 정책에 대한 효과성을 과학적으로 평가하여 왔다. J-PAL Lab의 주요정책평가 영역은 다음과 같다.

---

19) SBST 홈페이지의 내용참조, https://sbst.gov/, 검색일자. 2016.9.15

20) Abdul Latif Jameel Poverty Action Lab의 홈페이지 내용과 필자가 참가한 J-PAL Executive Course내용을 참조하여 구성
http://www.povertyactionlab.org, 검색일자 2015. 9. 15.

〈표 Ⅲ-3〉 J-PAL의 주요 정책평가 분야

| |
|---|
| 농업(Agriculture)<br>교육(Education)<br>환경 및 에너지(Environment & Energy)<br>재정(Finance)<br>보건(Health)<br>노동시장(Labor Market)<br>정부운영(Governance) |

출처: J-PAL 홈페이지, http://www.povertyactionlab.org/policy-lessons, 검색일자 2015. 9. 15.

## 나. 활동 및 성과

2003년 설립 이후 J-PAL은 (1) 개도국 빈곤개선을 위한 RCT 기반 정책평가 실시, (2) 정책 평가결과의 보급과 확산(Scale-up) (3) RCT 기반에 대한 교육프로그램 제공과 같은 활동을 벌여왔다. J-PAL Lab은 과학적인 정책평가를 위해 전 세계 지역사무소를 통해 각국 정부, 비정부기관, 대학과 파트너십을 구축하고 있으며 이러한 네트워크를 통해 RCT 평가결과의 확산과 보급, 제휴기관의 직원들에 대한 교육 등을 제공하고 있다. 이러한 교육 프로그램은 MIT의 J-PAL LAB과 각 지역 오피스에서 실시하는 Executive Training Course와 개별 파트너기관과 진행하는 Workshop Course가 있다.

〈표 Ⅲ-4〉 J-PAL RCT 주요 교육 내용

| 일차 | 주요 교육 내용 |
|---|---|
| Day 1 | Lecture 1: What is Evaluation?<br>Case Study 1: Program Theory(Group Discussion)<br>Lecture 2: Outcomes, Indicators, and Measuring Impact<br>Group Project: Choose topics for presentation(Theory of Change) |
| Day 2 | Group Project: Theory of Change<br>Case Study 2: Why Randomize?(Group Discussion)<br>Lecture 3: Impact Evaluation - Why Randomize?<br>Group Exercise 1: Mechanics of Randomization<br>Case Study 3: How to Randomize?(Group Discussion)<br>Lecture 4: How to Randomize?<br>Group Project: Evaluation Design |
| Day 3 | Group Exercise 2: Random Sampling and Law of Large Numbers<br>Lecture 5: Sampling and Sample Size<br>Case Study 4: Threats and Analysis(Group Discussion)<br>Lecture 6: Threats and Analysis<br>Group Project: Evaluation Design and Power Calculations |
| Day 4 | Lecture 7: Cost Effectiveness and Scaling Up<br>Group Exercise 3: Power Calculations and Sample Size<br>Lecture 8: Randomized Evaluation: Start-to-Finish<br>Group Project: Finalize presentation |
| Day 5 | Group Presentations<br>Course Wrap Up |

출처: J-PAL 홈페이지, http://www.povertyactionlab.org/course/agenda, 검색일자 2015. 9. 15.

J-PAL은 2003년 설립 이후 〈표 Ⅲ-5〉와 같이 다양한 활동을 펼쳐왔다. 2003년부터 2014년까지 약 520명의 교수과 함께 세계 60여개국 이상에서 2,610개의 RCT를 진행하여 왔다. 또한 2005년부터는 MIT J-PAL 본원 및 지역 오피스를 통하여 약 13,000여명의 각국 정부, 비영리기관, 민간기업, 학생들에게 RCT 관련 교육을 제공하여 무작위 추출에 의한 정책평가의 보급 및 전문가 육성에 기여하여 왔다.

〈표 Ⅲ-5〉 J-PAL RCT 정책의 주요성과

| 연도 | 평가 관련 교수 (Affiliated Professor) | RCT 평가회수 (Ongoing or Completed Evaluation) | 교육이수자 (Trained people) |
|---|---|---|---|
| 2003 | 4 | 33 | – |
| 2004 | 8 | 46 | – |
| 2005 | 11 | 48 (10) | 59 |
| 2006 | 13 | 75 (15) | 143 |
| 2007 | 18 | 98 (15) | 278 |
| 2008 | 30 | 181 (21) | 440 |
| 2009 | 44 | 214 (30) | 616 |
| 2010 | 55 | 235 (40) | 857 |
| 2011 | 64 | 302 (50) | 1173 |
| 2012 | 70 | 356 (52) | 1474 |
| 2013 | 92 | 439 (54) | 1620 |
| 2014 | 111 | 583 (62) | 6643 |
| Total | 520 | 2610 | 13303 |

주: ( ) country

출처: J-PAL 홈페이지, http://www.povertyactionlab.org/History, 검색일자 2015. 9. 15.

## 다. 주요 RCT 정책 평가

J-PAL의 주요 정책평가에는 약 3천 300만명의 학생들에게 혜택이 돌아간 인디아의 'Remedial Education'이 있다. 이 실험은 2001년에 시작되어 2009년에 종료된 사업이다. 전통적으로 인디아는 시골지역의 문맹률이 높았는데 6~14세의 아동의 약 47%가 Grade 2 수준의 글을 읽지 못하고 67%의 아동들이 기초계산을 할 수 없었다. J-PAL은 인디아 Pratham 지역의 200개 대상 학교 중 무작위로 추출한 절반의 학교에서 Remedial Education 프로그램을 시행하였다. Remedial Education 프로그램에서는 일기와 계산에서 기초학력 미달 학생들을 위하여 보조선생님을 고용하고 저학력 학생들을 위한 방과 후 보충수업을 실시하였다. 'Remedial Education'에서 제공하는 보충수업에 참가한 학생들은 보충수업을 참가하지 않은 학생보다 읽기와 수학시험 성적이 7.7%나 더 향상되었다.

Remedial Education이 RCT 평가에서 효과가 있는 것으로 평가되었기 때문에 Pratham 정부는 지역 내 전체 학교 3,300만명의 학생을 대상으로 'Read India' 도입하여 Remedial Education에 실시한 읽기 및 계산 방과 후 보충수업을 시행하였다.

# Ⅳ. 우리나라의 RCT 현황과 적용 가능성 탐색

## 1. 우리나라의 사회실험에 의한 정책평가 사례

RCT와 같은 사회실험을 통하여 정책의 정확한 효과를 측정하고 이를 예산배정과 사업의 효과성 제고에 활용한 영국이나 미국과 달리 우리나라에서 사회실험을 정책평가에 활용한 사례는 매우 적은 편이다. 이는, 사회실험을 수행하는데 여러 가지 법적, 제도적, 윤리적 측면에서 많은 장애요인이 있기 때문이기도 하지만, 정책평가 전문가 집단에서 사회실험에 대한 이해가 낮은 것도 하나의 원인이다. 따라서 사회실험을 통한 정책평가는 개별부처의 몇몇 단위사업에서 사회실험 전문가의 개입으로 매우 드물게 실시되어 왔다. 본 연구에서는 고용노동부의 사회보험료 지원사업인 '두루누리 사업'과 문화체육관광부의 '여행 바우처 사업'의 정책 실험 사례를 소개한다.

### 가. 두루누리 사업

두루누리 사업은 고용노동부가 저소득층의 사회보험 사각지대를 해소하기 위해 2012년에 실시하였다. 두루누리 사업은 10명 미만의 사업장에 근무하는 월 보수 125만원 이하인 근로자와 고용주들에게 고용보험과 국민연금 납부액의 일부[21]를 지원하는 제도다. 이 사업은 2단계로 구성되었다. 1단계는 시범사업으로서 2012년 2월부터 6월 사이에 일부 시도에 한하여 실시되었다. 2단계는 본사업으로서 2012년 7월부터 시범사업의 성과를 기초로 전국적으로 확대되었다. 두루누리 사업 시범사업의 평가는 보기 드물게 실험평가 방법으로 진행되었다. 시범사업에서 전국 16개 광역시도의 기초자치단

21) 월 보수 105만원 미만의 근로자와 고용주에게는 보험료 납부액의 1/2을, 105만~125만원 미만의 근로자와 고용주에게는 보험료 납부액의 1/3을 지원한다(유경준 외, 2015).

체 2곳을 선정한 후 1곳은 처치지역(Treatment) 다른 한곳은 통제지역으로 두었다.

시범사업 종료 후 사업의 효과를 평가하기 위해 처지지역과 통제지역의 고용보험과 국민연금의 가입자 수의 변화를 이중차분법을 사용하여(Differences in Difference) 측정하였다. 그 결과 〈표 Ⅳ-1〉에서 볼 수 있듯이 처치지역의 고용보험 가입자 수는 7,907명이 증가하여 시범사업을 통해 약 2.4% 정도 고용보험이 증가한 것으로 평가되었다. 국민연금의 가입자 수도 시범기간 동안 9,484명이 증가하여 약 3.1% 정도 증가한 것으로 분석되었다. 시범사업 결과 정책이 효과가 있는 것으로 분석되었기 때문에 전국적으로 확대된 두루누리 본사업이 시작되었고, 본사업도 고용보험과 국민연금의 가입자 수를 확대[22]시킨 것으로 분석되었다. 이러한 두루누리 사업의 성과는 엄밀한 평가 없이 본사업 시행의 정치적 명분을 획득하기 위해 시행되고 있는 우리나라 시범사업의 성과를 정확히 평가하여 이를 본사업의 확대시행과 연계시켰다는 측면에서 중요한 사례라 할 수 있다.

**〈표 Ⅳ-1〉 두루누리 사업의 효과(Differences in Differences)**

(단위: 명)

| | 시기/처치집단 | 시범사업 지역 | 후보 지역 | 차 분 |
|---|---|---|---|---|
| 고용보험 피보험자 수 변화 | 처치 이전 5개월 (2011/02~2011/06) | 8,751 | 5,202 | 3,549 |
| | 처치 적용 5개월 (2012/02~2012/06) | 22,601 | 11,145 | 11,456 |
| | 차분 | 13,850 | 5,943 | 7,907 |
| 국민연금 가입자 수 변화 | 처치 이전 5개월 (2011/02~2011/06) | 13,901 | 8,671 | 5,230 |
| | 처치 적용 5개월 (2012/02~2012/06) | 21,855 | 7,141 | 14,714 |
| | 차분 | 7,954 | −1,530 | 9,484 |

출처: 유경준 · 강창희 · 최바울, 「사회보험료 지원사업(두루누리 사업)의 효과: 현대성과평가론의 적용」, 2015, p. 12, 14.

22) 본사업은 고용보험 피보험자 수를 약 2.68%, 국민연금 피보험자 수를 약 2.04% 증가시킨 것으로 분석되었다(유경준 외, 2015).

### 나. 여행 바우처 사업

여행 바우처 사업은 사회적, 경제적 제약으로 여행이 어려운 저소득층에게 여행 바우처를 지급하여 여행을 독려하고 개인과 가정에서 삶의 만족도를 증가시키기 위한 사업이다. 여행 바우처는 2005년부터 문화체육관광부에서 시작하였으며 각 지자체는 사업시행자로서 문화체육관광부로부터 여행 바우처 재원을 받아 지역의 저소득층 주민들에게 여행 바우처를 발급하고 있다. 바우처의 지원액은 개별여행은 1인당 15만원, 가족여행은 최대 30만원을 지원한다. 신청자격 요건은 기초생활수급자, 법정 차상위계층, 우선돌봄 차상위 가구가 대상이다. 지원대상을 선정하는 방식은 각 지자체에 맡겨져 있다. 여행 바우처 사업의 예산이 한정되어 있기 때문에 대부분의 지자체에서는 신청자 중에서 수혜대상자를 선착순 방식(First Come, First Served)으로 선발하거나 소득이 낮은 순으로 선발해 왔다. 이러한 선발방식은 선정과정이 간단하지만 정책평가 시 내생성(Endogeneity)으로 인하여 사업효과를 정확하게 추정하기 어렵다.

그러나 예외적으로 서울시는 지원대상자 선발을 신청자 중에서 무작위 추첨방식(Randomized Lottery)으로 진행하였다. 무작위 추첨방식은 신청자 중에서 당첨자가 처치집단이 되고 미당첨자는 비교집단이 된다는 점에서 자연스럽게 RCT 실험평가를 활용할 수 있다. 서울시는 2011년 여행 바우처 사업의 신청자 5,547명의 절반 정도를 추첨을 통하여 선발하여 여행 바우처를 지급하였고 2011년 사업종료 후 바우처를 지급받은 신청자 그룹과 탈락그룹의 여행 여부, 여행시간, 만족도를 비교하였다. 그 결과, 바우처를 지급받은 사람들의 여행참여, 여행일 수, 여가·문화 만족도가 증가한 것으로 분석되었다.[23] 여행 바우처 사업은 한정된 예산으로 모든 정책대상자를 수혜자로 선발할 수 없을 때 무작위 추첨을 통한 실험평가를 통해 정책의 효과를 엄밀하게 평가했다는 점에서 RCT에 관한 의미 있는 사례[24]이다.

---

23) 바우처의 지급으로 여행 및 만족도가 증가하는 것은 정책효과라기보다 현물 보조의 직접적인 효과이기 때문에, 정책대상자의 정책처치에 따른 행동변화를 예측하기 위해서 바우처의 지급이 종료된 후에도 여행일 수와 만족도가 변화하였는지에 대한 추가연구가 필요하다.

## 2. 사회실험의 적용 가능성 탐색

우리나라에서는 RCT와 같은 사회실험이 거의 활성화되어있지 않다. 그러나 사회실험은 엄밀하게 설계되었을 때 정책의 효과를 가장 정확하게 평가할 수 있는 도구이다. 또한 평가를 통하여 사업을 중단하거나 개선할 수 있다는 점에서 재정의 낭비를 막을 수 있는 효과적인 정책수단이 된다. 본 연구에서는 사회실험이 활용될 수 있는 정책평가의 범위를 사전평가와 사후평가, 그리고 정책분야별로 탐색하고자 한다.

### 가. 사전평가

사회실험을 가장 효과적으로 적용할 수 있는 평가단계는 사전평가이다. 본사업 시행 전에 실험을 통하여 미래의 효과를 평가하고 평가결과에 근거하여 사업의 시행 여부를 결정할 수 있기 때문이다. 우리나라에서 사전평가는 예비타당성조사와 시범사업의 형태로 실시되어 왔다. 예비타당성조사는 총 사업비가 500억원 이상이면서 국가재정지원 규모가 300억원 이상인 건설사업, 정보화사업, 국가연구개발사업과 중기재정지출이 500억원 이상인 사회복지, 보건, 교육, 노동, 문화 및 관광, 환경보호, 농림해양수산, 산업·중소기업 분야의 사업에 걸쳐 실시한다. 예비타당성조사가 대규모 재정사업을 효과를 계량적으로 추정하여 무분별한 재정사업을 막는 데 기여해 왔다는 평가를 받고 있지만, 추정의 정확성은 여전히 많은 논란이 되고 있다. 특히, 건설이나 SOC 사업 외에, 교육, 복지, 노동, 문화 사업 등은 B/C나 B/E 분석과 같은 계량기법으로 편익(Benefit)이나 효과(Effectiveness)를 정확히 추정하기 어렵다.

이러한 한계를 보완하기 위하여 예비타당성조사는 경제적 효과 외에 정책적 효과, 지역균형발전, 기술적 효과를 함께 평가한 후 각 평가항목에 대한 가중치를 부여하여 최종 대안을 선택하는 AHP 의사결정기법에 의해 사업의

24) 박상곤, 「관광정책 평가방법 및 사례발표: 여행바우처 사업의 효과분석을 중심으로」, Working Paper, 한국행정학회, 2015.

시행 여부를 결정한다. 비교적 편익이 명확히 측정되는 건설사업과 달리 기타 재정사업의 경제성 효과 추정의 정확성은 추정의 어려움으로 인하여 떨어진다. 게다가 〈표 Ⅳ-2〉에서 보듯 기타 재정사업의 경우 정책적 분석의 비율이 50~70%를 넘는다. 비계량적으로 측정되는 정책적 분석의 가중치가 50%를 넘기 때문에 기타 재정사업의 예비타당성조사 예측 분석의 정확성은 현저하게 떨어질 가능성이 높다. 이러한 문제에도 불구하고 500억원 미만의 고용, 복지, 문화, 예술 등 소규모 비건축 사업에 간이 예비타당성조사 도입을 검토하고 있어 정확한 사전평가 방법론의 정립이 시급한 실정이다.

〈표 Ⅳ-2〉 평가항목의 가중치 범위

| 구분 | | 경제성 분석 | 정책적 분석 | 지역균형 발전 | 기술성 분석 |
|---|---|---|---|---|---|
| 건설사업 | | 40~50% | 25~35% | 20~30% | – |
| 기타 재정사업 | B/C 분석 시 | 40~50% | 50~60% | – | – |
| | E/C 분석 시 | 20~40% | 60~80% | – | – |
| R&D 정보화 | B/C 분석 시 | 40~50% | 20~30% | – | 30~40% |
| | E/C 분석 시 | 30~40% | | – | 40~50% |

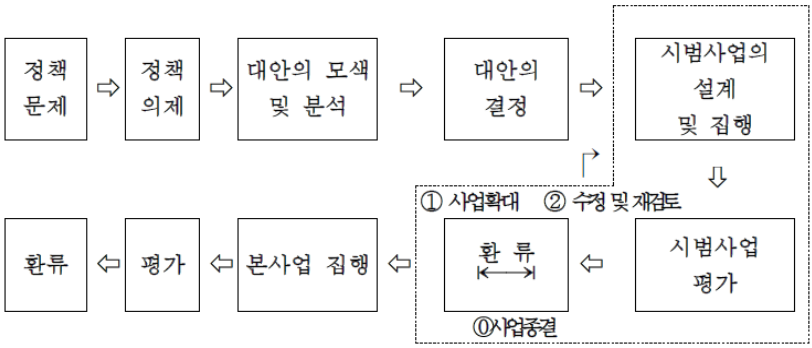
출처: 한국개발연구원, 「2014 KDI 예비타당성조사 지침 제 38조」를 토대로 저자가 표 작성.

이러한 예비타당성조사의 한계로 인하여 개별부처에 많이 사용되고 있는 방법이 시범사업(Pilot Projects)이다. 시범사업은 이삼열 외(2009, p. 13)는 시범사업의 정의를 "본 정책의 전면적인 집행에 앞서 정책의 효과성을 검증하고 공공자원의 낭비 등을 피해 정책의 효율성을 높이는 것을 목적으로, 특정 정책의 효과나 작동기제를 사전에 측정 또는 관찰하기 위한 엄격한 사전설계를 바탕으로 집행되는 비교적 작은 규모의 사업"이라고 정의한다.

시범사업은 엄격한 사전설계에 의하여 본정책 실시 전에 작은 규모로 사업을 실시하고 그 효과에 기반하여 본사업 시행 여부를 결정한다는 점에서 가장 효과적인 증거기반 사업평가(Evidence based policy) 수단이 될 수 있다. 특히, 시험사업은 사회실험을 위한 좋은 환경을 제공한다. 시범사업 내

에서 처치집단과 비교집단을 구분하거나 시범사업의 정책수혜자와 시범사업의 수혜자가 아닌 일반 정책대상자를 처치집단과 비교집단으로 설정하여 효과를 비교할 수 있기 때문이다.

[그림 Ⅳ-1] 시범사업이 고려된 정책과정




출처: 이삼열 · 정의룡 · 이은하, 「시범사업에 관한 탐색적 연구: 보건복지가족부 사업을 중심으로」, 2009, p. 16.

시범사업에서 사회실험을 적용할 수 있음에도 불구하고 시범사업에서 엄밀한 사회실험을 통해 평가한 경우는 매우 드문 편이다. 예를 들어 이삼열 외(2009)는 시범사업의 시행이 법으로 명문화[25]된 보건복지부에서 실시된 2002~2009년까지 시범사업의 수와 평가방법을 분석하였다. 분석결과 보건복지부는 7년 동안 62개의 시범사업을 실시했는데 그중 21개 시범사업의 평가서가 존재하였다.

21개 평가서 중에서, RCT와 같이 엄밀하게 설계된 처치-비교집단 사전사후 비교설계방식으로 평가한 시범사업 평가는 1개[26]만 존재(전체 2%)하였다. 나머지 대다수는 단일집단 사후설계(81%)와 단일집단 사전사후 비교설

25) 「보건의료 기본법」 제44조(보건의료 시범사업) 1. 국가 및 지방자치단체는 새로운 보건의료 제도를 시행하기 위하여 필요한 경우에는 시범사업을 실시할 수 있다. 2. 국가 및 지방자치단체는 제1항의 규정에 의한 시범사업을 실시한 때에는 그 결과를 평가하여 새로이 시행될 보건의료제도에 반영해야 한다.

26) 진실험 방식에 의해 평가는 과학적인 실험설계가 가능한 병원에서의 금연프로그램의 효과에 대해 행해졌다(이삼열 외, 2009).

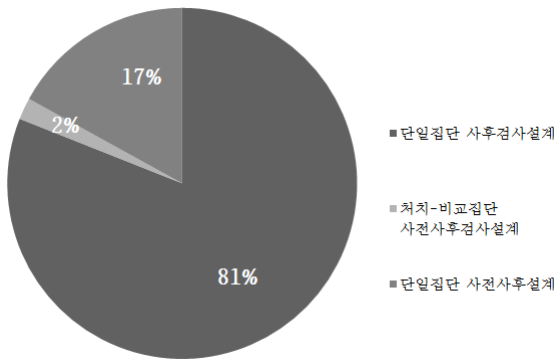
계(17%)로 평가하였다. 이는 많은 경우 시범사업의 평가가 제대로 실시되지 않고 있으며, 평가를 진행하더라도 RCT와 같은 엄밀한 정책평가가 실시되지 않고 있음을 보여준다. 이런 점에서 우리나라의 시범사업은 사전검증을 통한 사업의 효율성제고라는 시범사업 본래의 취지를 살리지 못하고 있는 실정이다.

〈표 Ⅳ-3〉 보건복지부 시범사업 시행횟수

| 연도 | 시범사업의 개수 | 시범사업 평가서 수 |
|---|---|---|
| 2002 | 4 | 0 |
| 2003 | 8 | 1 |
| 2004 | 4 | 2 |
| 2005 | 13 | 7 |
| 2006 | 6 | 2 |
| 2007 | 7 | 4 |
| 2008 | 20 | 5 |
| 총계 | 62 | 21 |

출처: 이삼열 · 정의룡 · 이은하, 「시범사업에 관한 탐색적 연구: 보건복지가족부 사업을 중심으로」, 2009, p. 32.

[그림 Ⅳ-2] 보건복지부 시범사업 평가방법론 비율




출처: 이삼열 · 정의룡 · 이은하, 「시범사업에 관한 탐색적 연구: 보건복지가족부 사업을 중심으로」, 2009, p. 33.을 수정

따라서 예비타당성조사와 시범사업의 한계를 극복하기 위하서는 재정사업 사전평가에 있어 RCT에 의한 실험평가를 전제로 한 시범사업을 제도화하는 것을 검토할 필요가 있다. 특히, 대규모 예산이 들어가는 비건설 분야 재정사업의 경우 편익을 정확하게 예측하기 어렵기 때문에 RCT 평가기반 시범사업을 실시한 후 정확한 정책 효과가 있을 때만 본사업을 실시하는 것을 검토할 수 있다. 소규모 사업의 경우에도 기존의 활용 가능한 데이터가 있을 경우에는 시범사업의 저비용 RCT 평가 설계를 통해 정책의 효과를 정확히 검증할 수 있을 것이다.

### 나. 사후평가

RCT와 같은 무작위처지－비교집단 실험설계는 사전평가 외에 본사업의 평가를 위해서도 사용할 수 있다. 재정사업의 경우 대표적인 사후평가는 재정사업 자율평가와 심층평가이다. 재정사업 자율평가는 사업평가에 대한 평가지표[27]를 설정하고 각 부처로 하여금 적어도 3년 주기로 사업의 성과를 평가하고 이를 사업개선에 환류하도록 하고 있다. 재정사업 심층평가의 경우도, 사업의 성과가 낮은 사업 및 사업군에 대하여 전문가 그룹의 심층적인 평가[28]를 통하여 문제점을 발견하고 사업의 개선을 추구한다. 재정사업 자율평가와 심층평가는 우리나라의 정책평가 보급에 기여해 왔다는 평가[29]를 받고 있다.

이러한 성과에도 불구하고 성과정보의 질은 높지 않은 편이며 부처에서 사업개선을 위한 수단으로 잘 활용되지 않고 있다. 이는 평가자가 주관에 기초하여 체크리스트 방식으로 재정사업자율평가를 시행하고 있으며, 심층

27) 사업평가에 관련된 평가지표는 “성과/환류 사업이 효과적으로 수행되는지 점검하기 위한 사업평가를 실시하였는가?”와 “평가결과 및 외부지적사항을 사업구조개선에 환류하였는가?”이다(기획재정부, 「2013년도 재정사업자율평가 지침」, 2013).

28) 재정사업의 심층평가는 재정사업의 적절성, 효과성, 집행성과를 분석한다(한국개발연구원, 「2007년 재정사업심층평가 지침」, 2007).

29) 오영민, 「재정사업성과평가의 운영성과와 제도적 개선방안」, 『재정포럼』 12월호, 한국조세재정연구원, 2014.

평가의 경우도 RCT와 같은 실험평가를 통해 사업의 효과성을 엄밀하고 평가하고 있지 않기 때문이다. 〈표 Ⅳ-4〉에서 보듯 개별사업에 대해 심층평가를 진행하였던 2005~2009년까지 RCT에 기반하여 무작위로 처치집단과 비교집단을 설정하여 진실험 형태의 평가를 하는 경우는 거의 없는 실정이다. 처치집단과 비교집단을 설정하지 않거나 무작위추출을 통한 두 집단 간의 동질성이 확보되지 않은 준실험 방식이 3건, DID(Differences in Differences) 방식이 4건 진행되었을 뿐이다.

〈표 Ⅳ-4〉 재정사업 심층평가 평가방법론

| 연도 | 평가명 | 평가방법론 |
|---|---|---|
| 2005 | 노인 일자리 사업 | 문헌 |
| 2005 | 지역전략산업 진흥사업 | 설문조사, 현장방문, 토론 |
| 2005 | 해외취업 지원사업 | 준실험모형, 설문조사 |
| 2006 | 광역관광기반시설 확충사업 | CVM(가상상황평가법) |
| 2006 | 국가어항 건설사업 | VAR 모형, 횡단면, 시계열, 패널분석 |
| 2006 | 농산물유통시설 효율화지원사업 | 다른 대상과 비교 분석, 사례분석 |
| 2006 | 농어촌의료서비스 개선사업 | 문헌, 조사, 전문가 의견 수렴, 시계열분석, DID |
| 2006 | 문화콘텐츠 진흥사업 | 사업 성과지표에 의한 성과분석 |
| 2006 | 미취업청년 취업지원사업 | 비교집단을 통한 성과평가 |
| 2006 | 어업구조조정사업 | 문헌, 데이터 분석 |
| 2006 | 자활근로사업 | 회귀분석 |
| 2006 | 지식정보자원 관리사업 | 성과지표를 통한 정량평가 |
| 2006 | 해외마케팅 지원사업 | t-test, 설문조사 |
| 2007 | 기본보조금 지원 사업 | 이중차감법(DID), 설문조사 |
| 2007 | 농지규모화사업 | (비)모수적 분석, 직접조사 |
| 2007 | 실업자 직업훈련사업 | 통제집단 비교평가, 비용편익분석 |
| 2007 | 어업인 정책보험 사업 | 문헌, 유사제도와 비교 |
| 2007 | 운행차 저공해화 사업 | 통계분석, 비용편익분석 |
| 2007 | 일선수협 경영개선 지원사업 | t-test, DID |
| 2007 | 재래시장 활성화사업 | 문헌, 실태조사, 설문조사 |
| 2007 | 중소기업 기술혁신개발사업 | 회귀분석, 설문조사 |
| 2007 | 지방대학 혁신역량 강화사업 | DID, 설문조사 |

〈표 Ⅳ-4〉의 계속

| 년도 | 평가명 | 평가방법론 |
|---|---|---|
| 2008 | 국민임대주택 건설지원 | 문헌, 심층인터뷰 |
| 2008 | 군 복지시설 확보운영사업 | 문헌, 만족도조사 |
| 2008 | 대단위 농업종합개발사업 | 문헌 |
| 2008 | 사업주 직능개발지원 | 실증분석 |
| 2008 | 산전후 휴가/육아휴직 지원사업 | 문헌 |
| 2008 | 아동청소년 방과 후 돌봄 서비스 | 문헌 |
| 2008 | 조기 재취업수당사업 | 문헌 |
| 2008 | 지방대 장학금지원사업 | 설문조사 |
| 2008 | 친환경농업 인프라지원사업 | 설문조사 |
| 2008 | 클린사업장 조성사업 | 문헌 |
| 2008 | 풍수해보험사업 | 문헌 |
| 2008 | 해외자원개발사업 | 문헌 |
| 2009 | 고용유지 지원금사업 | 동태적선형패널모형 |
| 2009 | 기초노령연금 지원사업 | 횡단면, 패널분석 |
| 2009 | 대학 구조개혁지원 | 문헌, 성과지표 달성도 |
| 2009 | 수산비축 및 수매지원 | 문헌 |
| 2009 | 외국인투자유치 | 문헌 |
| 2009 | 장애인 사회활동지원 | 성과목표부합도 등 |
| 2009 | 재해예방시설지원 | 자체평가결과 분석, 회귀분석, 방문 |
| 2009 | 중기 모태조합출자 | 문헌 |
| 2009 | 창업기업 투자보조금사업 | 설문조사, 통계분석 |
| 2009 | 한국국제협력단사업 | 문헌 |

출처: 한국조세재정연구원 재정사업 심층평가 내부자료

이는 RCT와 같은 엄밀한 실험평가가 윤리적, 제도적, 비용적 측면에서 여러 제약이 있지만 우리나라의 대표적인 사후평가에서 전혀 활용이 되고 있지 않는 현실을 보여준다. 따라서 증거기반 정책평가의 보급을 위해 RCT가 적용 가능한 평가방식을 개발하고 시행이 용이한 정책분야를 발굴할 필요가 있다. 이를 위해 현재 사업군으로 평가하여 실험평가가 불가능한 심층평가 대상 중 일부 사업을 개별사업평가로 전환이 필요하다. 이러한 사후평가는 전체 정책대상집단에 대하여 사업을 실시하는 본사업의 성격상 특정그룹이 배제되지 않도록 다중처치실험(Multiple Treatments), 단계별 처치(Phase-in Treatment), 교대처치(Rotation Treatment)와 같은 평가 설계를 검토할 수 있

다. 또한 특정 사업이 예산제약이 있어 전체대상자에게 사업을 시행할 수 없다면 정책수혜자를 대상자 중에서 무작위 추첨을 통해 선발하여 평가하는 것을 검토할 수 있을 것이다. 이와 같은 엄밀한 실험평가를 통해 얻어진 평가정보를 토대로 사업을 종료하거나 사업의 추진방식을 변경하여 예산을 절약하거나 사업성과의 개선을 도모할 수 있다.

### 다. 정책분야

RCT와 같은 실험평가가 적용될 수 있는 정책분야는 넓다. 미국 OMB[30]는 사회실험의 평가가 적용될 수 있는 전제조건을 세 가지로 제시하고 다음의 조건만 충족된다면 대부분의 정책분야에서 RCT와 같은 실험평가를 실시할 수 있음을 밝히고 있다.

(1) 사업의 참가자와 비참가자를 무작위 추출로 두 집단 또는 다중 집단으로 구분할 수 있으며, 이런 무작위 추출로 구성될 샘플의 크기가 충분히 큰 사업
(2) 각 집단 간에 다른 사업추진방식을 적용할 수 있는 사업
(3) 각 집단별로 다른 사업추진방식을 적용함으로 인하여 발생한 사업의 결과(Outcomes)을 측정할 수 있는 사업

이론적으로 위와 같은 전제조건이 충족된다면 모든 정부 정책사업에서 사회실험을 통한 정책평가가 가능하지만 현실적으로 사회실험은 교육이나 복지사업에서 주로 실행되어 왔다. 〈표 Ⅳ-5〉와 같이 OMB는 미국에서 RCT를 실행했던 대표적인 정책분야를 소개하였는데, 주로 교육, 보건, 복지, 치안분야에서 사용해 오고 있음을 알 수 있다.

30) OMB, "What Constitutes Strong Evidence of a Program's Effectiveness?," 2004, p. 8.

〈표 Ⅳ-5〉 RCT 기반 평가 정책분야

| 정책분야 | 교육 | 보건 | 복지 | 치안 | 재정 | 기타 |
|---|---|---|---|---|---|---|
| 실행횟수 | 3 | 5 | 4 | 3 | 1 | 2 |

출처: OMB, "What Constitutes Strong Evidence of a Program's Effectiveness?," 2004, pp. 9~10.

우리나라의 경우에도 교육, 보건복지, 고용·노동, 문화·체육, 재정, 해외원조 사업 등에서 사회실험의 적용이 가능할 것으로 판단된다. 이들 분야의 재정사업 평가에서 처치집단과 비교집단의 구분이 가능하고 객관적 데이터를 통해 정책의 효과를 명확하게 파악할 수 있기 때문이다. 먼저, 교육 분야의 경우, 학교정책, 교육방법, 학생상담 프로그램에서 실험적 평가방법론을 적용할 수 있을 것으로 보인다. 구체적으로, 최근 시행된 자유학기제나 방과 후 돌봄 서비스와 같은 사업의 경우 시범사업을 실시하여 정책시행 학교와 기존학교를 처치그룹과 비교그룹으로 구분한 후 효과가 있을 경우 전체학교로 확대하는 방법을 검토할 수 있을 것이다.

보건·복지 분야도 RCT와 같은 실험평가를 활발히 적용할 수 있는 분야이다. 각종 보건 제도설계, 보건서비스 및 교육 프로그램, 저소득층 공적부조 및 자활 관련 사업은 무작위 추출에 의한 처치-비교집단의 실험평가 설계가 가능하다. 또한 실제로 해외국가에서는 이 분야에서 가장 활발하게 정책평가가 활성화되고 있다. 적용 가능한 예로서 최근 정책의 효과에 의문이 발생하고 있는 무상보육 서비스의 경우, 보편적 무상보육 외에 차등보육지원 프로그램을 개발하여 차별화된 지원 프로그램의 효과를 비교하는 실험설계를 통해 정책의 효과를 정확히 검증할 수 있을 것이다.

고용·노동 분야 또한 RCT와 같은 사회실험 방식의 정책평가를 유용하게 활용할 수 있는 분야이다. 현재 우리나라는 다양한 고용 및 노동 지원 프로그램을 운영하고 있으며 이 분야의 공공지출은 전체 GDP 대비 0.77%[31]에 이르고 있다. 이러한 높은 지출에도 불구하고 이 분야에 대한 평가는 엄밀

31) 홍승현·원종학, 『적극적 노동정책의 재정효율성 평가방법에 관한 연구』, 2013, p. 30.

하게 이루어지지 않고 있는 편이다. 예를 들어 사업의 효과에 대해 많은 논란이 있는 고용보조 청년 인턴사업의 효과는 RCT와 같은 실험평가로 정확하게 측정할 수 있을 것이다. 청년인턴을 고용하기를 원하는 기업의 신청을 받은 후 최종지원대상 기업을 무작위 추첨으로 선정한 후 선정기업과 미선정기업의 인턴채용비율과 고용증가율을 비교하는 방법으로 사업의 효과를 정확하게 검증할 수 있을 것이다.

문화·체육 분야도 RCT에 의한 사회실험평가를 적절히 활용할 수 있는 분야이다. 해외와는 달리 최근 정부는 국민의 문화생활 향유와 이 분야의 고용확대를 위해 다양한 문화 및 체육 육성 지원사업을 진행하고 있기 때문이다. 국민의 문화 및 체육활동 참가율을 높이기 위한 문화 바우처 사업이나 생활체육 프로그램이 대표적 예다. 예를 들어 이와 같은 문화·체육 분야 정책사업의 효과성을 검증하기 위하여 문화 바우처의 정책수혜자를 무작위 추첨방식으로 선정하여 바우처를 지급받은 사람들과 지급받지 않는 사람들의 문화이용률 수준을 비교하는 방식으로 문화 바우처의 효과를 평가할 수 있을 것이다.

마지막으로 정책금융, 세제, 해외원조 또한 RCT에 의한 진실험평가를 적용할 수 있는 사업분야이다. 이미 해외에서는 이 분야에서 실험을 통한 정책평가를 활발히 진행하였다. 체납자의 납부율을 높이기 위한 체납고지서 메시지 및 디자인 실험은 영국과 미국에서 실시된 이 분야의 대표적인 실험의 예이다. 특히, 다양한 해외원조 사업을 수행하고 있는 UN이나 월드비전과 같은 국제기구들은 개발도상국에 대한 원조의 효과성을 정확하게 측정하기 위하여 RCT에 의한 실험평가를 실시하고 있다. 우리나라의 경우도 한국국제협력단(KOICA)를 통하여 다양한 해외원조 사업을 실시하고 있기 때문에 원조 수원국의 동의를 받아 해외원조 프로그램 수원국에서 무작위 추출 처치–비교집단 실험평가[32]를 통해 특정 원조사업의 효과를 정확히 검증할 수 있을 것이다.

32) RCT에 의한 실험평가 가능성을 탐색하기 위해 KOICA는 MIT 대학의 J-Poverty Action Lab를 통해 RCT 관련 단기 교육연수를 진행하였다.

〈표 Ⅳ-6〉 RCT에 의한 사회실험이 적용 가능한 정책분야

| 분야 | 적용가능 사업 분야 | 사업의 예 |
|---|---|---|
| 교육 | 학교 관련 정책<br>교수법 및 교육 방식<br>학생 상담 프로그램 | 방과 후 돌봄 서비스<br>자유학기제<br>원어민 교사 |
| 보건 · 복지 | 건강 관련 제도설계,<br>건강서비스 및 교육프로그램<br>저소득층 공적 부조<br>자활지원 사업 | 금연교육 프로그램<br>무상보육사업 |
| 고용 · 노동 | 고용 및 실업 보조사업,<br>직업훈련 사업<br>취업알선 사업 | 실업자 훈련 프로그램<br>청년인턴 프로그램 |
| 문화 · 체육 | 문화프로그램 지원 사업<br>국민체육 활성화 사업 | 여행, 공연 바우처 사업<br>생활체육 지원 사업 |
| 재정지원 | 투융자 등 정책금융<br>세금고지 및 체납 관련 사업 | 중소기업 창업자금 지원<br>세금고지서 차등구성 |
| 해외원조 | 해외원조 프로그램 | ODA 수원국 내의 실험평가 |

전술한 것처럼 다양한 분야에서 사회실험이 가능하지만 몇몇 정책분야는 사회실험이 어렵거나 불가능하다. 특정 정책분야는 무작위 추출을 통한 처치-비교집단 설정이 불가능하거나 처치의 효과를 측정하는 것이 매우 어렵기 때문이다. 〈표 Ⅳ-7〉과 같이 실험평가가 어려운 분야는 SOC, 국방, 안전 · 재해, R&D 분야이다. 먼저 건설사업이 주된 사업이 되는 SOC 분야는 건설공사에 들어가는 높은 비용으로 인해 비교집단을 구성하기 어렵다. 따라서 정책평가는 건설로 발생하는 편익을 추정하고 이를 비용과 비교하는 B/C 분석이 많이 이용된다. 국방정책도 비교집단의 구성이 어려운 정책영역이다. 군사작전의 경우, 정책결정의 시급성으로 인해 비교집단 구성을 통한 실험이 원칙적으로 불가능하며, 무기구매 및 개발의 경우도 소용 비용이 매우 높기 때문에 비교집단을 설정하기가 어렵다. 안전 · 재해 분야의 경우도 처치-비교집단 구성을 통한 실험평가가 불가능한 정책영역이다. 이론상으로 재해피해에 대한 보상을 처치집단과 실험집단에 달리하여 상이한 재해복구

정책의 효과를 측정하는 것이 가능할지라도 재해로 고통 받고 있는 주민들을 차등지원하거나 실험대상으로 삼았다는 윤리적 비난을 극복하기 어렵기 때문이다. 마지막으로 R&D 분야의 경우도 실험을 통한 정책평가가 어려운 분야이다. R&D의 특성상 성과창출에 오랜 시간이 걸리기 때문에 처치의 성과를 측정하기 어렵다.

〈표 Ⅳ-7〉 RCT에 의한 사회실험이 적용 불가능한 정책분야

| 분야 | 프그로램 | 장애 및 불가요인 |
|---|---|---|
| SOC | 도로, 철도, 항만 등 사회기반시설건설 | 비교집단 설정불가 |
| 국방 | 군사작전, 무기개발 및 구매 | 비교집단 설정불가 |
| 안전재해 | 재해지원금 | 집단 간 다른 처치 불가 |
| R&D | 원천/응용기술개발 | 성과측정의 어려움 |

## 3. 사회실험의 장애요인과 대처방안

전술한 바와 같이 다양한 정책분야의 사전 및 사후평가에서 RCT와 같은 사회실험을 통한 정책평가가 가능할지라도 현실에서는 여러 가지 장애요인 때문에 실험을 원활하게 실시하기 어려운 경우가 많다. 이런 장애요인은 사회실험을 둘러싼 제도적, 윤리적, 실험적 한계에서 발생한다. 이런 이유 때문에 우리나라에서는 영국이나 미국과 달리 사회실험이 활성화 되어 있지 않다. 따라서 이러한 장애요인을 극복할 수 있는 실질적 대처방안들을 제시하는 것이 사회실험을 통한 정책평가를 활성화시킬 수 있는 방법이 될 것이다.

### 가. 제도적 장애요인

우리나라에서 사회실험이 활성화되지 않은 근본적인 이유는 사회실험을 둘러싼 제도적 장애요인이 매우 많기 때문이다. 이러한 제도적 장애요인으로 들 수 있는 것이 법적규정의 미비이다. 우선 재정사업 성과평가의 총괄규

정이 포함되어 있는 「국가재정법」에는 사회실험에 대한 규정이 없다. 또한 사회실험을 적용할 수 있는 개별 시범사업의 경우에도 사회실험을 적용할 수 있는 평가 및 환류에 관한 법적 규정이 거의 없는 실정[33]이다. 따라서 일정규모 이상의 특정 분야 재정사업의 효과성을 사회실험을 통해 검증하는 규정을 「국가재정법」과 각 개별 사업 법령을 통해 마련할 필요가 있다.

다른 제도적 장애요인으로서 정부 내에 RCT 방식의 효과에 대한 충분한 이해를 가지고 있는 전문가가 거의 없다는 점을 들 수 있다. 실험방식의 정책평가에 대해 윤리적 문제와 같은 실행절차의 어려움 때문에 막연히 거부감을 표시하는 경우가 많다. 따라서 사회실험을 포함한 정책평가방법에 대한 교육을 확대하여 RCT의 장단점에 대해 정책담당자들이 충분히 인지하게 하는 노력이 필요하다. 또한 순환보직제도의 특성상, 사업담당 공무원은 사업의 정확한 효과가 나타나기 전에 자리를 옮길 확률이 높다. 따라서 장기적인 시각을 가지고 사업에 대한 엄밀한 평가를 통해 사업을 개선하려는 유인이 없다. 더구나 잦은 이동 때문에 사회실험과 같은 전문적인 성과평가제도에 대한 전문성을 쌓을 기회도 적다. 정책적 중요도가 높거나 예산이 많이 들어가는 대규모 사업의 경우 순환주기를 연장하여 사업관리자가 장기적 안목을 가지고 사업을 평가하는 문화가 정착될 때 사회실험과 같은 과학적이고 장기적인 평가시스템이 활성화될 수 있다.

마지막으로 제도적 장애요인은 사회실험에 들어가는 관련 자료의 부족과 이로 인한 비용의 문제이다. 사회실험의 효과를 측정하는데 많은 자료가 필요하다. 그러나 기존 행정자료가 구축되어 있지 않을 경우, 자료를 생성하는데 많은 비용이 들어간다. 이런 이유로, 전술한 것처럼 미국은 저비용 RCT를 활성화하기 위해 다양한 자료공유 프로그램을 두고 있다. 우리나라의 경우 부처 간 자료공유가 되지 않고 있기 때문에 사회실험의 보급과 확산에 어려움이 있다. 따라서 부처 간 자료를 공유할 수 있는 시스템을 만들고 평가와 관련된 자료의 공개와 공유의무를 법적으로 제도화할 필요가 있다.

---

33) 「보건의료 기본법」 제44조는 시범사업의 평가 및 환류규정까지 포함하고 있다(이삼열 외, 2009)

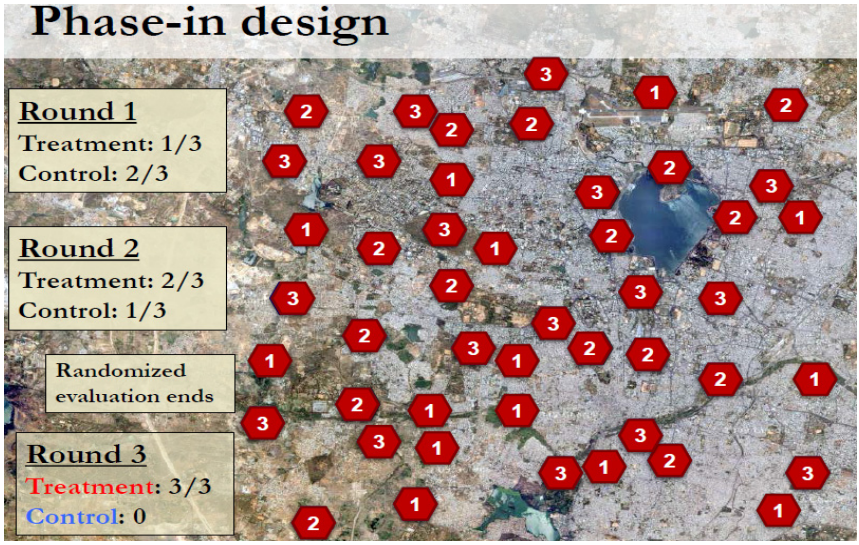
### 나. 윤리적인 문제

사회실험에 대한 가장 많은 문제제기가 윤리적인 문제이다. 윤리적인 문제는 사람이나 사람이 집합체인 조직을 실험이나 처치 대상으로 하는 데서 발생한다. 이러한 윤리적인 문제는 특정 개인이나 집단이 처치집단이 되거나 처지집단에 포함되지 않아 정책의 수혜자가 되지 않았을 때 발생한다. 처치집단에 포함되어 발생하는 윤리적인 문제는 인체에 부작용을 줄 수 있는 신약개발과 같은 자연실험에서 주로 발생한다. 특정 사업이나 제도를 가지고 시행하는 사회실험에서는 비교집단의 개인이나 조직이 정책의 수혜를 받지 못할 때 윤리문제가 생긴다. 비교집단에 속한 특정 개인이나 조직이 정책과 제도가 모든 대상자에게 평등하게 적용해야 한다고 주장할 때 문제가 발생하기 때문이다. 특히 사업대상자 선정 규정이나 시범사업 대상자에 대한 명확한 규정이 없을 때는 정책대상자 선정에 논란이 발생할 수 있다. 따라서 실험방식의 정책평가를 염두에 둔다면 사업시행 전에 사업의 대상자와 평가에 관한 규정을 명확히 할 필요가 있다. 이러한 규정의 예로서, 해외국가에서 활성화되어 있는 IRB(Institutional Review Board)를 제도화하는 방법을 검토해 볼 수 있다. IRB는 실험방식을 포함한 모든 연구의 참가자들을 윤리적으로 보호하기 위한 제도적 장치로서 연구의 절차에 법적, 윤리적 문제점이 있는지를 사전에 심사한다. 실험평가 전에 실험의 내용과 방법을 IRB를 통해 승인받게 한다면 윤리적인 문제를 어느 정도 사전에 점검할 수 있을 것이다. 또한 IRB 승인으로 실험평가의 신뢰성을 높여 실험에 대한 거부감을 줄이고 참가율을 높일 수 있다.

윤리문제 해결을 위한 다른 대안으로, 시차를 두고 모든 정책대상자에게 사업을 진행하는 단계별 처치(Phase-in)나 교대처치(Rotation) 실험설계방식을 검토할 수 있다. [그림 Ⅳ-3]에서 보듯이 단계별 처치 실험방식은 시차를 두고 비교집단에 속했던 그룹이 처치집단이 되어 사업의 적용을 받게 되는 방식이다. [그림 Ⅳ-4]처럼 교대처지실험방식은 실험종료 후 처치집단과 비교집단을 교대하여 비교집단에게 사업을 적용하는 실험방식을 말한다. 단계별 또는 교대처치 실험설계는 비용이 많이 들고 사업의 효과가 없더라도 비교

집단에 사업을 시행해야 하는 단점이 있다. 그러나 이런 단점에도 불구하고 사회실험에 대한 윤리적인 문제제기가 크거나 실험의 평가결과를 가지고 사업을 종료하거나 변경해야 할 필요성이 클 때 사용할 수 있는 방법이다.

[그림 Ⅳ-3] 단계별 처치 실험설계(Phase-in Design)

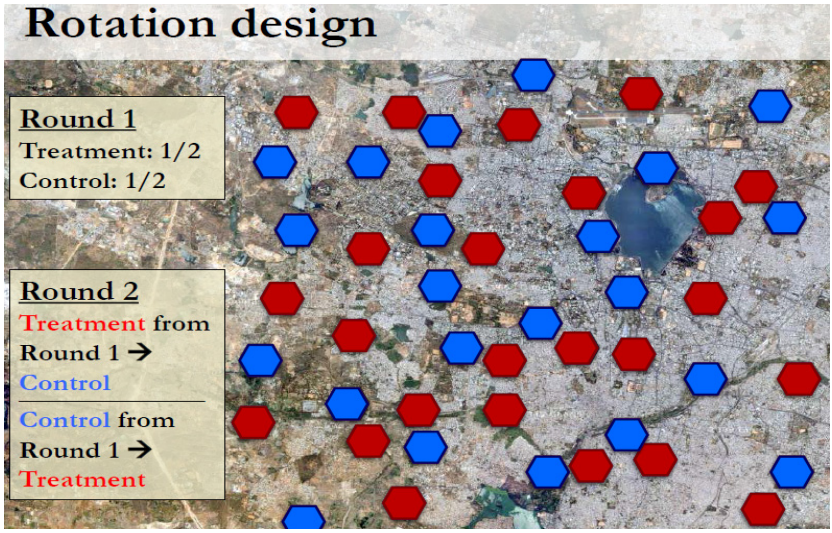



출처: J-PAL RCT Executive Course Lecture Notes

### 다. 평가과정의 오류

RCT와 같은 진실험 방식의 사업평가방식이 준실험이나 비실험방식의 다른 평가방식보다 내적인 타당성(Internal Validity)이 높다고 할지라도 여전히 실험과정에서 오류(Bias)가 발생할 가능성 있다. 이런 오류는 실험평가의 내적인 타당성을 떨어뜨려 사회실험 평가를 보급하는 데 장애요인이 될 수 있다. 내적인 타당성을 위협하는 오류의 종류에는 선택오류(Selection Bias), 실험이탈(Attribution), 실험의 전이(Spill over Contamination), 실험 오류, 실험환경의 변화 등이 있다.

[그림 Ⅳ-4] 교대처치 실험설계(Rotation Design)




출처: J-PAL RCT Executive Course Lecture Notes

(1) **선택오류**

선택오류는 실험의 내적 타당성을 위협하는 대표적인 오류로서 처치집단과 비교집단의 동질성이 확보되지 않았을 때 발생한다. 두 집단 간의 동질성이 확보되지 않는다면 정책의 처치효과가 아니라 집단의 특성의 차이에 의해 정책효과가 발생할 수 있기 때문이다. 예를 들어, 특정 수업 프로그램의 학생의 학업성취도 개선효과에 대한 실험에서 처치집단에 부모가 부자인 학생들의 비율이 높다면 프로그램의 효과가 아닌 부모의 경제력이 학생들의 성적향상에 영향을 끼칠 가능성이 높기 때문이다. 선택오류는 무작위 추출(Randomized Sampling)이 이루어지지 않을 때 발생하는데 정부정책에서 정책대상자 집단을 처치집단과 비교집단으로 나누어 선발하기는 쉽지 않다.

이런 장애요인을 극복하는 방법으로 정책대상자를 추첨(Lottery)으로 선발하는 방법을 검토할 수 있다. 시범사업의 경우, 전체 정책대상자 내에서 시범사업 대상자를 추첨으로 결정한 후 사업의 수혜를 받는 사람들을 처치집단으로 하여 평가를 진행할 수 있다. 본사업 평가의 경우도 전체 사업 대상

자에 대한 신청을 받은 후 수혜자로 추첨된 신청자를 처치집단으로, 탈락한 신청자를 비교집단으로 설정하여 실험방식의 평가를 진행할 수 있다. 이와 같은 추첨방식은 모든 대상자에게 정책을 시행하는 보편적 사업에는 적용할 수 없고, 일부 대상을 선정하여 실시하는 시범사업이나 예산제약으로 정책대상자 일부에게만 사업시행이 가능한 본사업에서 사용할 수 있다.

(2) **이탈**(Attribution)

RCT와 같은 실험적 정책평가방법에서 나타날 수 있는 또 다른 문제는 정책실험 시행 중 정책대상자의 이탈이다. 이런 이탈현상은 실험의 처치집단이 되는 정책대상자를 강제배정이 아닌 임의추첨으로 정했기 때문에 발생한다. 이탈의 폭이 작을 때는 문제가 되지 않지만 매우 클 때는 실험의 정확성에 오류가 발생할 가능성이 높아진다. 이런 오류를 해결하는 방법으로 ITT(Intention to Treat)와 ToT(Treatment on the Treated)를 사용할 수 있다. ITT는 이탈자가 적을 때 사용하는 방법으로 이탈자를 고려하지 않고 실험의 효과를 추정하는 방법이다. 실제 정책과정에서 정책대상자가 이탈하는 것은 자연스러운 현상이기 때문에 정책 실험과정에서의 탈락도 자연스런 과정으로 간주하고 처치의 효과를 계산한다. 〈표 Ⅳ-8〉에서 ITT에 의한 정책효과는 실험종료 후 처치집단 45명의 결과값에서 비교집단 47명의 결과값의 차감하여 구한다.

〈표 Ⅳ-8〉 'Intention to Treat(ITT)'의 예

| | 처치집단(Treatment) | 비교집단(Comparison) |
|---|---|---|
| 실험시작 | 50 | 50 |
| 실험종료 | 45 | 47 |
| 정책효과 | ITT = Y(T)-Y(C) | |

ToT는 정책실험 도중 이탈자가 많이 발생할 때 사용한다. 많은 이탈자로 인하여 실험의 효과를 정확히 추정할 수 없을 때, 실제 실험으로 발생한 정

책효과(ITT)를 처치집단과 비교집단의 참가자가 처지를 받을 확률로 나눈 값으로 정책효과를 추정하는 방식이다.

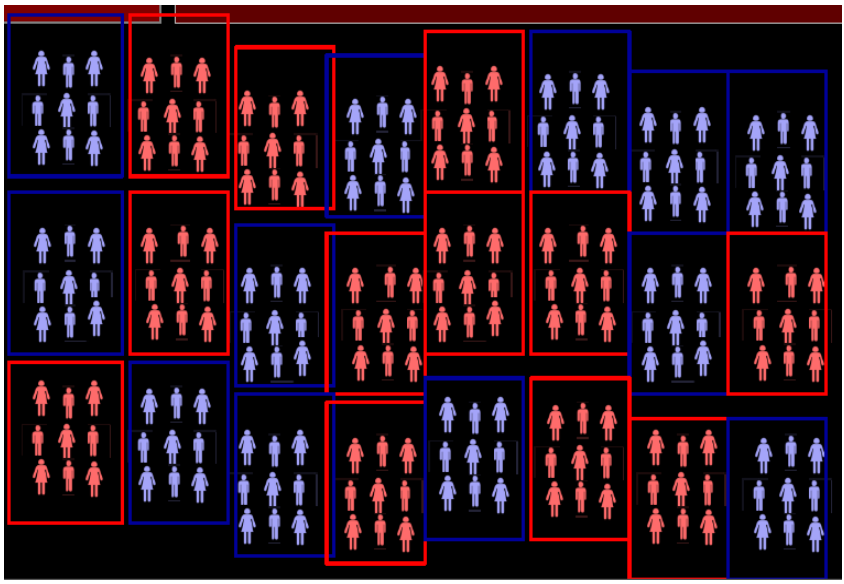
〈표 Ⅳ-9〉 'Treatment on the Treated(ToT)'의 예

| | 처치집단(Treatment) | 비교집단(Comparison) |
|---|---|---|
| 실험시작 | 50 | 50 |
| 실험종료 | 38 | 42 |
| 정책효과 | ToT = (Y(T)-Y(C))/(Prob[treated/T]-Prob[treated/C]) | |

(3) **전이 또는 오염**(Spillover or Contamination)

사회실험과정에서 가장 흔히 발생하는 내적 타당성의 위협요인은 처치의 전이 또는 오염이다. 이는 처치집단에게만 실시한 처치의 효과가 비교집단에 전이되어 처치의 효과를 오염시킬 때 생긴다. 예를 들어 새로운 학습방법에 대한 정책실험을 같은 학교에서 처치집단과 통제집단으로 나누어 진행할 때 통제집단의 학생들이 처치집단의 학생들과 교류하여 새로운 학습방법을 습득할 때 발생한다. 전이 또는 오염에 의한 오차를 방지하기 위한 방법으로는 군집 무작위 추출(Cluster Random Sampling)이나 차단(Blocking) 실험설계방식을 사용할 수 있다. 군집 무작위 추출이란 개인을 단위로 무작위 추출을 실시하지 않고 집단별로 무작위 추출을 실시하여 개인 간의 접촉으로 인한 처치효과의 전이나 오염을 막는다.

[그림 Ⅳ-5] 군집 무작위 추출(Cluster Random Sampling)의 예



출처: J-PAL RCT Executive Course Lecture Notes

이를 전술한 학습방법의 효과에 대한 평가를 위해 적용할 경우 학생 개인이 아니라 학교별로 무작위 추출을 하는 것이다. 단점은 군집단위로 추출을 해야 하기 때문에 실험설계가 복잡하고 샘플의 수가 많아져 비용이 많이 드는 점이다. 또 다른 방법은 차단 실험설계방식을 들 수 있다. 이 설계방식은 차단된 집단에서 한 사람씩 정책대상자를 선정한 후 대상자를 처치집단과 비교집단에 배정하는 방식이다. 대상자가 서로 다른 집단에서 선발되었기 때문에 대상자 간의 교류가능성이 작아 처치의 전이나 오염이 발생할 가능성이 줄어든다.

### (4) 실험오류

실험과정의 특성 자체로 인하여 오류가 발생할 수 있다. 이러한 오류는 실험참가자의 의도적 행동변화와 실험으로 인한 비의도적인 환경변화에 의해 발생한다. 먼저 실험참가자의 행동변화는 처지집단에 속한 참가자들의

호손효과(Hawthorn Effect)와 통제집단에 속한 참가자들의 존 헨리효과(John Henry Effect)가 있다. 호손효과는 처치집단의 참가자들이 정책이 기대하는 효과에 맞추어 행동을 변화시키는 것을 말한다. 반대로 존 헨리효과는 비교집단에 속한 참가자들이 처치집단보다 더 많은 노력을 할 때 발생한다. 예를 들어 새로운 학습방법의 학업성적에 대한 효과를 무작위로 처치학교와 비교학교로 나누어 정책실험을 실시할 때, 호손효과는 처치학교에 속한 학교들이 정책의 효과를 높이기 위하여 학업성적 향상에 더 많은 노력을 기울임으로써 발생한다. 반대로, 존 헨리효과는 비교집단에 속한 학교들이 처치집단의 학교에 뒤처지지 않기 위하여 학업성적 향상을 위해 부가적으로 노력을 기울임으로써 학업성적이 향상될 때 나타난다. 이러한 참가자의 의도적 행동변화는 실험의 대상자가 스스로 실험의 대상이 되고 있음을 인지하여 의식적으로 행동을 변화시킬 때 발생한다. 이와 같은 실험오류는 정책효과에 의해 부가적인 혜택이나 손해가 발생하는 교육정책의 실험상황에서 많이 발생한다. 따라서 이런 오류를 줄이기 위해서는 가급적 실험상황을 사전에 고지하지 않음으로써 참가자들이 최대한 자연스러운 상황에서 정책실험에 참가할 수 있도록 유도해야 한다.

실험으로 인한 환경변화도 정책효과의 정확한 추정을 어렵게 만드는 장애요인이다. 이는 정책실험으로 인하여 정책효과에 영향을 끼치는 환경이 의도하지 않게 변화할 때 발생한다. 예를 들어, 순찰의 범죄예방에 대한 효과를 측정하는 정책실험을 위해 순찰구역을 무작위로 처치구역과 비교구역으로 나누어 처치구역에만 순찰을 실시할 경우, 범죄자들이 비교구역에 몰려 정책의 정확한 효과를 추정할 수 없게 된다. 의도하지 않은 환경변화에 대처하기 위해서는 다중처치실험과 같은 설계를 실시할 수 있다. 순찰 유무의 효과보다는 순찰의 시간대, 순찰인원 수, 순찰방법별로 다른 종류의 정책 처치를 실시하여 그 효과를 비교할 수 있다.

### 라. 외적인 타당성(External Validity)

마지막으로 외적인 타당성(External Validity)은 정책실험의 장애요인이 될 수 있다. 외적 타당도란 연구결과가 일반화될 수 있는지를 나타낸 정도를 의미한다. 외적인 타당성은 실험대상집단(처치 및 통제집단)의 성격이 모집단의 성격과 유사할 때 확보될 수 있다. 그러나 실험환경을 둘러싼 특수성 때문에 실험결과를 일반화하기 어려운 경우가 발생한다. 실제로 같은 정책수단을 가지고 다른 실험환경에서 실시한 실험결과가 상이한 경우가 많이 발생한다. 특히, 문제가 되는 것은 시범사업의 연구결과를 본사업으로 확대 시행했을 때 시범사업의 효과가 본 사업에서 유사하게 나타날 수 있는가가 논란이 된다.

이러한 문제를 해결하기 위해서는 시범사업의 경우 시범사업 대상자가 본 사업 대상자의 전체표본을 대표할 수 있도록 정확하고 충분하게 선정해야 한다. 이를 위해, 대상자를 층화 추출방식(Stratified Sampling)을 통해 정확히 선별하고 통계적 정당성을 확보할 수 있도록 충분한 수의 표본을 선정할 필요가 있다. 또한 표본의 대표성이 확보되었다 할지라도 정책효과의 보다 정확한 추정을 위해 다른 연구결과를 참고하거나 실험참가자에 대한 질적 연구, 준실험적 또는 비실험적 정책평가결과와의 비교를 통하여 정책효과가 유사하게 나올 때 본사업의 확대시행을 검토해야 한다.

# Ⅴ. RCT를 통한 실제 정책평가 설계 사례

본 장에서는 RCT 또는 그와 유사한 준실험 평가방식으로 평가를 진행한 국내외 사례를 선정하고 시사점을 간략히 제시하였다. 국내사례는 희망리본 사업 사례로서 RCT가 완벽히 적용된 사업은 아니지만 처지집단과 통계적으로 인구통계학적 차이가 없는 비교집단을 구성함으로써 비교한 준실험적 방식의 평가사례이다. 해외사례는 교육정책평가에서 엄격한 RCT 평가방법론을 준수하고 있는 Institute of Education Science의 보충수업효과에 대한 RCT 평가와 국내 정책평가에의 시사점을 제시하고 있다.

## 1. 희망리본 사업 사례

우리나라에서는 시범사업이라는 수단이 본사업을 시작하기 전에 사업의 효과성을 검증하는 수단으로 활용되었다고 볼 수 있다. 소개하고자 하는 희망리본 사업은 신규사업 도입을 위한 사전검증의 목적으로 시도된 사업이라기보다는, 기존에 시행되고 있는 사업의 효과성을 개선하기 위해 시도한 새로운 사업 방식의 검증에 목적을 둔 시도였다. 그러나 본질적으로는 새로운 형태의 사업을 추진하기 전에 시범적으로 사업 모형을 테스트하고 검증하려는 의도였다는 점에서, RCT를 통한 신규사업 사전검증이라는 목적과 동일한 의도를 가지고 있다.

본 연구에서는 시범사업으로서 추진된 희망리본 사업에서 사업평가가 어떤 방식으로 추진되었으며, 우리나라의 현실적 제약조건하에서 원래 의도한 바와 다르게 사업평가가 구현된 원인에 대해 살펴보고자 한다. 시범사업을 통해 본사업화의 가능성을 점검하고 주요 쟁점을 확인하며 대응책을 마련하

여 본사업화를 도모한다는 이상적인 계획을 가지고 출발했다는 점에서 희망리본 사업은 RCT를 통한 사업의 사전평가라는 본 연구와 의도를 같이 한다. 다만 RCT가 가능한 사업이었음에도 불구하고, 차선의 방법인 준실험 방법을 통해 사업평가를 진행했다는 한계를 가지고 있다.

기초수급권자를 포함하는 취약계층에게 제공되는 근로연계형 복지사업인 희망리본 사업은, 기존의 자활사업의 한계를 극복하고자 새로운 형태의 시범사업으로 추진되었다. 그리고 새로운 형태의 사업의 효과성을 검증하고, 본 사업화 과정에서 발생할 수 있는 정책적 쟁점을 사전에 확인하고 대책을 마련하고자 하는 취지에서, 제한된 지역에서 시범사업으로 4년간 시행되었다.

시범사업 추진 초기에는 제대로 된 RCT 방식을 적용한 평가를 통해, 이 시범사업을 둘러싼 논쟁에 객관적 증거로써 대처하고자 하는 의도를 가지고 있었다. 그러나 시범사업 도입 과정에서 정책 담당자의 교체와 재원확보의 미비로 인해, 준실험적인 방식의 성과평가가 이루어졌다. 본 고에서는 희망리본 사업 성과평가 실행 과정의 현실적 장애를 살펴보고 RCT 제도화의 필요성을 논하고자 한다. 그리고 희망리본 사업에 RCT를 적용했을 경우, 어떻게 사업을 기획하고 어떠한 쟁점이 발생하고 다루었을지에 대한 가상적인 시나리오도 제시하고자 한다. RCT 적용의 가상 시나리오를 보여준다는 의미가 있다.

### 가. 희망리본 사업의 도입 배경

희망리본사업은 2000년대 초반에 도입된 자활사업의 개선 방안을 점검하기 위해 시범사업으로 추진되었다. 자활사업은 "생산적 복지"를 정책 목표로 하여 「기초생활보장법」과 연결된 사업으로 도입되었다. 엄밀하게는 1996년 말에 자활사업 자체로 산발적으로 추진되어 왔으며, 2000년의 「기초생활보장법」의 도입과 더불어, 전국적인 사업으로 확대되었다. 기초생활보장법에 의해 기초생활수급자에게 소득 보조를 하는 대신, 자활사업에 참여하여 근로 의무를 수행하도록 한 것이다.

자활사업의 목표가 무엇인지에 대해서는 논란의 여지가 있다. 자활센터에 출석하여 근로에 참여하고 공동사업에 참여하는 것 자체가 목표인지, 아니면 자활사업을 통해 경제적 자립을 이루는 것이 목표인지에 대해, 사업 내부 이해관계자와 외부 이해관계자 사이에 괴리가 있다. 생산적 복지의 정책 목표에 비추어 보면, 자활사업에 참여하여 경제적으로 자립하는 것이 사업의 목표임을 부정하기는 어렵다고 판단된다. 사업 참여자의 경제적 자립이 정책 목표라면, 중요한 성과지표는 사업 참여자의 탈수급율이다. 탈수급율이라는 기준으로 보면, 기존의 자활사업의 성과가 지극히 미약하다고 알려져 있다.

이러한 사업 성과의 부진에는 사업추진 방식의 문제점이 배경에 깔려있다. 전국의 시군구 마다 자활센터를 지정하고, 매년 고정액에 가까운 예산을 배정하는 체제로 운영되고 있다. 지역자활센터에 대한 평가가 있기는 하지만, 실질적으로 평가에 의한 환류체제에는 미치지 못하는 한계를 가지고 있다. 이런 사업추진 방식으로 인해, 서비스 제공 기관 간의 경쟁이나 서비스 제공기관에 대한 성과관리가 거의 작동하지 않고 있다는 문제가 발생하고 있다.

이러한 기존 자활사업 추진 방식에 대한 비판에 대응하여, 2009년에 민간 서비스 수행기관을 공모하고 성과계약을 도입하는 시범사업을 추진하게 된다. 이 시범사업의 희망리본 사업이며, 희망리본 사업의 원래 취지는 기존 자활사업의 운영방식을 개선하는 방안을 테스트하는 것이었다. 희망리본사업의 성과를 객관적으로 검증하고, 새로운 사업방식의 위험요소를 사전에 파악하여 본 사업화를 추진하자는 취지로 시범사업이 추진되었다.

### 나. 희망리본 사업의 성과평가 방법과 성과

희망리본 사업의 추진 방식은 다음과 같은 특징을 가지고 있다. 첫째, 사업 수행기관을 선정하기 위해 공모 과정을 거쳤다. 비영리 기관 뿐 아니라 영리기관도 참여할 수 있도록 하였다. 둘째, 성과계약을 도입하여, 명시적으로 취업률, 취업유지율, 탈수급율에 대응하는 성과급을 제시하였다. 사업 대

상은 확대하여, 기존 자활사업 참여 대상 수급자 뿐 아니라 차상위 계층까지도 포함하였다. 그러나 여전히 기초생활 수급자가 차지하는 비율이 80% 이상으로서 대다수를 차지하였다.

희망리본 사업의 성과는 다음과 같다. 2009년에는 이미 시행된 자활사업들의 취업률과 탈수급률은 각각 13.7%, 7.7%이다. 반면에 희망리본 사업의 경우, 45.4%, 11.7%에 이르고 있다. 취업률은 3배의 정도의 차이가 있으며, 희망리본 사업은 탈수급율에 있어서 1.5배 이상의 성과를 보이고 있다.[34]

이러한 성과의 차이는, 시범사업 2차 연도가 되면 더욱 증폭된다. 그 이유는 서비스 제공 기관들의 사업 노하우가 향상되었기 때문이라고 판단된다. 기존 자활사업들의 취업률은 15.8%, 탈수급률은 7.2%로 파악되고 있다. 희망리본 사업의 경우, 각각 46.2%, 22.8%에 이른다. 희망리본 사업의 취업률이 기존 자활사업과의 3배에 이르고, 탈수급률도 3배 차이를 보이고 있다.

이러한 성과의 차이에 대해 반론이 있을 수 있다. 희망리본 사업의 성과가 우월하다는 점에 대한 반론으로서, (1) 희망리본 사업의 사업참여자가 취업능력에 있어서 더욱 우월하다, (2) 희망리본 사업의 목적이 취업률과 탈수급율로 명확히 제시된 반면 자활사업의 경우 분명하지 않았다 , (3) 희망리본 사업의 비용이 더욱 비싸다 등이 있을 수 있다. 희망리본과 자활사업의 사업 참여자가 서로 다르다는 주장은 일부 일리가 있을 수 있으나, 두 사업의 뚜렷한 성과 차이를 모두 설명하기는 어렵다. 희망리본 사업에 대한 지난 4년간의 평가 보고서[35]에 따르면, 조건부 수급자, 조건부과 제외자 등이 사업 참여자 중에 차지하는 비율이 자활사업과 유사하다고 보고되고 있다.

사업의 목표 자체가 희망리본 사업의 경우 명시적으로 사업 수행기관에게 공유된 반면, 자활사업의 경우는 모호하였다는 주장도 일부 타당성이 존재하지만 사업성과가 차이가 발생하는 것을 모두 설명한다고 보기는 어렵다. 희망리본 사업의 평가 보고서는, 자활사업 중에서 사업의 목적이 경제적 자

34) 희망리본사업의 성과는 참여자 중 수급자만을 대상으로 한 수치이다. 실적 자료는 노대명(「희망리본 시범사업과 자활사업 연계방안」, 2012), 박노욱 외(「희망리본 사업의 본사업화 방안」, 2012)에서 왔다.

35) 한국조세연구원과 보건복지부 중앙자활센터에서 평가 보고서 발간

립이라고 비교적 명확한 시장진입형 사업과 희망리본 사업의 성과 차이가 통계적으로 유의하게 있다는 점을 보여주고 있다.

그 뿐 아니라 자활사업과 희망리본 사업의 목표에 차이가 있다는 주장이 취업률 기준으로는 타당할 수 있으나, 탈수급률 기준으로는 타당하지 않다. 왜냐하면 자활사업과 희망리본사업의 궁극적인 정책 목표는 참여자들의 경제적 자립이기 때문이다. 사업 참여자들의 경제적 자립이라는 정책 목표 기준으로 보면, 취업연계에 초점을 둔 희망리본사업이 자활기업 생성에 초점을 둔 자활사업보다 효과적임을 시사한다.[36]

참여자 1인당 소요된 비용에 차이가 있다는 비판도 표면적으로는 정당하지만, 직접 사업 비용 뿐 아니라 사업으로 인한 탈수급 효과까지 고려한다면, 희망리본사업의 비용이 아주 크다고 보기는 어렵다. 자활사업 임금 일부분이 생계급여를 절감하는 부분을 고려하면, 자활사업의 1인당 연비용(2011년 기준)[37]은 약 267만원으로 추정된다. 반면에, 희망리본 사업의 1인당 연비용은 약 358만원으로 추정된다. 탈수급에 따른 기초생활 급여가 절감되는 부분을 고려한다면, 희망리본 사업이 경제적으로 비효율적인 사업이라고 보기 어렵다.[38]

희망리본 사업의 성과를 요약하면, 희망리본 사업의 성과가 취업률이나 탈수급 기준으로 보면 자활사업보다 높으며, 그 원인을 사업 참여자의 직업

---

36) 경제적 자립 효과를 보다 종합적으로 판단하기 위해서는 단기적인 효과뿐 아니라, 중장기적인 효과도 동시에 고려해야 할 것이다. 현재 시점에서 사업의 성과를 본다면, 적어도 단기적으로는 취업연계 사업이 효과적이라는 점은 분명하다. 희망리본 사업의 중장기적 성과를 판단하기는 이른 시점이기 때문에, 중장기적 성과 비교는 어렵다. 다만, 기존 자활사업의 경우, 중장기적으로도 대외적으로 보여줄 만한 성과를 창출하고 있지는 못하다는 한계를 가지고 있다.

37) 노대명(「희망리본 시범사업과 자활사업 연계방안」, 2012), 박노욱 외(「희망리본 사업의 본사업화 방안」, 2012)

38) 취업성공 패키지 사업의 1인당 비용은 기존 자활사업과 희망리본 사업의 중간에 위치하고 있다고 추정된다. 취업성공 패키지 사업의 경우, 사례관리자 1인이 연간 100명에게 서비스를 제공한다는 기준으로 예산이 책정되고, 희망리본 사업의 경우는, 연간 40-50명에게 서비스를 제공한다는 기준으로 운영되고 있다. 중장기적인 성과의 차이를 고려했을 경우 희망리본 사업이 취업성공패키지 사업보다 비용이 더 매우 비싸다고 단정하기 어려운 측면이 있다.

능력의 차이, 사업 목적 자체의 차이, 참여자 1인당 재정투자의 차이 등 만으로는 설명하기 어렵다는 것이다. 성과의 상당 부분은 사업 수행기관의 공모와 성과계약 방식의 도입을 통한 사업 목적의 명확화와 그에 따른 서비스 제공 기관들의 서비스 향상을 통해 이루어진 것으로 판단된다. 「기초생활보장법」의 급여구조 자체가 가지는 정책 환경의 한계가 바뀌지 않았지만, 사업 추진방식을 바꾸는 것만으로도 성과를 제고할 수 있었다는 점에서 정책적 시사점이 크다고 판단된다.[39]

이상이 시범사업인 희망리본 사업, 기존 자활사업과 노동부에서 시행한 취업성공 패키지 사업과의 비교 분석을 통해 성과를 평가한 주된 내용이다. 실제 4년간의 시범사업 과정에서 생산된 평가 보고서에는 기존 자활사업 중 희망리본 사업과 성격이 가장 유사하다고 판단되는 시장진입형 자활사업의 참여 대상자와의 비교 분석을 통해, 시범사업과 기존 사업과의 성과를 비교분석하였다. 그리고 두 집단 간의 이질성 여부를 확인하기 위해 인구학적 특성을 비롯한 소득, 학력 등을 포함해서 취업률이나 탈수급률에 영향을 미칠 개연성이 있는 요소들을 통제하여 분석하였다. 일종의 준실험적인 평가 방식을 채택한 것이다.

### 다. 시범사업으로서의 희망리본 사업의 한계와 시사점

기존 자활사업의 개선 모형으로서 성과관리형 사업인 희망리본 사업을 시범사업으로 추진하였다. 그리고 원래 계획으로는 무작위 추출로 비교집단과 실험집단을 선정하여 사업의 효과성을 분석하려고 하였다. 이러한 취지로 외국과 국내의 연구기관이 협업하여 시범사업을 디자인하고 평가를 하고자 하는 시도도 있었다. 그러나 실제적으로는 RCT 평가 방식이 시도되지 못했고, 가장 유사한 목적을 가진 기존 자활사업 유형의 사업 참여자와 비교 평가하는 방식으로 사업의 성과평가가 이루어졌다. 시범사업 초기에 사업의 효과성이 뚜렷이 드러나는 상황이 전개되어 사업의 효과성 자체를 엄밀히

39) 희망리본 시범사업 후, 보건복지부는 탈수급을 유인하기 위하여 급여구조를 전환하는 일을 시행하였다.

검증하는 것은 쟁점이 아니게 되었다. 희망리본 사업 자체만으로 볼 때는, RCT 평가 방식의 적용 여부가 중요한 쟁점이 아닌 것으로 되었다는 것이다. 다만 본사업화 시 고려해야 할 정책적 쟁점들은 부각되었으며, 본사업화 시 계약 구조의 개선이 이루어졌다.

이런 희망리본 사업 자체에 있어서 RCT 적용 여부는 현실적 쟁점이 아니었지만, RCT 방식을 적용하는 데에 있어서 고려해야 할 과제는 찾아볼 수 있다. 가장 큰 교훈은 RCT가 제도화되지 않으면, 부처 자체의 입장에서는 비용과 시간이 많이 소요될 개연성이 높은 RCT를 자발적으로 시도할 가능성이 낮다는 것이다. 희망리본 사업에서도 원래의 사업 기획자들은 선도적인 사례로 RCT를 시도하고자 했지만, 보직이동에 의해 새로운 사업담당자들이 사업을 담당하게 되었다. 그리고 새로운 사업담당자는 RCT에 대한 중요성을 뚜렷이 인식하지는 못하였다. 그리고 예산확보에 있어서도 적극적인 노력을 기울이지 않음으로써 RCT를 자연스럽게 포기하게 되었다. RCT로 실제 사업의 효과성을 검증하여, 신규 도입 여부 또는 확대 여부를 결정하고자 하는 수단으로 활용하고자 한다면, 재원 확보를 포함한 제도화가 중요함을 시사한다.

현재의 시범사업이 가지는 한계도 희망리본 사례를 통해 볼 수 있다. 원래는 희망리본 사업이 기존 자활사업의 사업 모형을 개선하고자 하는 시도로 추진되었으나, 결과적으로 별도의 신규사업으로 자리매김하였다. 시범사업 기간이 지나가면서 자연스럽게 담당 부처 내부에서 새로운 별도의 사업화가 되었으며, 중앙 예산 당국에서도 시범사업으로서의 희망리본 사업의 취지를 모니터링하지 못하는 결과를 낳았다. 희망리본 사업의 경우, 성과 자체는 예상을 뛰어넘는 결과를 창출하여 사업 모형 자체의 효과성에 대해서는 의문이 없었지만, 기존 사업의 개선 모형으로서의 역할이 아니라 추가되는 새로운 사업이 됨으로써, 시범사업의 취지를 유지하지 못했다는 아쉬움이 있다. 여기서도 시범사업의 제도화의 중요성을 볼 수 있다. 중앙기관이 시범사업 자체를 제도화하고, 시범사업의 목적에 부합하는 모니터링과 평가를 통해 시범사업의 취지에 부합하는 의사결정을 하는 체제 수립이 필요하

다. 특히 우리나라와 같이 순환보직이 정례화되어 있는 체제에서는, 제도화 및 이를 모니터링하고 환류할 수 있는 일관성 있는 실행 체제를 마련하는 것이 필수적이다.

### 라. 희망리본 사업의 RCT 적용 가상 사례

실제 사업에의 RCT 적용 사례를 가상으로라도 살펴보기 위해, 희망리본 사업에 RCT를 적용하는 가상 사례를 상정해 보자. 첫 번째 단계는 희망리본 사업에 RCT를 적용하는 것이 적절한지에 대한 정책적 윤리적 판단이 필요하다. 희망리본 사업에 RCT를 적용하기 위해서는 자활사업 대상자 중 일부는 기존의 자활사업에 배치하고, 일부는 희망리본 사업에 배치하여 서비스를 제공하여야 하는데, 배치 과정이 전적으로 무작위 과정을 통해 이루어져야 한다. 희망리본 사업의 수혜자가 기존 자활사업과 다른 직접적인 특혜를 받는 것은 아니며, 다른 유인 구조를 가진 서비스 제공기관의 서비스를 받는다는 점만 다르다. 그러므로 자활사업과 희망리본 사업에 사업 참여자를 무작위로 배치한다고 해서 윤리적인 문제가 발생한다고 보기는 어렵다.

기존 자활사업과 희망리본 사업에 참여할 대상자를 무작위로 추출하여 배치하는 것과 동시에 지역적인 배분도 중요하다. 특히 취업연계 복지사업의 경우, 지역의 경제적 상황이 성과창출에 중요한 영향을 미친다. 그러므로 동일한 지역에서 두 사업이 동시에 진행되도록 디자인하는 것이 중요하다. 희망리본 사업의 경우, 초년도에는 경기도와 부산에서 사업이 진행되었다. RCT를 적용할 경우, 이미 자활사업이 진행되고 있는 지역인 경기도와 부산에서 동일하게 희망리본 사업을 적용하는 것이 필요하다. 동일한 지역에서 진행된 두 사업 간의 성과를 비교 분석함으로써 희망리본 사업이 기존의 자활사업보다 우월한 성과를 창출하는지를 판단할 수 있다.

희망리본 사업의 또 하나의 쟁점은 일자리가 많은 도시지역에서만 가능한 사업이 아니냐는 것이다. 그러므로 다양한 지역을 대상으로 사업을 진행해서 지역에 따른 사업의 적합성을 진단할 필요가 있다. 농촌지역에서도 과연 취업연계형 복지사업인 희망리본 사업이 효과성이 있을 수 있는지를 판단할

수 있도록 지역 선정을 다양하게 할 필요가 있다.

RCT를 통해 희망리본 사업 자체가 가지는 다른 쟁점에 대한 분석도 시도할 수 있다. 예를 들어, 취업난이도를 어떤 방식으로 성과계약에 반영하는 것이 적절한지에 대해서도 몇 가지 모형을 시도해 볼 수 있다. 희망리본 사업의 경우, 영리 기업도 사업에 참여할 수 있도록 하는 노력이 있었다. 그러므로 사업수행기관의 특성에 따른 성과를 비교분석할 수 있도록 사업 수행기관을 다양하게 선정하는 것도 가능하다. 그럼으로써 영리, 비영리, 지자체 등 다양한 형태의 사업 수행기관 간의 성과를 비교 분석하는 것이 가능하게 될 것이다.

취업연계 복지사업의 성과는 중기적으로 나타날 개연성이 높으므로 적어도 3년 정도 사업을 추진하여 성과를 판단하는 것이 적절하다. 그리고 RCT를 통한 시범사업 추진 결과를 바탕으로 기존 자활사업의 개편 여부와 새로운 사업 모형에 대한 합의 도출이 가능할 것이다.

RCT를 기반으로 한 시범사업을 제대로 추진하기 위해서는, 전문연구기관이나 전문가와 중앙기관(중앙 예산 당국)의 역할이 중요하며, 사업 부처의 적극적인 참여가 필수적이다. 시범사업 목적의 명확화, 사업의 특성에 부합하는 RCT의 적절한 디자인, 사업 추진 및 사업 추진과정의 모니터링, 사업 결과의 적절한 환류 등이 일관성 있게 이루어지는 체제 마련이 필요하다.

## 2. 미국 연방교육부 교육과학협회의 EMSP 사업[40)]

2002년에 「교육과학개혁법」(Education Sciences Reform Act of 2002)에 의해 미국 연방교육부(U.S. Department of Education) 산하에 설립된 교육과학협회(Institute of Education Sciences)는 미국 교육의 현황과 정책 프로그램의 효과성 등을 연구하기 위해 RCT 사용을 장려하는 기관이다. 기관 설립 이후 RCT를 활용한 다양한 연구결과가 발표되었는데 여기서는 그 중

40) 이 부분의 내용은 미국 교육과학협회 보고서 Snipes 외(2015)를 요약한 것이다. 더욱 자세한 내용은 이 보고서를 참고하면 된다.

에서 여름방학 동안 제공되는 수학 보충수업 프로그램의 효과성을 연구한 보고서를 소개한다. 이는 앞서 언급한 우리나라의 다양한 보충수업 프로그램과 유사한 성격을 띠고 있으므로 실제로 RCT를 활용해 어떻게 보충수업 프로그램의 효과성을 평가하는지 살펴볼 수 있는 좋은 사례이다.

Elevate Math Summer Program(EMSP)은 실리콘밸리교육협회(Silicon Valley Education Foundation)가 8학년에 진학하는 학생들을 대상으로 설계한 프로그램이다. 이 프로그램은 네 가지 하위프로그램으로 구성되어 있는데 그중 가장 핵심은 여름방학 동안 4주에 걸쳐(19일, 총 75시간) 진행되는 준비수업이다. 이 준비수업은 8학년에 진학하는 학생 중 6학년 때 치룬 600점 만점의 캘리포니아 표준시험(California Standard Test; CST)의 수학과목에서 325~360점을 받은 학생들을 대상으로 한다. 과거에 이 점수대를 받은 학생들은 8학년 때 대수학I(Algebra I) 과목을 이수해야 했는데, 이들 중 추가적인 도움 없이 8학년 CST에서 350점 이상을 받은 학생은 절반에 불과했다. 최근 교육과정 개정으로 6학년 CST에서 325~360점을 받은 학생들은 대수학과 다른 관계 과목을 통합한 공통핵심수학(Common Core Math) 수업을 이수해야 하는데, 이는 이전의 대수학I 과목만큼 어려울 것으로 예상되기 때문에 학생들에 대한 보충수업이 필요한 상황이다. EMSP는 이러한 학생들을 대상으로 여름방학 동안에 보충수업을 진행하는데 연구진은 이러한 보충수업이 학생들의 수학성취도와 대수학준비도(algebra readiness)에 어떠한 영향을 미쳤는지 분석하고자 했다.

이 연구에서 대답하고자 하는 질문은 크게 3가지이다.

① 첫째, EMSP가 8학년에 진학하는 학생들의 수학 성취도와 대수학 준비도에 미친 영향은 무엇인가?

② 둘째, EMSP가 프로그램 커리큘럼과 가장 관계있는 수학 영역에서 학생들의 수학 성취도에 미친 영향은 무엇인가?

③ 셋째, EMSP가 8학년에 진학하는 학생들의 수학에 대한 관심과 자신감에 미친 영향은 무엇인가?

이러한 질문에 답하기 위해 연구진은 2014년 여름 6개 지구(district) 8개 학교에서 RCT를 시행했다. 총 496명의 8학년 진학 예정 학생들이 RCT에 참여했는데 이 중 19명의 학생은 초기 자료조사에서 참여를 포기해 477명의 학생들을 대상으로 무작위처치 여부를 결정했다. 이 학생들의 6학년 CST 점수는 아래 표와 같다.

〈표 Ⅴ-1〉 Elevate Math Summer Program의 RCT에 참여한 학생들의 6학년 CST 성적 분포

| | Far below basic (150~252점) | Below basic (253~299점) | Basic | | Proficient | | Advanced (415~600점) | Unknown |
|---|---|---|---|---|---|---|---|---|
| | | | Low (300~324점) | High (325~349점) | Low (350~360점) | High (361~414점) | | |
| 비율 (%) | 0.84 | 8.60 | 19.92 | 34.80 | 17.82 | 14.05 | 2.52 | 1.47 |
| 학생 수 (총 477명) | 4 | 41 | 95 | 166 | 85 | 67 | 12 | 7 |

자료: Snipes, Huang, Jaquet and Finkelstein(2015), p. A-3

각 학교별 무작위 추첨을 통해 총 477명의 참가자 중 239명의 학생은 처치집단에, 나머지 238명의 학생은 비교집단에 포함시켰다. 처치집단의 학생들은 첫 4주 동안 보충수업을 듣게 되고, 비교집단의 학생들은 첫 4주 보충수업이 끝난 직후 4주 동안 같은 보충수업에 참여하기로 했다.

한편 연구진은 수학 성취도를 측정하기 위해 수학진단시험 프로젝트(Mathematics Diagnostic Testing Project; MDTP)의 대수학 준비시험(Algebra Readiness test) 결과를 사용했다. 이 시험은 각 보충수업의 첫 날과 마지막 날에 치러졌다. 그리고 MDTP는 총 7개의 영역으로 구성되어 있는데, 대수학 준비도는 이 중 세 개 이상의 영역에서 통과했는지를 기준으로 삼았다.

또한 두 번째 질문에 답하기 위해 MDTP의 7개 영역 중 EMSP의 커리큘럼과 가장 가까운 영역으로 다음 세 가지 영역을 선택했다: ① decimals, their operations and applications, and percent, ② literals and equations,

③ geometric measurement and coordinate geometry. 그리고 세 번째 질문의 수학과목에 대한 관심과 자신감을 측정하기 위해 실리콘밸리교육협회에서 운영 중인 학생인식 설문조사를 이용했다. 또한 보다 정확한 추정값을 얻기 위해 6학년 CST 결과를 관찰변수로 회귀식에 포함시켰다.

보충수업 프로그램의 효과는 처치집단이 첫 번째 보충수업 마지막 날 치룬 MDTP 시험결과에서 비교집단이 두 번째 보충수업의 첫 날 치룬 MDTP 시험결과를 차감함으로 계산했다. 따라서 여기서 측정한 프로그램의 효과성은 두 가지 결과, 즉 처치집단이 보충수업을 통해 얻은 성적변화와 비교집단이 첫 4주 동안 학업을 하지 않은 결과가 결합된 것이다. 이 과정에서 처치집단 총 239명 중 165명만이 시험을 치렀고, 비교집단에서는 총 238명 중 184명만이 시험을 치렀다. 효과성 분석은 MDTP 시험을 치룬 학생만을 대상으로 하였고 이탈자가 처치집단과 비교집단에 유의미한 차이를 불러왔는지를 검토하기 위해 6학년 CST 결과를 이용해 시험전 동등성테스트(baseline equivalence test)를 시행했다. 그 결과 연구진은 이탈자를 제외한 이후 처치집단과 비교집단 사이의 차이는 통계학적으로 유의미하지 않다는 결론을 내렸다.[41]

연구 결과 EMSP는 전반적으로 학생들의 수학 성취도와 대수학 준비도를 향상시킨 것으로 드러났다. 먼저 총 45점 만점의 MDTP 시험에서 처치집단은 평균 21점을 받았으며, 비교집단은 평균 17점을 받았다. MDTP의 7개 영역 중 최소 3개 영역에서의 통과를 기준으로 했던 대수학 준비도에서도 처치집단은 29%의 학생이 기준을 통과한 반면, 비교집단에서는 12%의 학생만이 기준을 통과했다. 이러한 두 집단 간의 차이는 통계학적으로 유의미한 것으로 판정됐다.

뿐만 아니라 EMSP는 EMSP 커리큘럼과 관계있는 세 개의 영역에 대해서도 긍정적인 영향을 미친 것으로 나타났다. 하지만 이러한 결과에도 불구하고 보충수업 종료 이후에 실시된 시험결과에서 대부분의 학생들은 8학년 대수학 과정을 이수할 준비가 되지 않은 것으로 판정됐다. 처치집단의 학

41) 이탈자에 대한 자세한 분석은 본 보고서를 참고하도록 한다.

생들이 비교집단의 학생들보다 더 나은 성적을 받았지만, 8학년 대수학I 과목을 통과할 수 있는 점수에는 못 미치는 것으로 드러났다. 이는 EMSP 참여 학생들이 설계된 프로그램 이외에 추가적인 도움이 필요하다는 것을 의미한다.

그리고 연구진은 위에서 추정된 EMSP가 MDTP 시험결과에 미친 영향 중 약 1/3 이상은 비교집단의 학생들이 첫 4주 동안 학업을 하지 않은 결과임을 밝혔다. 처치집단과 비교집단이 보충수업 참여 이전에 유사한 집단으로 설계했기 때문에 비교집단이 두 번째 보충수업 참여 이전에 치룬 시험결과는 처치집단이 보충수업을 참여하지 않았을 가상의 경우 4주 후의 성적으로 생각할 수 있다. 처치집단이 첫 번째 보충수업 이전에 치룬 시험성적의 평균은 비교집단이 두 번째 보충수업 이전에 치룬 시험성적의 평균보다 1.47점 높았는데 이는 EMSP가 MTDP 점수에 미친 영향의 약 37%를 차지한다.

그리고 마지막으로 설문조사 결과 EMSP는 처치집단의 학생들이 수학에 대해 더 많은 관심을 갖게 하지만 이는 통계학적으로 유의미하지 않았으며, 수학에 대한 자심감에 미치는 영향은 없는 것으로 판정됐다. RCT를 활용한 이 연구결과는 여름방학 동안 제공되는 수학 보충수업의 효과와 한계가 무엇인지 비교적 객관적으로 보여주고 있다.

하지만 연구진은 이러한 연구결과의 한계점도 명시하고 있다. 먼저 이 연구는 RCT에 참여한 지역의 학교를 대상으로 실시되었기 때문에 이와 유사한 지역과 학생들에게만 일반화가 가능하다. 그리고 교과과정 개정으로 8학년 학생들은 이제 공통핵심수학을 이수하기 때문에 대수학I에 대한 본 연구의 결론을 그대로 공통핵심수학에 적용할 때에는 주의를 필요로 한다.

그리고 주로 6학년 중 CST 점수가 낮은 학생들이 RCT에서 이탈했으며, 특히 비교집단에서 이러한 현상이 더욱 두드러지게 나타났기 때문에 효과성 분석 결과를 해석할 때 주의를 요한다. 이탈자로 인한 처치집단과 비교집단 사이의 차이는 통계학적으로 유의미하지 않다고 판정했지만, 연구진이 자료에서 확인할 수 없는 변수가 성과변수에 영향을 미치지 않았는지 검토가 필요하다. 또한 약 27%의 학생들이 RCT에서 이탈했기 때문에 이탈자와 실제 참여자 사

이의 유의미한 차이가 있다면 연구결과의 일반화에 제한이 있을 수 있다.

마지막으로 이 연구결과는 단기에만 적용가능하기 때문에 장기적으로 보충수업이 어떠한 영향을 미치는지, 그리고 수학 전반에 대한 학습능력에 어떠한 영향을 미치는지에 대해서는 추가적인 연구가 필요하다.

### 1) 우리나라 교육 정책에 대한 시사점

앞서 언급한 바와 같이 교육부가 2015년 2월에 발표한 '제2차 수학교육 종합계획(2015~2019)'에 따르면 학습부진 학생들을 대상으로 '맞춤형 수학멘토링' 프로그램을 활성화할 계획이다. 이 계획에 따르면 맞춤형 수학멘토링 프로그램은 학업성취도를 향상시킬 뿐만 아니라 수학에 대한 관심과 동기유발을 촉진하는 방향으로 운영될 계획이다. 하지만 이러한 프로그램을 운영하기 전에 그 효과성에 대한 객관적인 평가는 이루어지지 않은 것으로 보인다. 2012년 1월에 발표된 '수학교육 선진화 방안'에도 이와 비슷한 프로그램이 시행되었지만 그 효과성에 대한 심도 있는 평가는 이루어지지 않았다. 멘토링의 효과성을 분석한 연구는 있었지만 이러한 연구 또한 멘토링이 학업성취도나 자신감 등에 미친 영향을 RCT를 이용해 분석하지는 않았다.[42][43]

하지만 위의 미국 교육과학협회의 예에서 확인할 수 있듯이, RCT를 이용한 프로그램의 효과성 연구는 프로그램의 영향력을 객관적으로 추정할 뿐만 아니라 그 설계방법에 따라 추후 프로그램을 수정 · 보완하는데 유용한 정보를 이끌어낼 수도 있다. 특히, 방학 보충수업은 학업성취도를 향상시키는 효과뿐만 아니라 방학 동안에 학생들이 학업을 꾸준히 지속시켜주는 역할도 하고 있다. 따라서 미국 사례처럼 처치집단과 비교집단이 각각 보충수업 전후로 시험을 치루게 하면 보충수업의 두 가지 효과를 구분해 어느 요인이 더욱 학업성취도 향상에 효과적인지 분석할 수 있다.

---

42) 멘토링이 수학학습부진아의 수학학습태도에 미친 영향에 대한 연구로 권수진 외(2014)가 있지만 이 연구에서는 RCT 방법을 사용하지 않았다.

43) 연구진이 알고 있는 RCT를 이용한 교육 프로그램 평가 연구는 신을진·이일화(2010)가 있다.

또한 위의 미국 사례에서 RCT를 시행한 방법도 좋은 참고사례가 될 수 있다. 우리나라에서 현재 '방과 후 학교' 프로그램을 운영 중인데 이는 기본적으로 학생들의 자발적인 참여를 원칙으로 하고 있다. 따라서 만약 방과 후 학교의 효과성 평가를 위해 참여 학생과 그렇지 않은 학생들을 단순히 비교할 경우 선택편향 문제를 해결할 수 없기 때문에 그 결과는 신뢰성을 갖기 어렵다. 하지만 미국 사례처럼 처치집단과 비교집단의 프로그램 참여 시기만을 무작위로 배정하면 모든 지원자가 프로그램의 수혜를 받는 동시에 프로그램 담당자와 연구진은 프로그램의 효과를 객관적으로 분석할 수 있게 된다. 이 방법은 큰 추가비용 없이 RCT를 시행할 수 있는 방법으로 바로 현장에서 적용할 수 있을 것으로 보인다.

# Ⅵ. RCT를 통한 재정사업 사전검증체계 도입방안

이상에서 RCT에 대한 개념 소개와 우리나라에서의 활용현황, 적용과정에서의 장애요인, RCT를 통한 주요 평가설계를 사례를 통해 살펴보았다. 본 장에서는 RCT를 통한 재정사업 사전검증체계 도입방안에 대한 정책적 대안을 제시한다.

## 1. 제도적 토대 구축

재정사업 사전검증체계 강화를 위해 RCT와 같은 과학적인 평가방법을 적용하기 위해서는 원활한 평가수행을 위한 법적, 제도적 토대가 필요하다. 제도적 안정성 없이 비교적 새로운 RCT와 같은 평가기법이 정착되기 어렵기 때문이다. 특히, RCT를 시행하는 데 있어서는 관련 공무원의 평가자체에 대한 이해가 낮아 실행이 어려울 뿐만 아니라 윤리적인 문제나 정책대상자의 심리적 거부감 등으로 평가의 실행이 어려운 경우가 많기 때문이다. 따라서 과학적인 평가방법에 대한 법적, 제도적 절차를 명확히 하여 제도의 지속가능성을 높일 필요가 있다.

### 가. 법적인 근거마련

우선, 재정사업 사전 검증평가를 위해 예비타당성조사가 「국가재정법」과 「국가재정법 시행령」에 규정되어 있는 것처럼 RCT에 대한 규정을 관련 법규에 명문화할 필요가 있다. 현재 「국가재정법」 제38조에는 재정사업 예비타당성조사 규정이 명문화되어 있으며 「국가재정법 시행령」 제13조에는 예비타당성조사의 실시와 면제요건이 규정되어 있다. 따라서 대규모 예산이

들어가면서 기존의 예비타당성조사의 평가기법으로 정확한 사업효과를 예측하기 어려운 노동, 복지, 교육 사업에 대하여 RCT 기반 정책평가를 의무화하는 규정을 「국가재정법」과 시행령에 신설할 필요가 있다. 미국의 경우 2002년에 'Eduction Science Reform Act'와 같은 법률 제정을 통해 교육프로그램의 효과성을 평가하는데 있어 RCT와 같은 Scientifically Based Research를 실시할 것을 명문화하였다. 이 법은 모든 교육 정책 및 프로그램의 평가를 과학적 방법론에 근거하여 실시하도록 규정하고 있다. 이 법률 18조 19조에는 과학적 연구와 평가(Scientifically Based Research or Evaluation)의 정의를 정책효과의 엄밀한 인과관계를 검증할 수 있는 무작위 추출에 기반한 실험설계 및 그에 상응하는 연구설계로 규정하여 RCT에 대한 법적 근거를 제공하고 있다.

〈표 Ⅵ-1〉 미국 'Eduction Science Reform Act of 2002'의 RCT 규정

(18) SCIENTIFICALLY BASED RESEARCH STANDARDS.—(A) The term "scientifically based research standards" means research standards that—
(i) apply rigorous, systematic, and objective methodology to obtain reliable and valid knowledge relevant to education activities and programs; and
(ii) present findings and make claims that are appropriate to and supported by the methods that have been employed.
(B) The term includes, appropriate to the research being conducted—
(i) *employing systematic, empirical methods that draw on observation or experiment;*
(ii) involving data analyses that are adequate to support the general findings;
(iii) relying on measurements or observational methods that provide reliable data;
(iv) *making claims of causal relationships only in random assignment experiments or other designs(to the extent such designs substantially eliminate plausible competing explanations for the obtained results);*
(v) ensuring that studies and methods are presented in sufficient detail and clarity to allow for replication or at a minimum, to offer the opportunity to build systematically on the findings of the research;
(vi) obtaining acceptance by a peer-reviewed journal or approval by a panel of independent experts through a comparably rigorous, objective, and scientific review; and
(vii) using research designs and methods appropriate to the research question posed.
(19) SCIENTIFICALLY VALID EDUCATION EVALUATION.—The term "scientifically valid education evaluation" means an evaluation that—
(A) adheres to the highest possible standards of quality with respect to research design and statistical analysis;

〈표 Ⅵ-1〉의 계속

(B) *provides an adequate description of the programs evaluated and, to the extent possible, examines the relationship between program implementation and program impacts;*
(C) provides an analysis of the results achieved by the program with respect to its projected effects;
(D) *employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible; and*
(E) may study program implementation through a combination of scientifically valid and reliable methods.

출처: IES 홈페이지,http://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf, 검색일자 2015. 10. 15.

우리의 경우도 RCT 기반 정책평가를 안정적으로 실시하기 위해서는 「국가재정법」 등에 관련 규정을 명문화할 필요가 있다. 이러한 규정에는 RCT 평가를 위한 예비사업, 평가대상 사업의 범위와 기본절차, 예비사업의 효과성을 과학적으로 검증할 수 있는 평가 설계 유형에 대한 개괄적인 규정을 포함시켜야 한다. 〈표 Ⅵ-2〉는 RCT 기반 정책평가의 기본이 되는 증거기반 사전타당성 평가가 「국가재정법」에 규정된 예를 보여준다.

〈표 Ⅵ-2〉 RCT 기반 사전타당성 평가규정의 예

제38조(예비타당성조사) 기획재정부장관은 총사업비가 500억원 이상이고 국가의 재정지원 규모가 300억원 이상인 신규 사업으로서 다음 각 호의 어느 하나에 해당하는 대규모사업에 대한 예산을 편성하기 위하여 미리 예비타당성조사를 실시하고, 그 결과를 요약하여 국회 소관 상임위원회와 예산결산특별위원회에 제출하여야 한다. 다만, 제4호의 사업은 제28조에 따라 제출된 중기사업계획서에 의한 재정지출이 500억원 이상 수반되는 신규 사업으로 한다.

신설: 제39조(예비사업 타당성 평가) 기획재정부장관은 제38조의 예비타당성조사 대상사업으로서 예비타당성조사에 의해 사업의 효과성 검증이 곤란하거나 불가능한 사업에 대하여 사업의 실시 전에 예비사업을 실시하고 예비사업의 효과성을 입증하는 타당성 평가를 실시하여야 한다. 다만, 사업의 시행부서에서 예비사업 타당성 평가에 준하는 평가를 실시하였을 경우 기획재정부의 승인을 거쳐 평가를 면제할 수 있다.

사업의 효과성에 대한 인과관계를 과학적으로 검증할 수 있는 평가는 다음과 같다.
1. 무작위 추출에 의한 처치-비교집단 효과성 비교 평가
2. 1항에 준하는 사업효과의 인과성을 확보할 수 있는 평가
3. 1항과 2항의 예비사업 타당성 평가의 대상사업, 수행체계, 분석방법, 평가환류에 관한 구체적인 절차는 대통령령과 예비사업 타당성 평가지침에 따른다.

### 나. 행정자료공유 제도화

RCT와 같은 실험평가를 활성화하기 위해 필요한 제도적 방안으로 자료공유시스템을 구축하는 것도 중요한 과제이다. 개별 실험이 효과적으로 진행되기 위해서는 비용문제를 고려하지 않을 수 없는데, 다양한 기존 행정자료를 활용할 수 있다면 비용을 크게 절감할 수 있기 때문이다. 미국의 경우 RCT를 비롯한 정책평가를 활성화하기 위하여 자료공유시스템을 구축하고 각 부처가 보유하고 있는 자료들을 서로 공유하여 평가의 비용을 최소화하고 있다. 우리나라의 경우 조세지출이나 재정사업의 심층평가에서 부처의 자료제공의 미협조로 인해 어려움을 겪고 있는 경우가 많은데, 이는 RCT를 통한 실험평가에도 유사하게 나타날 확률이 높다. 따라서 「국가재정법 시행령」에 각 부처의 자료제출 공유의무를 명문화하고 행정자료를 통합관리할 수 있는 시스템을 평가전담기관에 설치해야한다. 또한 제도이행의 실효성을 높이기 위하여 부처의 자료공유 실태를 정기적으로 점검하고 이를 공개하거나 자료공유 노력을 정부업무평가 또는 행정관리역량평가에 반영할 필요가 있다.

### 다. 전담조직 설치

다음으로, RCT 기반 정책평가를 재정사업의 평가에 제도화하기 위해서는 해외 국가에서와 같이 전담조직 설치가 필요하다. 이러한 전담조직은 현재 재정사업 사전사후 평가를 담당하고 있는 국책연구원에 설치할 수 있다. 다만, 기존 예비타당성조사를 수행하고 있는 KDI의 경우 기존 예비타당성조사와 상충될 수 있고 RCT 기반 평가가 재정사업의 사전평가 외에 심층평가와 같은 사후평가에서도 적용될 수 있기 때문에 다른 국책기관이나 해외국가와 같이 별도의 평가전문 기관 설립을 통해서 실시하는 것이 바람직하다. 별도의 전문기관은 내각 소속의 정부조직에서 독립기관으로 된 영국의 BIT의 형태, 미국 대통령 직속하에 공무원 조직인 Social and Behavioral Science Team, MIT의 J-PAL과 같은 순수민간기관의 형태를 검토할 수 있다. 이러한

전문기관은 정책평가, 교육, 보급 확산, 자료관리 네 가지 핵심 기능을 수행해야 한다.

(1) RCT 기반 정책평가 실시
(2) RCT 기반 정책평가 역량배양을 위한 교육프로그램 운영
(3) RCT 기반 정책평가의 보급과 확산을 위한 제도 및 파트너십 구축
(4) 정책성과에 대한 통합 행정자료 및 DB 관리

## 2. RCT 평가실행을 위한 수행체계와 절차의 명확화

RCT 평가를 위한 법적 근거와 전담조직을 설치하였다면 RCT 평가실행을 위한 구체적인 수행체계와 절차를 마련해야 한다. 예비타당성조사의 경우도 별도의 지침을 통해 평가 실행을 위한 구체적인 수행체계와 절차를 규정하고 있다. RCT를 통한 재정사업 평가의 경우도 실행과정에 장애요인이 많기 때문에 명확하고 구체적인 절차를 작성할 필요가 있다. 이러한 절차에는 평가대상사업선정, 평가수행체계, 평가분석방법, 평가결과 공개와 활용에 대한 내용을 포함시켜야 한다.

### 가. 평가대상사업 선정기준 명확화

먼저, RCT 기반 재정사업 평가대상이 되는 사업의 범위를 명확하게 선정할 필요가 있다. 이는 예비타당성조사와 같이 RCT에 의한 엄밀한 평가 설계가 불가능하거나 시행에 있어 정책적인 고려가 필요한 사업영역이 존재하기 때문이다. 이러한 특성을 고려하여 RCT를 통한 재정사업 사전검증은 예비타당성조사와 같이 총 사업비가 500억원 이상(국가재정지원 300억원 이상)이면서 예비타당성조사에 의해 정확한 편익의 도출이 어렵거나 불가능한 사업에 대해서 실시하는 것을 검토할 수 있다. 이러한 사업들은 주로 정확한 편익의 추정이 어려워 비용/편익분석 대신 비용/효과분석을 실시하고,

AHP 분석에서 정책적 효과에 가중치를 높게 부여하는 고용, 복지, 노동, 교육 등의 사회서비스이다.

다만, 이러한 사회서비스에 대하여 RCT를 통한 사전 효과성 평가가 가능하다고 할지라도 여러 현실적인 여러 장애요인들로 인하여 평가가 불가능할 수 있다. 전술한 바와 같이, 사업수혜자의 반대가 심해 무작위 추출을 통한 비교집단 설정이 불가능하거나, 비교집단이 설정 가능하더라도 처치집단과 비교집단 간 다른 처치에 대한 참가자들의 반대로 인하여 평가가 불가능 할 수도 있다. 물론, RCT 설계가 불가능할 경우 준실험이나 비실험설계를 통하여 예비사업의 효과성에 대한 인과관계를 추론할 수 있지만 효과성을 측정하는 데이터 수집이 불가능하거나 비용이 많이 들 경우 정책효과의 인과관계를 정확하게 분석하는 평가는 어렵게 된다. 따라서 평가의 실행과정에서 현실적인 장애요인을 극복할 수 없는 경우 〈표 Ⅵ-3〉과 같이 RCT를 통한 예비사업 효과성 사전평가를 면제할 필요가 있다.

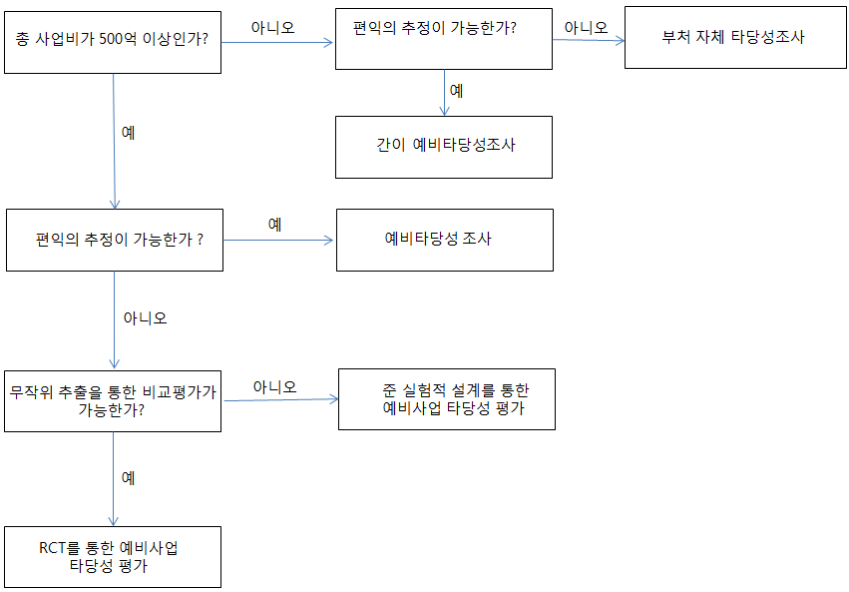
**〈표 Ⅵ-3〉 예비사업 타당성 평가 면제사유 규정의 예**

| |
|---|
| 1. 공공청사, 교정시설, 초 · 중등 교육시설의 신 · 증축 사업 |
| 2. 문화재 복원사업 |
| 3. 국가안보에 관계되거나 보안을 요하는 국방 관련 사업 |
| ..... |
| 8. 법령에 따라 추진하여야 하는 사업 |
| 9. 출연 · 보조기관의 인건비 및 경상비 지원, 융자 사업 등과 같이 예비타당성조사의 실익이 없는 사업 |
| 10. 지역 균형발전, 긴급한 경제 · 사회적 상황 대응 등을 위하여 국가 정책적으로 추진이 필요한 사업 |
| 11. *예비사업 타당성 평가의 실행이 불가능한 사업. 예비사업 타당성 평가의 실행이 불가능한 사업의 구체적인 유형은 대통령령과 예비사업 타당성 평가지침에 따른다.* |

RCT 기반 사전평가 대상 사업 선정에서 마지막으로 고려해야 할 것은 선정기준의 객관성을 확보하는 것이다. 어떤 사업을 예비타당성조사로 할 것인지 혹은 RCT 평가로 할 것인지, 아니면 정성적인 방법으로 사전평가를 실시할 것인지를 결정하는 것은 개별사업의 성격과 실행환경에 따라 매우 다르기 때문이다. 이를 위해 우선 RCT 평가대상 선정에 대한 구체적인 기준

을 제시하고, 이러한 기준에 따라 평가대상 사업을 결정할 독립적인 평가사업선정위원회를 설치할 필요가 있다. 독립적인 평가대상사업선정위원회의 선정기준은 세부지침을 통해 구체화해야 하지만 [그림 Ⅵ-1]과 같이 사업규모, 편익추정 여부, 무작위 추출 여부를 통해 RCT 기반 평가를 위한 최종사업을 선정할 수 있다.

[그림 Ⅵ-1] RCT 기반 예비사업 타당성 평가대상 사업 선정절차 예시




### 나. 평가수행체계의 구체화

평가대상사업 선정기준을 확정하면 평가를 수행하는 체계에 대한 절차를 마련해야 한다. 평가수행체계는 평가의 기획 및 감독기관, 평가시행기관, 연구진 선정, 연구기간을 통하여 구체화해야 한다. RCT 정책평가는 기획재정부, 평가실행기관, 개별부처 협업을 통해서만 수행할 수 있다. 기획재정부에서 RCT 평가 대상사업과 평가에 대한 예산을 확정하면 평가실행기관은 사업부서와 사전에 면밀한 평가계획을 수립하여 예비사업을 실시하고 사전에

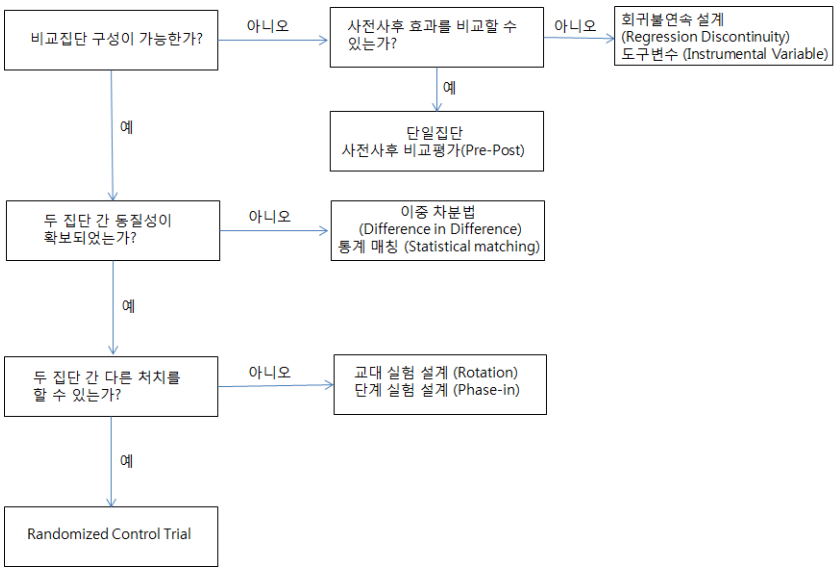
계획된 평가계획의 의거하여 예비사업의 효과성을 평가한다.

예를 들어, 평가실행기관은 연구진을 구성하고 연구진은 개별부처 담당자와 공동으로 평가의 장애요인, 처치/비교 실행방법, 사업 효과성 측정방법을 포함한 구체적인 평가계획서를 확정한다. 평가계획서가 확정되면 실제 예비사업을 처치 집단에 시행하고 그 효과를 비교집단과 비교하여 사업의 최종 효과성을 판단한다. 평가 수행기간은 예산주기와 사업효과성 발생시간을 고려하여 1년으로 설정하되, 사업의 성격과 유형에 따라 6개월에서 2년까지 탄력적으로 조정할 수 있도록 한다. 다만 기획재정부에 의한 RCT 사전평가가 개별부처의 평가에 대한 관심과 역량을 낮출 수 있기 때문에, 부처 자체적으로 시범사업 실시 후 사업의 효과성을 RCT와 그에 준하는 평가절차를 통해 입증하였을 경우 기획재정부에 의한 평가를 생략한다. 개별부처가 실시한 평가절차의 정확성에 대한 승인은 전술한 평가대상사업 선정위원회에서 실시한다.

### 다. 평가방법의 확정

RCT 기반 예비사업 평가의 분석방법도 엄밀한 RCT 평가를 위해 매우 중요한 요소이다. 일단, RCT 평가기법을 적용할 대상사업과 연구진, 사업부서를 정하면 어떤 평가방법을 사용할 것인가를 확정해야 한다. 무작위 추출을 통한 엄밀한 RCT 평가를 수행하기 위해서는 (1) 처치집단과 비교집단의 구성 (2) 집단 간 동질성의 확보 (3) 비교집단 간 다른 처치라는 세 가지 전제조건이 충족되어야 한다. [그림 Ⅳ-2]는 RCT에 의한 사전평가에서 고려할 수 있는 평가방법의 유형과 조건을 보여준다.

[그림 Ⅵ-2] RCT 평가방법 유형과 조건




### (1) 처치집단과 비교집단의 구성

RCT의 첫 번째 단계는 서로 다른 집단 간에 정책효과를 비교할 수 있는 비교집단을 구성하는 것이다. 그러나 비교집단 설정이 불가능한 경우가 많다. 주민의 반발이나 법 규정에 의해 자격 기준이 넘는 모든 정책수혜자들에게 정책을 시행해야 하는 경우가 발생하기 때문이다. 이 경우 RCT에 의한 정책평가가 불가능하며, 사업의 효과를 검증하는 준실험적 또는 비실험적 평가방법을 사용해야 한다. 정책시행 전에 정책 대상 집단의 성과를 알 수 있다면 정책시행 후에 성과를 측정하여 비교하는 단일집단 사전사후 비교법을 사용할 수 있다. 다만, 이 경우 비교집단이 설정되지 않았기 때문에 다른 요인에 의한 정책효과 발생을 통제할 수 없게 된다. 정책 대상 집단의 사전성과를 사전에 확보할 수 없다면 사업의 최종 효과성을 회귀분석과 같은 비실험적 방법으로 검증할 수 있는 회귀불연속 설계(Regression Discontinuity Design)나 도구변수법(Instrumental Variable Design)을 사용할 수 있다.

### (2) 두 집단 간 동질성 확보

처치집단과 비교집단을 구성했다고 할지라도 두 집단 간에 동질성을 확보해야 한다. 처치집단과 비교집단에 동질성을 확보하지 못하면 사업의 효과가 아니라 특정 집단의 특성에 의해 정책효과가 발생할 수 있기 때문이다. 집단 간 동질성은 추첨을 통한 무작위 추출로 확보할 수 있지만 주민의 선호나 형평성 문제 때문에 수혜대상자 결정에 있어 선착순 배정이 선호되고 있다. 이 경우, 선착순 신청자를 처치집단, 신청하지 않은 사람들을 비교집단으로 하여 정책효과를 검증할 수 있지만 선착순으로 신청한 사람들은 원래부터 정책선호도가 높은 집단이기 때문에 사업시행의 효과로 정책효과가 발생했는지를 파악하기 어렵다.

두 집단 간의 동질성이 확보되지 않았을 경우 사용가능한 평가 설계로는 통계적 매칭법[44](Statistical Matching)과 이중차분법(Difference in Difference; DID)가 있다. 통계적 매칭법은 정책처치에 참가하지 않은 사람들 중에서 처치집단으로 뽑힌 사람들의 특성과 유사한 사람들을 인위적으로 선발하여 비교집단을 구성하는 것이다. 매칭법은 처치집단과 유사한 비교집단을 선발할지라도 인위적으로 비교집단을 설정한다는 점에서 정확한 인과관계검증에 한계가 있다.

매칭법 외에 고려할 수 있는 평가기법이 이중차분법이다. 이중차분법에서는 정책의 참가집단과 비참가집단의 정책처치 시행 전과 시행 후 정책효과의 차이를 각각 측정한 후 두 집단 간의 차이를 다시 한번 차감하여 정책의 효과를 산출한다. 사업 전후의 정책참가자와 비참가자의 성과변화를 차감함으로써 보다 비교적 정확하게 정책효과를 측정할 수 있다. 그러나 정책참가자의 특성이 비참가자보다 사업 전과 사업 후의 성과변화에 더 큰 영향을 끼친다면 정책효과의 정확한 인과관계 검증이 어렵다는 한계가 있다.

---

44) 통계적 매칭법에는 정책대상자와 통계적으로 특성이 정확히 일치하는 사람만을 선발하여 비교그룹에 매칭하는 'Exact Matching'과 성향이 유사한 사람을 선발하여 매칭하는 'Propensity Score Matching'이 있다.

### (3) 대상 집단 간 상이한 정책처치의 실행

처치집단과 비교집단을 구성한 후 집단 간 동질성을 확보했다 하더라도 처치집단에만 특정 정책을 시행할 수 없는 상황이 발생할 수도 있다. 처치집단의 구성원들이 정책처치를 거부하거나 비교집단의 구성원들이 정책수혜를 요구할 때 발생한다. 자연과학과는 달리 정책의 수혜를 받는 사회과학의 RCT 평가에서 정책처치를 거부하는 경우는 드문 편이지만, 비교집단이 동일한 처치를 통한 정책수혜를 요구하는 경우는 빈번하게 발생한다. 이렇듯 대상 집단 간 상이한 정책처치의 실행이 어려울 경우 고려할 수 있는 방법이 교대실험(Rotation Design)이나 단계실험 설계(Phase-in)이다. 이러한 실험설계를 통하여 처치집단에 대한 정책처치 후에 교대로 또는 단계적으로 비교집단에도 처치집단과 똑같은 정책처치를 실시한다.

### (4) RCT에 의한 정책평가

앞서 언급한 세 가지 조건이 확보된다면 RCT에 의한 예비사업 평가를 실시할 수 있다. 여기서 중요한 것은 정책효과의 충분한 통계적 파워를 확보할 수 있는 정책대상 집단의 충분한 표본 수와 사전사후 비교를 위한 대상 집단의 사전성과 데이터를 확보하는 것이다. 표본의 크기는 통계적 파워의 수준과 정책처치로 달성하는 정책효과의 크기로 결정한다. 예를 들어 학생들의 성적을 향상시키기 위해 제안된 예비사업의 참가자 수는 통계적 파워의 수준을 높일수록 향상시키고자 하는 성적의 폭이 클수록 증가한다. RCT에 의한 평가 설계로 진행하는 예비사업에서 참가자의 수를 정하는 것은 평가예산을 결정하는 데 중요한 변수가 된다. 연구진과 사업담당자는 적정한 통계적 파워[45]와 정책효과의 크기를 사전에 결정한 후 참가자의 수를 정하고 그에 맞추어 예산을 편성할 수 있다.

45) 일반적으로 사용되는 적정한 통계적 파워의 수준은 80%이다.

### 라. 평가결과의 공개와 본사업 시행결정

일단 예비사업이 종료되면 처치집단과 비교집단 간 정책효과의 차이를 분석한 후 평가결과를 공개해야 한다. 이러한 공개절차는 예비타당성조사 절차에 따른다. 평과결과, 처치를 받은 정책대상자 집단의 성과가 비교집단의 성과보다 높게 나타났다면 본사업의 시행을 결정할 수 있다. 다만 RCT 평가결과의 외적 타당성을 확보하기 위해, 연구진은 회귀분석과 같은 비실험적 평가분석이나 참가자 인터뷰를 통한 질적분석 결과와 최종 평가결과를 비교해야 한다. 비교결과, 결과가 유사하다면 기획재정부는 본사업의 시행을 최종 승인하고 예산을 배정한다.

## 3. RCT 활성화를 위한 대내외 환경조성

RCT와 같은 전문적인 정책 평가를 원활히 운영하기 위해서는 행정부 내부의 사업담당자과 조직의 역량이 요구된다. 또한 RCT와 관련된 윤리적인 문제와 실행의 어려움으로 인하여 국회 및 국민에 대한 홍보를 통해 외부의 지지를 이끌어 내야 한다. 이렇게 행정부 내부와 국민으로부터 RCT 정책평가에 대한 이해와 지지를 얻은 후에, RCT 평가에 대한 전문가 풀을 구축하여 지속적인 증거기반 정책평가에 대한 지식과 경험을 공유해야 한다.

### 가. 행정부 내부의 평가 역량강화

앞서 언급했듯이 RCT 정책평가 운영에는 여러 장애요인과 전문성이 요구된다. 이는 RCT 평가가 내생성을 줄여 정확한 정책효과를 측정할 수 있는 장점에도 불구하고 실무에서 크게 활성화되지 않은 원인이 되었다. 더구나 연구진에 의해 수행되는 예비타당성조사와 달리, RCT에서는 실제 정책 처치가 실행되기 때문에 사업담당자의 RCT에 대한 깊은 이해를 바탕으로 한 적극적인 참여가 성공적인 정책평가를 위해 필수적이다. 이를 위해 사업담당 공무원의 RCT 정책평가에 대한 이해도를 높이기 위한 역량강화가 요구된다.

역량강화를 위해서는 MIT의 J-PAL의 훈련과정을 참고할 필요가 있다. J-PAL에서는 사례토론, 실제 적용실습이 포함된 5일 전일제 강의를 통하여 RCT에 대한 강도 높은 교육프로그램을 제공하고 있다. 우리나라의 경우 RCT가 재정사업의 사전평가로 제도화되면 전담기관에서 실제 RCT 평가와 더불어 담당자에 대한 교육훈련을 실시해야 한다. 또한 RCT가 가능하지 않은 다양한 사업들이 존재하기 때문에 RCT 평가 교육 외에 다양한 정책평가 기법에 대한 교육과정 프로그램을 개발하여 사업담당자의 평가관련 역량을 높여야 한다.

특히, 평가의 역량강화를 위해서는 교육훈련의 강화와 함께 희망리본의 사례에서 언급하였듯이 현재 순환보직으로 인해 1~2년에 머무르고 있는 사업담당자의 임기를 적어도 예비사업이나 시범사업의 평가가 종료되는 기간까지 연장할 필요가 있다. 정책평가에 대한 이해도를 높이고 정책효과를 산출하기 위해서는 일정시간이 필요하기 때문이다. 따라서 대규모 예산이 들어가는 신규 사업의 경우, RCT에 대한 이해와 사업타당성을 입증하기 위해 사업책임자의 임기를 보장해주는 유연한 인사제도가 필요하다.

### 나. 국회 및 대국민 홍보

RCT 정책평가의 보급에 가장 큰 장애요인인 정책대상자에게 차별적으로 정책처치를 하는 윤리적인 문제이다. 이를 위해 RCT가 사람을 대상으로 하는 실험이 아니라 정책의 효과를 과학적으로 검증하는 평가기법이라는 사실을 국회 및 국민들에게 적극적으로 홍보할 필요가 있다. 영국의 BIT는 RCT를 통해 도출한 평가결과를 적극적으로 언론을 통해 홍보하고 있다. 우리나라의 경우 청년고용, 무상보육과 급식, BK 21사업 등 사회적으로 논란이 된 대규모 정책사업이 있었지만, 정책효과에 대한 과학적 증거를 통해 진지하게 토론한 적은 거의 없었다. 따라서 실증기반 정책결정과 평가를 확산시키기 위해 대규모 예산이 들어가지만 정책효과에 논란이 있는 사업의 경우, 예비사업 단계에서부터 지속적으로 국회와 국민들에게 홍보 및 설득작업을 실시할 필요가 있다.

이런 과정을 거쳐 RCT 정책평가에 대한 국민들의 이해도가 높아지면 장기적으로 'PAYGO'와 더불어 대규모 재정을 수반하면서 정책효과가 의심되는 의원입법에 대하여 법안제출 전에 반드시 RCT 평가결과를 제출할 것을 의무화하는 방안을 검토할 수 있을 것이다. 더불어, 정책 실험평가에 대한 대국민 수용도를 높이기 위해 실제 RCT 평가과정에서 발생할 수 있는 법적, 윤리적인 문제를 사전에 심사할 수 있는 IRB[46](Institutional Review Board)를 정부 내에 설치하는 것을 검토할 필요가 있다.

〈표 Ⅵ-4〉 Institutional Review Board(IRB)의 예

**목적**: IRB 설치의 주요 목적은 모든 생명의과학연구(Biomedical Research)의 윤리적 과학적 타당성을 심의하여 연구대상자를 보호하기 위함이다. 이를 위해 연구자로부터 연구계획서, 연구대상자 설명문 및 동의서, 증례기록서 등 관련 자료를 받아 심의 및 승인하여 적절한 연구진행이 이루어질 수 있도록 관리 및 감독하는 역할을 수행한다. 또한 IRB는 연구대상자의 보호 문제와 함께 관련 법률을 준수하여 연구를 진행할 수 있게 연구자를 보호하는 역할도 수행한다. 연구자들에게 연구윤리지침을 준수하면서 연구가 진행될 수 있도록 필요한 정보를 제공하여 IRB의 사전승인을 받게 한다.

**권한과 의무**: IRB는 연구계획 및 승인된 연구에 대해 다음 각 호의 권한을 가진다.

1) IRB는 외부 기관 또는 연구자로부터 심의 요청된 연구에 대하여 다음의 사항을 심의한다.
   ① 연구계획서의 윤리적 · 과학적 타당성
   ② 연구대상자 등으로부터 적법한 절차에 따라 동의를 받았는지 여부
   ③ 연구대상자 등의 안전에 관한 사항
   ④ 연구대상자 등의 개인정보보호 대책
   ⑤ 그 밖에 생명윤리 및 안전에 관해 필요한 사항
2) IRB는 연구계획의 검토를 위해 연구자에게 추가적인 정보 제공을 요구할 수 있다.
3) IRB는 필요한 경우 위원회에서 승인된 연구과제의 수행 중 진행과정 및 결과에 대하여 조사 · 감독한다.
4) 그 밖에 생명윤리 및 안전을 위한 다음 각 목의 활동을 한다.
   ① 해당 기관의 연구자 및 종사자 교육
   ② 취약한 연구대상자 등의 보호 대책 수립
   ③ 연구자를 위한 윤리지침 마련

46) 연구에 참여하는 연구대상자의 윤리적인, 과학적 문제발생을 연구 시작 전에 심사하여 연구계획을 승인하는 기구로서, 해외 연구기관에 설치되어 있다. 우리나라의 경우 임상실험 과정에서 윤리적인 문제가 발생할 수 있는 생명공학 및 병원연구소에 설치되어 있다.

〈표 Ⅵ-4〉의 계속

5) IRB는 [본 기관에서 수행 중인 연구의 진행과정 및 결과에 대한 조사, 점검 업무]와 [생명윤리 및 안전을 위한 본 기관 연구자 및 종사자 교육 업무]를 임상연구윤리센터에 위임할 수 있다.
6) IRB는 해당 연구의 연구대상자에게 미치는 위험 정도에 따라 연구책임자에게 적절한 주기로 보고하도록 하고 이를 지속심의한다. 지속심의 주기는 최대 1년을 넘길 수 없다.
7) IRB는 승인된 연구의 수행과정에서 연구대상자 등을 적절하게 보호할 수 있어야 한다.
8) IRB는 수행 중인 연구에 대한 조사·감독 중 예기하지 않았던 중대한 위험 또는 중대한 관련 법률의 위반 등 생명윤리 및 안전에 중대한 위협이 있다고 판단되는 경우, 심의를 거쳐 해당 연구에 대한 제한·중지 혹은 보류를 결정할 수 있다.
9) IRB는 그 밖에 보건복지부 장관이 정한 고시에 따른 업무를 수행하여 해당 사항에 대한 적절한 조치를 취할 수 있다.

출처: 서울대학교병원 의생명연구원 홈페이지,
http://hrpp.snuh.org/irb/introirb/_/singlecont/view.do, 검색일자 2015. 10. 15.

### 다. RCT 평가 관련 전문가 네트워크 구축을 통한 지식공유

마지막으로, RCT 평가의 보급과 확산을 위해서는 관련 전문가 네트워크를 구축하여 지식과 경험을 공유하는 것이 필요하다. 사실 정책평가에 있어 RCT와 같은 실험설계방법에 대한 기초지식은 석사과정의 연구 설계 수업에서 다루고 있는 내용이다. 실무에서도 재정사업 심층평가 지침을 통하여 RCT의 방법론이 이론적으로 설명되고 있다. 그러나 이론적인 지식 외에 RCT 평가과정에서 발생할 수 있는 장애요인을 해결하고 실행방법을 구체적으로 적용할 수 있는 평가전문가는 국내에 매우 부족한 실정이다. 이러한 전문가의 부족은 RCT 평가가 미국이나 영국에 비해 국내에서 활성화되지 않고 있는 이유 중의 하나이다. 따라서 RCT를 통한 재정사업 평가의 보급을 위해 평가 관련 전문가 네트워크 구축과 지식공유가 필요하다. 이러한 전문가 그룹은 연구논문 작성을 위한 사회실험 자료 분석뿐만 아니라 실제 정책과 사업을 분석하여 RCT에 기반한 연구 및 평가계획을 수립하고 이를 현실 정책평가에 활용할 수 있는 방법과 절차를 제시할 수 있다. 또한 평가 관련 연구 및 실제 평가과정에서 얻어진 지식과 경험을 전문가 그룹에서 적극적으로 공유하여 RCT 기반 정책평가 확산에 기여할 수 있다.

# 참고문헌


강창희 · 이정민 · 이석배 · 김세움, 『관광정책 및 관광사업 프로그램 평가방법』, 문화체육관광부, 2013.

권수진 · 임대근 · 류현아, 「멘토링을 통한 수학학습부진아의 수학학습태도 변화에 대한 사례연구」, 『East Asian Mathematical Journal』, Vol. 30, No. 2, 영남수학회, 2014, pp. 123-148.

기획재정부, 「2013 재정사업자율평가 지침」, 2013.

노대명, 「희망리본 시범사업과 자활사업 연계 방안: 공급기관 확충을 중심으로」, 『한국조세연구원 세미나자료』, 2012, pp. 37-57.

박노욱 외, 「희망리본사업의 본 사업화 방안」, 『희망리본 시범사업 성과평가 및 본 사업 추진방안』, 한국조세연구원, 2012, pp. 54-122.

박상곤, 「관광정책 평가방법 및 사례발표: 여행바우처 사업의 효과분석을 중심으로」, 『한국행정학회 동계학술발표논문집』, 2013, pp. 154-197.

신을진 · 이일화, 「학습코칭프로그램이 학습부진아의 학습전략에 미치는 효과」, 『아시아교육연구』 Vol. 11, No. 4, 서울대학교 교육연구소, 2010, pp. 145-165.

오영민, 「재정사업성과평가의 운영성과와 제도적 개선방안」, 『재정포럼』 12월호, 한국조세재정연구원, 2014, pp. 6-22.

유경준 · 강창희 · 최바울, 「사회보험료 지원사업(두루누리 사업)의 효과: 현대성과평가론의 적용」, 『한국재정학회 추계학술대회 논문집』, 2015, pp. 1-16.

이삼열 · 정의룡 · 이은하, 「시범사업에 관한 탐색적 연구: 보건복지가족부 사업을 중심으로」, Working Paper, 한국행정학회, 2015, pp. 1-40.

한국개발연구원, 「2007 재정사업심층평가 지침」, 2007.

_____, 「2014 KDI 예비타당성조사 지침」, 2013.

한국조세재정연구원, 재정사업 심층평가 내부자료.
홍승현 · 원종학, 『적극적 노동정책의 재정효율성 평가방법에 관한 연구』, 한국조세재정연구원, 2013.
Abdul Latif Jameel Poverty Action Lab, Executive Course Lecture Notes.
Behavioral Insight Team, "Applying behavioural insights to reduce fraud, error and debt," 2012.
_____, "Applying Behavioural Insights to Organ Donation: preliminary results from a randomised controlled trial," 2015.
_____, "the behavioural Insights Team Update Report 2012-2015," 2015.
Duflo, E., Glennerster, R. and Kremer, M., "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, Vol. 4, Chapter 61, 2007.
Office of Management and Budget, "What Constitutes Strong Evidence of a Program's Effectiveness," 2004.
_____, "Memorandum to the Heads of Department and Agencies," 2013.
Snipes, J., Huang, C., Jaquet, K. and Finkelstein, N., "The Effects of the Elevate Math Summer Program on Math Achievement and Algebra Readiness," *Making an Impact, Institute of Education Sciences*, US Department of Education, July 2015.

**〈웹사이트〉**

서울대학교병원 의생명연구원 홈페이지, http://hrpp.snuh.org/irb/introirb/_/singlecont/view.do, 검색일자 2015. 10. 15.
Abdul Latif Jameel Poverty Action Lab, http://www.povertyactionlab.org, 검색일자 2015. 9. 15.
BBC, Nudge unit sold off to charity and employees, http://www.bbc.com/news/uk-politics-26030205, 2014, 검색일자

2015. 9. 15.

Behavioral Insight Team,

http://www.behaviouralinsights.co.uk/people, 검색일자 2015. 9. 15.

Coalition for Evidence-based Policy

http://coalition4evidence.org/low-cost-rct-competition/, 검색일자. 2015.9.15

Institute of Science Education

http://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf, 검색일자 2015. 10. 15.

Social and Behavioral Science Team, https://sbst.gov/, 검색일자 2016. 9. 15

# 부록: J-PAL Executive Course 강의내용[47)]

## 1. 평가의 정의(What is Evaluation?)

□ 프로그램 평가(Program Evaluation)

○ 프로그램 설계·수행 과정의 책임성(accountability) 확보를 위한 목적에서 평가 실시

- 프로그램 설계·수행 과정에는 정책결정자, 기부자, 연구자 등 다양한 이해관계자가 참여하며,
- 프로그램의 영향 혹은 결과가 효과적인지 여부는 평가를 통해 충족될 수 있음

□ 프로그램 평가(Program Evaluation)의 5가지 구성 요소

① 요구 분석(Needs Assessment)

② 프로그램 이론 평가(Program Theory Assessment)

③ 과정 평가(Process Evaluation)

④ 영향 평가(Impact Evaluation)

⑤ 비용 효과 분석(Cost Effectiveness analysis)

□ 요구 분석(Needs Assessment): 문제 정의하기

○ 체계적 접근방법(계획-실행-평가)에 의거하여 프로그램을 개발하는 과정에서, 관련된 개인·집단의 요구를 파악하고 해결하기 위하여 관련 자료를 수집·활용하는 기법

○ 현재 상태를 점검하고 이용 가능한 자원을 확인하며, 예상되는 문제

47) 저자가 수강한 Abdul Latif Jameel Poverty Action Lab의 「Executive Course Lecture Notes」의 내용을 정리하여 구성

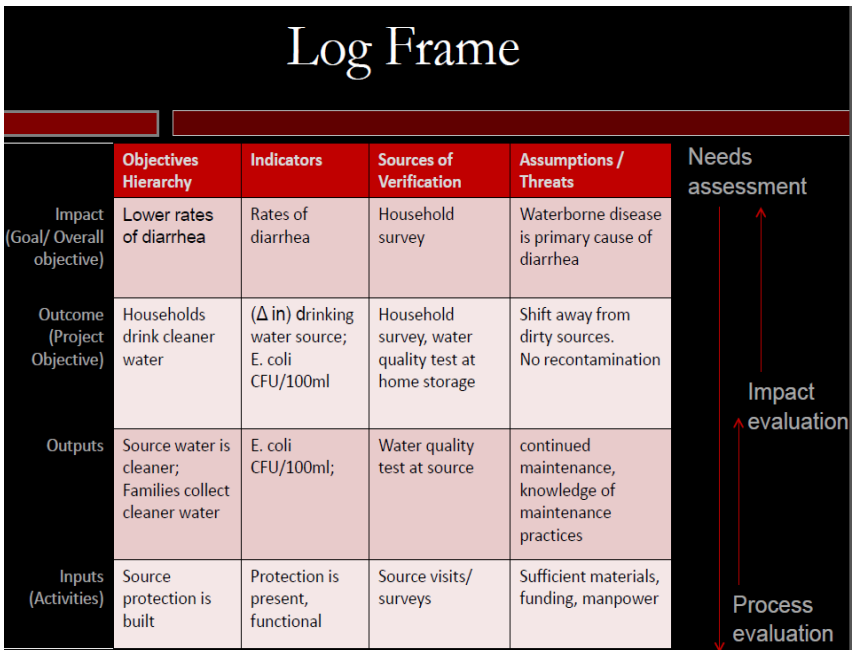
점 및 해결책 등을 분석함

□ 프로그램 이론 평가(Program theory assessment): 변화를 위한 청사진

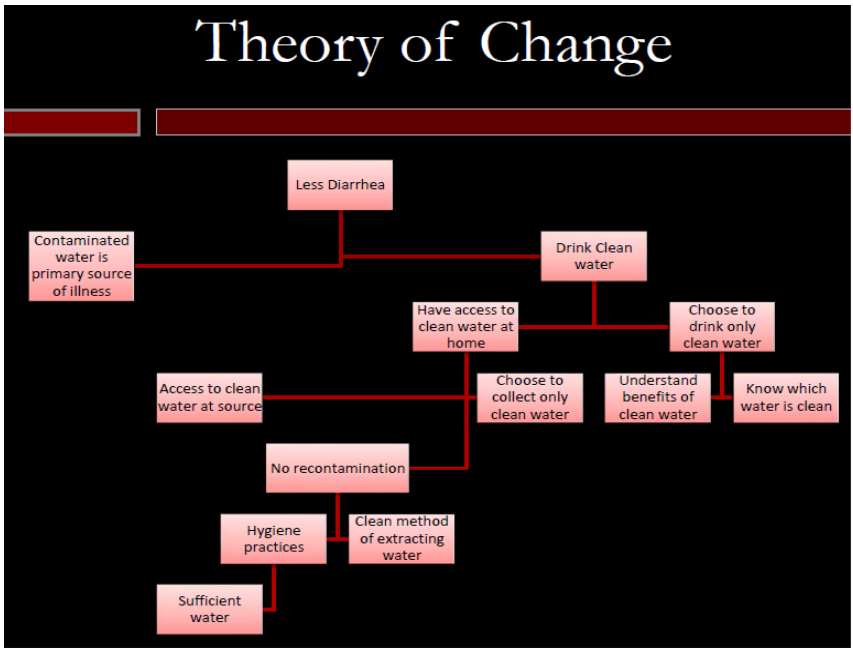
○ 프로그램이 어떤 과정 또는 절차를 통해 작동하고 실행되는가를 개념화하여 표현하는 기법

- '이론상으로, 프로그램이 문제를 어떻게 해결하는가?'와 관련됨

○ 논리적 프레임워크(LogFrame, LFA)



○ 변화 이론(theory of change)




○ 결과 프레임워크(results framework)

○ 결과 지도(outcome mapping)

□ 과정 평가(Process Evaluation)

○ 프로그램이 진행되는 '과정'에 대한 평가로, 시행상의 문제점을 파악하고 그 개선방안을 탐색하는 것을 목적으로 함

- 프로그램의 운영 및 활동에 대한 분석을 근거로 프로그램 내용을 수정·변경하거나 프로그램의 중단·축소·유지·확대 여부를 결정하는데 도움을 줌
- 예시: 원래의 계획대로 수요/공급 활동이 이루어졌는가?

□ 영향 평가(Impact Evaluation)

○ 프로그램의 실시로 인하여 유발되는 효과, 변화 등에 초점을 두는 평가를 의미함

- 영향(primary outcome, impact) 측정법: '프로그램 부재 시 무슨 일이 발생하였는가?'로 판단
- 프로그램 참여자(실험집단)를 비참여자(비교집단)와 비교하여 프로그램의 상대적 효과성 측정
- 프로그램의 영향=(프로그램으로 인해 발생한 일(변화)) - (프로그램이 시행되지 않았을 때 발생하였을 일)

○ 영향(impact)과 과정(process)의 차이점
- 과정(process): 무슨 일이 발생했는지를 서술
- 영향(impact): 무슨 일이 발생했는지와 프로그램 부재 시 무엇이 발생했을 것인지를 비교(그 차이를 서술)

○ 무작위 평가(randomized evaluation)
- 영향평가의 주된 어려움은 '프로그램에 참여하지 않았던 사람들'을 찾아내는 것이라 할 수 있음
- 프로그램이 시행되지 않은 상태와 비교해보았을 때, 얼마나 많은 사람들이 프로그램으로 변화하였는지를 측정→counterfactual

○ 사후가정(counterfactual)의 구성
- counterfactual: 프로그램이 시행되지 않은 상태를 가정→통계적으로 동일한 실험군과 비교군을 설정하여 프로그램의 효과성을 살펴볼 수 있음
- 예시: 프로그램에 참여하지 않은 사람들이 설사병을 앓은 수준이 '염소(chlorine)가 없었던 상태에서 사람들이 설사병을 앓았던 수준'과 동일할 것이라는 가정을 한다면, 프로그램에 참여하지 않은 사람들을 측정할 수 있음→이를 비교집단으로 가정

□ 비용 효과 분석(Cost Effectiveness analysis)

○ 여러 정책대안 가운데 가장 효과적인 대안을 찾고자 각 대안이 초래할 비용과 효과를 비교·분석하는 기법
- 동일한 효과를 달성하는데 다양한 방법들 간의 비교를 통해 가장 비용이 적게 드는 방법을 찾거나,

- 규모 및 비용이 정해진 경우, 비슷한 수준의 비용을 지출하나 가장 높은 수준의 산출 혹은 결과를 달성하는 대안을 찾는 방법을 활용

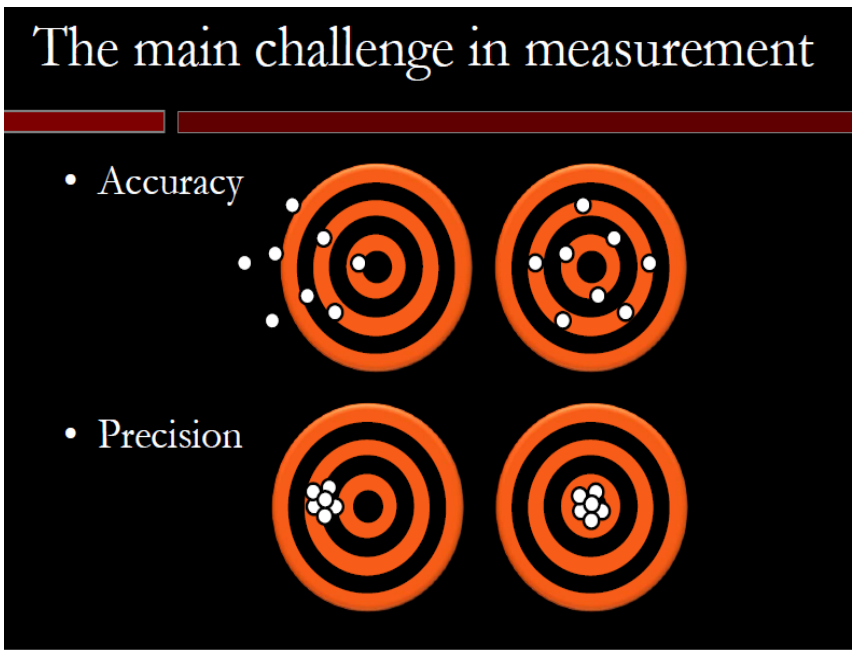
## 2. 측정지표(Measurement: Outcomes, Impact, and Indicators)

□ 무엇을 측정하는가?: 변화 이론, 결과, 지표

□ 정확성과 정밀성

○ 정확성(accuracy): 실험 결과의 질에 대한 평가 척도(진실을 말하는가) → 실험 시스템의 문제

○ 정밀성(precision): 실험 방법의 질에 대한 평가 척도(동일한 결론이 반복되는가) → 실험자의 기술, 능숙도 등에 달린 문제



□ 타당성(validity)과 신뢰성(reliability)

○ 타당성: '특정한 개념이나 속성을 측정하기 위해 개발한 지표가 그 속성을 정확히 반영할 수 있는가'를 의미(e.g. IQ 테스트 → 지능)

○ 편향되지 않은 답을 얻기 위한 방법: 이하 현상을 제거

- 사회적 바람직성에 의한 편향(social desirability bias): 사람들이 사회적으로 인정받을 수 있는 방향으로 대답하는 경향
- framing effect: 질문이나 문제 제시 방법(틀)에 따라 사람들의 선택이나 판단이 달라지는 현상
- 회상 편중(recall bias)
- 기준점 편향(anchoring bias): 처음 제시된 정보가 기준점이 되어 판단에 영향을 미치는 현상

○ 신뢰성: 연구대상을 반복 측정했을 때 일관성 있는 결과를 보이는 정도를 의미함(vs. "noisy").

○ 결과 측정의 어려움

- 사람들이 잘 모르는 것들: 특히 시간에 걸쳐 추정하는 어떤 것. 오류 및 부족한(질이 나쁜) 평가를 기억하는 경향 → 해결전략: 일관성 검사, 동일한 지표에 대한 여러 번의 측정
- 사람들이 말하기 원치 않는 것들: 알코올, 약물 사용 등과 같이 사회적으로 "위험"하거나 고통스러운 것 → 해결전략: 무거운 질문부터 시작하지 말 것. 응답자의 사생활 보장, 가능하다면 간접적으로 정보 얻기 등
- 추상적인 개념: 협상력, 사회 통합 등 지표를 측정하기 어려움 → 해결전략: '추상적인 개념' 측정 시 3단계를 거침(추상적인 개념 정의, 개념 측정의 역할을 하는 결과 선택, 결과 측정에 적합한 질문 설계
- (항상) 직접적으로 측정되지 않는 것들: 부패, 사기, 차별 등과 같이 직접적으로 질문하거나 관찰할 수 없는 결과를 측정하는 경우 → 해결전략: 기존문헌의 검토(literature review) – 기존의 연구내용

들을 사전에 검토
- 가장 직접적으로 잘 측정되는 것들: 말보다 행동으로 나타나는 행동 선호도 → 해결전략: 구체적인 규약 개발, 동일한 개인 환경 하에서 행해진 행동 측정에 대한 자료 수집

□ 자료원(sources of data): 행정 자료, 기타 2차 자료, 1차 자료

□ 자료 수집: 품질 관리, 조사원 교육, 조사원의 성별 구성, 자료 보안, 비용 등을 고려

## 3. 무작위 추출(Why Randomize?)

□ 사후 가정(counterfactual)
○ 프로그램 참여자가 프로그램의 부재 하에서 경험한 상태를 의미(즉, 프로그램에 참가하지 않은 상태)
- 특정 인구집단과 유사한 성질을 가진 집단을 대조군으로 삼아 연구 진행
- 실제로는 같지 않지만 우리는 counterfactual한 결과를 살펴보기 위해서 노출된 집단(exposure group)과 비노출(non-exposure group)이 같은 성질을 가진 인구집단이기를 기대함
- '프로그램 미참가자들'을 찾아내는 일이 현실적으로 어려움

○ counterfactual 구성: 일반적으로 프로그램에 참여하지 않은 개인으로 구성된 집단을 선택. 이 집단은 대조군(control group) 또는 비교군(comparison group)으로 불림. 영향 평가 설계에서 이 집단을 어떻게 선택하느냐가 중요함
- 비교 집단의 선택: 한 가지를 제외한 모든 점에서 참여자 집단은 정확히 동일하도록(exactly like) 집단 선택 → 그들의 프로그램에 대한 노출을 평가

- 프로그램 참여군과 대조군 사이의 결과에 있어 차이점에 기인한다(attribute)고 할 수 있기 위함
- 영향 평가 방법: 무작위 실험(Randomized Experiments, Randomized Controlled Trials(RCTs) 등), 비실험 또는 준실험 방법(Pre-Post (before vs. after), Simple Difference, Differences-in- Differences), 다중 회귀(Multivariate Regression), 단절 시계열(Interrupted Time Series) 등)

□ 무작위 실험의 기본 사례

○ 프로그램 지원자의 sample을 채취, 그들을 무작위로 할당(randomly assign)

- treatment group: 치료 제공
- control group: 평가 기간 동안 치료를 받지 못함
- 실험의 주요 이점: 실험 초기에 참여군과 대조군의 구성원이 통계적으로 다르지 않기 때문에, 이후 그들 사이에 발생하는 차이는 다른 요소가 아닌 프로그램에 기인할 수 있음

□ 몇 가지 변화

○ 여러 참여군으로 할당: 개인이나 가정 이외의 단위로 지정(건강센터, 학교, 지방자치단체, 마을)

○ 실험 구성의 주요 단계

1. 신중한 연구 설계
2. 치료군 혹은 대조군에 무작위로 인구 할당
3. 기준 자료(baseline data) 수집
4. 무작위 할당에 대한 검증
5. 실험의 무결성(integrity of experiment)이 타협되지 않도록, 과정에 대한 모니터링 실시
6. 치료군과 대조군 간의 후속 자료(follow-up data) 수집
7. 치료군의 평균 결과와 대조군의 평균 결과를 비교함으로써 프로

그램 영향 추정

8. 프로그램의 영향이 통계적/현실적으로 중요한지 평가

□ 개념 논증(conceptual argument) 개념 변수

○ 적절히 설계·구성된 무작위 실험은 프로그램의 영향을 측정하는 가장 신뢰할 만한(most credible) 방법

○ 왜 "가장 신뢰할 만한(most credible)"인가?

- 실험 초기에 치료군과 대조군이 체계적으로 다르지 않기 때문에, 이후 그들 간에 발생하는 차이는 다른 요소가 아닌 프로그램에 기인할 수 있음
- 통계적으로 동일한 실험군과 비교군을 구성함으로써 프포그램이 성과가 차이가 있는지를 볼 수 있기, 성과에 영향을 끼칠 수 있는 외부요인들을 미리 없앨 수 있다는 장점이 있음: 편향(bias) 없는 영향 평가

□ 결론

○ conceptual argument: 프로그램의 영향을 평가하는 여러 가지 방법이 있으나, 적절히 설계·구성된 무작위 실험은 프로그램의 영향을 측정하는 가장 신뢰할 만한(most credible) 방법

○ Empirical argument: 다른 방법은 다른 영향 추정치를 생성할 수 있음

## 4. 무작위 추출의 방법(How to randomize)

□ Unit of randomization

○ 무작위 실험을 설계 시 그 실험 단위(분석단위)에 대한 적절한 선택이 매우 중요함(개인 혹은 일정 그룹 단위)

○ 적절한 실험 단위 설계를 위해서는 다음의 내용을 고려해야 함

- 프로그램(처치)의 대상이 되는 단위

- 실험 분석 단위

○ 결정 요인

- nature of the treatment
- 영향의 범위
- 가용 데이터 범위
- 통계적 검정력
- 일반적으로 가장 적절한 분석단위는 기본적으로 프로그램의 대상이 되는 단위(unit of intervention)임

□ 현실적인 제약

○ 실험 대상이 되는 프로그램의 정책방향, 실험대상 선정의 공정성 문제 등

○ 자원의 제약

○ 처지집단과 비교집단의 상호작용 등으로 인한 효과 측정 오류

○ 실행 방법

○ 샘플의 크기

□ Methods of randomization

○ 현실적인 제약에 따라 적절한 무작위 실험 방법을 선택할 필요가 있음

○ 무작위 실험에 적절히 적용할 수 있는 방법들을 간략히 소개함

〈무작위 추출의 유형〉

| 방법 | 적용 | 장점 | 단점 |
|---|---|---|---|
| basic lottery | 프로그램 대상자 혹은 지원자들이 많은 경우 | • 익숙함<br>• 이해가 쉬움<br>• 실험 설계가 간단함<br>• 대중적인 방법임 | • 비교집단의 데이터 수집이 용이하지 않음<br>• 특정 그룹별로 탈락률이 다를 수 있음(모집단과의 특성에 차이가 있을 수 있음) |
| phase-In | • 실험이 장시간 가능한 경우<br>• 결국 모든 대상자들에게 처치를 할 수 있는 경우 | • 이해가 쉬움<br>• 현실적 제약에 대한 이해(설명)가 쉬움<br>• 결국 모든 사람들에게 혜택이 돌아가기 때문에 비교집단이 실험 방법에 대해서 비교적 쉽게 수긍함 | 단기적인 효과만 관찰이 가능하기 때문에 장기적 효과에 대해 측정하기 어려움 |
| rotation | • 모든 사람이 적정시점에 어떤 혜택을 받아야 하는 경우<br>• 자원의 제약으로 동시에 프로그램의 혜택을 받기 어려운 경우 | • phase-In 방법보다 데이터 수집이 용이함 | 장기적 효과를 측정하기 어려움 |
| encouragement | • 모든 사람을 대상으로 프로그램이 수행되는 경우<br>• 참여자 비율이 적으나, 인센티브로 참여자를 모집하기 비교적 쉬운 경우 | 프로그램이 개인을 기준으로 제공되는 것은 아니어도 개인수준에서 무작위 배정이 가능함 | • 인센티브를 받게 되는 사람의 효과만 측정 가능함<br>• 충분한 자원(유인책)이 필요함<br>• 프로그램 참여에 대한 유인이 효과에 영향을 미칠 수 있음 |

□ Multiple arms and Stratification

○ Multiple arms

- 실험 설계 시 서로 다른 처치를 받는 다양한 처치집단을 동시에 설계하는 것이 가능함
- 이 경우 다양한 처치를 조합하여 실험함으로써 다양한 결과를 얻을 수 있다는 장점이 있음

○ Stratification
- 샘플의 양이 매우 적을 경우 유용함
- 샘플을 서로 다른 하위그룹으로 분류한 뒤에 하위그룹끼리 처치그룹과 비교그룹을 선정함
- 샘플 수가 적은 상황에서도 검정력을 높일 수 있다는 장점이 있음
- 다만, 하위그룹을 너무 많이 나눌 경우 실험이 복잡해질 가능성이 있어 주의할 필요가 있음

## 5. 샘플링 및 샘플링 크기(Sampling and Sample Size)

□ 대수의 법칙(The Law of Large Numbers)
○ 대수의 법칙이란, 어떤 표본을 관찰할 때 그 관찰 횟수가 많아질수록 그 평균값은 실제 평균에 가까워지는 법칙을 의미함
○ 일반적으로 표본의 수가 커질 경우 확률적으로 그 추정의 정밀도가 향상됨

□ 중심극한정리(The Central Limit Theorem)
○ 중심극한정리란, 표본의 수가 일정 수준 이상이 될 경우 표본평균의 분포가 평균은 모집단의 평균값, 분산은 모집단의 분산/표본의 크기 값을 가지게 되는 정규분포를 따른다는 법칙을 의미함
○ 이에 따르면, 표본의 수가 많을수록, 표본평균의 분산(변동성)은 줄어드는 경향을 보임

□ 정확도(Accuracy) vs. 정밀도(Precision)
○ 정확도(Randomization): 참값에 근접한 정도
- 일반적으로 systemetic error(bias)와 연관이 있음
○ 정밀도(sample size): 측정값들의 분포
- 측정의 재연성과 연관이 있으며, random error와 연관이 있음

□ 통계에서의 기본적인 고려사항

○ 결과에 대한 신뢰성

○ 샘플의 크기

□ 표준편차(Standard deviation)와 표준오차(Standard error)

○ 표준오차(Standard error)란, sampling distributions의 표준편차를 의미함

○ 분산: $\sigma^2 = \sum(\text{observation value-average})^2/N$

○ 표준편차: $\sigma$

○ standard error = $\sigma / \sqrt{N}$

○ 일반적으로 샘플 사이즈가 커질수록 SE는 작아지는 경향을 보임

□ 가설검증

○ 제1종 오류(Type Ⅰ error)

- 통계적 가설 검정 시 귀무가설이 옳음에도 이를 기각하는 오류
- 실제로 효과가 나타나지 않았으나, 효과가 있다고 판단하는 오류

○ 제2종 오류(Type Ⅱ error)

- 통계적 가설 검정 시 대립가설이 옳음에도 이를 기각하는 오류
- 실제로 효과가 있었음에도 효과가 없다고 판단하는 오류

○ 유의수준(significance level)

- 귀무가설이 참임에도 이를 기각할 확률
- 제1종 오류의 최대 허용 범위를 나타냄
- 일반적으로 5% 혹은 1% 수준으로 제한하고 있는 경우가 많음

□ 통계적 검정력(Power)

○ 검정력이란, 통계적 검정에서 귀무가설이 거짓일 때(대립가설이 참인 경우), 귀무가설을 기각시키는 확률임

○ 검정력을 결정하는 주요한 요인으로는 효과의 크기(+), 샘플의 크기(+), 분산, 처치그룹과 비교그룹의 비율, Clustering이 있음

○ ICC($\rho$ )

## 6. 실험적 방법론의 위협요인(Threats and Analysis)

□ Ramdomize Evaluation 설계 및 효과에 대한 분석 시 다음의 내용을 고려할 필요가 있음

- ○ Bias의 발생
- ○ 외적 타당도
- ○ 비용 효과성

□ 발생할 수 있는 Bias의 종류

- ○ Attribution bias
- ○ Spillover bias
- ○ Sample selection bias

□ Intention to Treat(ITT)와 Treatment on Treated(TOT)

- ○ Intention to Treat(ITT)
  - ITT란, 실제로 처치그룹 중 처치를 받은 사람을 대상으로 두 그룹(treatment group과 control group)을 비교하는 것이 아니라, 원래 배정된 처치그룹(treatment group)과 비교그룹(control group)을 비교하여 효과를 분석하는 방법임
  - 이를 위해서는 처치그룹과 비교그룹의 관찰 대상자의 모든 데이터를 수집할 수 있어야 하며, 처치 여부에 상관없이 원래 실험 설계 시 배정된 그룹으로 분류하여 그 효과에 대해서 분석해야 함
  - 이는 실험 효과 분석 시 Bias의 발생을 최소화하고, 최초 설계 시 의도한 효과에 대한 정확한 결과를 얻기 위함임
  - 하지만, 실제 실험을 수행할 경우 결측치의 발생 혹은 처치그룹과 비교그룹 간의 spillover에 의해 원래 측정하고자 하는 효과를 작게 혹은 크게 나타낼 가능성이 있음
- ○ Treatment on Treated(TOT) approach: Instrumental Variables
  - 정확한 효과 분석을 위해 도구변수를 활용하여 그 효과를 분석해 볼 수 있음

□ 외적 타당도

○ 외적 타당도란 연구결과가 일반화될 수 있는지를 나타낸 정도를 의미함
- 실험조사에 대한 반응성
- 연구표본의 대표성

○ 실험조사에 대한 반응성: 실험의 대상자가 스스로 실험의 대상이 되고 있음을 인지하고 이에 따라 나타나는 의식적 반응으로 이에 따라 연구결과에 영향을 미칠 수 있음
- Hawthorne effect
- John Henry effect

○ 연구표본의 대표성: 표본집단(실험집단)의 조건(성격)이 모집단의 성격과 유사해야 실험 결과를 일반화할 수 있음
- 대규모로 확대가 가능한 실험인지 여부, 실험 샘플의 대표성, 결과에 대한 민감도 등의 요인에 의해 결정됨

## 7. 실제 실험적 방법론의 적용사례

□ 선정사례: 브라질의 Game Changer에 대한 실험평가

□ 사례개요

○ 브라질의 도시빈민 지역의 아동들은 높은 빈곤율로 인한 문화적 혜택의 부재 때문에 낮은 학교성적, 퇴학, 범죄 등에 노출되어 있음

○ 이에 브라질의 도시빈민 지역의 비영리기관은 빈민지역 학생들의 방과 후 프로그램 개설하여 학생들의 학업성적을 높이고자 함

○ 단, 방과 후 프로그램은 아카데믹 프로그램이 아닌 스포츠, 무용강좌 등 Soft Skill을 키울 수 있는 프로그램 구성
- 최근 연구결과는 학생들의 학업성과의 중요한 요인으로서 직접적 학습지도와 학생들의 감정적 역량을 소개하고 있음
  ※ 감정적 역량(Soft Skill)의 주요개념

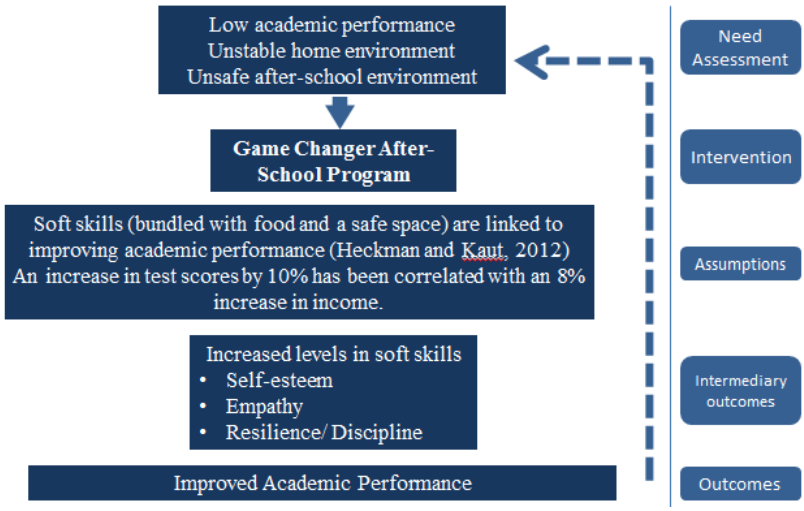
a. 자기존중감(Self-Esteem)

b. 감정이입(Empathy)

c. 수용성(Resilience)

d. 자기절제(Discipline)

※ 사례개요 개념도




□ 실험설계

○ 분석단위는 방과 후 프로그램에 참여하는 개인학생으로 설정

○ 분석대상 학생은 빈곤선 이하 학생을 대상으로 하고 분석대상 학생을 실험집단과 통제집단에 무작위로 배정

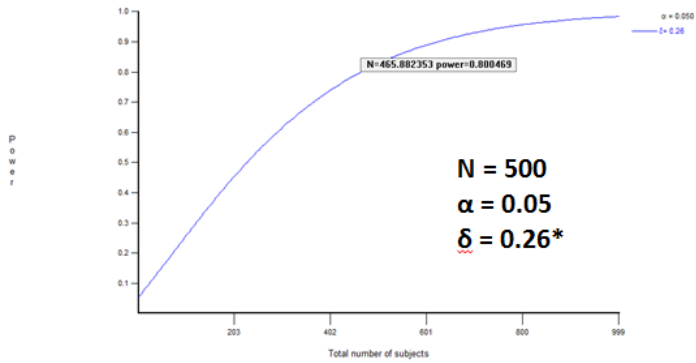
○ 방과 후 프로그램 1년 후에 학생들의 Soft skills와 시험성적이 얼마나 향상되었는가를 통계적으로 비교

| 결과 | 성과지표 | 자료원천 | 자료 생성주기 |
|---|---|---|---|
| 학업성과 | 시험성적 | 표준학력평가 성적 | 매년 |
| 감정능력 (Soft Skill) | 감정능력지수<br>• 유연성: Resilience (Hanson et al, 2004)<br>• 자기존중감: Self-esteem (Scheier et al, 1994)<br>• 감정이입: Empathy (John, 2012) | 설문조사 자료 | 매년 |

□ 샘플 수

○ 일정 수준의 통계적 유효성을 유지하기 위해 통계 파워의 계산을 통하여 적정 샘플 수를 산정

○ 상기 실험설계의 경우, 시험성적 10% 향상, 감정능력(Soft skill) 15% 증가를 가정할 때, 80%의 통계적 파워 수준에서 약 500명의 샘플이 필요할 것으로 계산됨



□ 실험의 위협요인

○ 상실요인: 실험 도중 실험 참가 학생 수가 포기하여 줄어들 가능성이 있음

- 저소득층이라 부모가 경제적으로 어려울 시 방과 후 프로그램을 이용하지 않고 근로행위에 종사할 가능성이 있음

○ Spillover Effect: 실험에 참여하지 않은 학생들과 부모들이 실험에 참여하는 친구나 가족을 통해 방과 후 프로그램을 인지하고 다른 기관에서 제공하는 방과 후 프로그램에 참가할 가능성이 존재함

- 다만, 상대적으로 감정능력을 키워주는 스포츠나 무용 방과 후 프로그램은 프로그램 참가비가 비싸기 때문에 저소득층의 통제집단 학생들이 별도의 프로그램에 참가할 가능성은 적음

국문요약

# 재정사업사전검증 체계 강화를 위한 연구
## – RCT 도입방안을 중심으로

오영민 · 박노욱 · 강희우

새로운 재정사업을 충분한 사전점검 없이 정치적인 요구에 의해 급격하게 도입하는 사례가 증가하면서, 시행착오로 인한 사회적 비용이 커지고 있다. 이러한 부작용을 막기 위해 사업시행 전에 예비타당성조사가 운영하고 있으나 정책적 효과를 정확히 계량화할 수 없는 교육인 복지사업에 대한 정확한 사전평가는 만족할 만한 수준이 아니다. 예비타당성조사가 나름대로 재정사업의 무분별한 시행을 막는 완충작용을 해왔을지라도 사업의 정확한 편익을 추정할 수 없는 교육 또는 복지사업에는 적용하기 어려운 한계가 있기 때문이다.

이런 의미에서 본 연구는 우리나라에서 최근 급속하게 확대되고 있는 복지나 고용 관련 사업을 사회실험 방식의 RCT 평가할 수 있는 가능성을 검토하였다. 본 연구에서는 우선 RCT 평가기법을 이론적으로 기술하고 준실험 방식의 여러 평가 기법을 소개하였다. 그 후 해외의 정부와 민간분야에서 RCT가 어떻게 정책평가에 활용되고 있는지를 소개하였다. 또한 아직 RCT가 크게 활용되고 있지 않은 우리나라의 적용가능성을 사전사후평가와 정책분야별로 탐색하였으며 실험의 적용과정에서 발생할 수 있는 장애요인과 대처방안을 제시하였다. 특히, 우리나라의 고용정책과 해외의 교육정책의 효과를 사회실험 방식으로 평가한 사례를 소개하였다. 마지막으로 우리나라 실정에서 RCT 평가를 도입할 수 있는 방안을 제도적, 기술적, 환경적 측면에서 구체적으로 제시하였다.

# The study on how to strengthen the ex-ante evaluation of financial programs
## - Introducing Randomized Control Trial (RCT)

Youngmin Oh · No Wook Park · Heewoo Kang

Nowadays, new financial programs have increased without making detailed ex ante evaluation because of political factors. These programs are causing social costs such as increased fiscal deficit. To prevent such side effects, the budget office has operated the ex-ante feasibility that evaluates the relevance of big new programs, but its effects are still questionable. Although the ex-ante feasibility test has contributed to saving wasteful budget by stopping ineffective programs, the benefits of educational or welfare program cannot be accurately measured due to the lack of numeric parameters.

In this sense, this study explores whether the effects of welfare or labor programs can be evaluated by the Randomized Control Trial(RCT). First, this study explains the theoretical issues of the RCT, and introduces several quasi-experimental policy evaluation methods. Next, this study mentions how the RCT is used in the foreign countries. This study also addresses how the RCT can be applicable to ex-ante and ex-post evaluation across policy areas in Korea. In particular, this study provides how to overcome various barriers in implementing the RCT policy evaluation. To help readers understand the experimental design, two cases of several

RCT-based policy evaluation are introduced. Lastly, the study provides how to adopt the RCT in the policy evaluation from the institutional, technical and environmental perspectives.

## 저자약력

**오영민**
연세대학교 행정학과 졸업
미국 Florida State University 행정학 박사
현, 한국조세재정연구원 부연구위원/평가제도팀장

**박노욱**
서울대학교 경제학과 졸업
미국 University of Michigan 경제학 박사
현, 한국조세재정연구원 재정성과평가센터 소장

**강희우**
서강대학교 수학, 경제학 복수 전공
미국 University of Wisconsin_Madison 경제학 박사
현, 한국조세재정연구원 부연구위원

**자료 수집 및 정리**

**신현태** 한국조세재정연구원 연구원

연구보고서 15-17

**재정사업 사전검증체계 강화를 위한 연구**
**- RCT 도입방안을 중심으로**

| | |
|---|---|
| 발 행 | 2015년 12월 31일 |
| 저 자 | 오영민·박노욱·강희우 |
| 발 행 인 | 박형수 |
| 발 행 처 | 한국조세재정연구원 |
| 주 소 | 30147 세종특별자치시 한누리대로 1924 |
| 전 화 | (044)414-2114(代) |
| 홈 페 이 지 | www.kipf.re.kr |
| 등 록 | 1993. 7. 15. 제2014-24호 |
| 정 가 | 6,000원 |
| 조판 및 인쇄 | 고려씨엔피 (02)2277-1508/9 |
| I S B N | 978-89-8191-803-3 93320 |