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

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

이 영 수 서강대학교 심리학과 박사과정

석 혜 원† 서강대학교 심리학과 부교수

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

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

---

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

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

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

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

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

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

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

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

차이점수 모형과 공분산분석 모형

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

차이점수 모형

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

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

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

공분산분석 모형

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

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

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

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

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

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

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

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

Lord의 역설

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

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

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

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

는 학기 초와 학기 말에 각각 0.89, 0.90으로 나타났다.

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

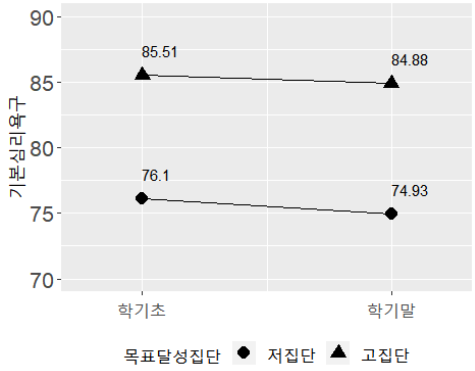

그림 1. 목표달성 집단별 기본심리욕구 평균 변화

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

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

Lord의 역설이 발생하는 조건

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

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

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

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

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

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

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

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

차이점수에 대한 논쟁

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

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

차이점수의 신뢰도

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

차이점수와 사전점수 간 부적 상관

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

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

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

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

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

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

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

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

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

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

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

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

인과추론 맥락에서 차이점수 모형과 공분산분석 모형의 비교

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

인과추론 모형

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

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

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

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

2) 반복측정 설계의 경우, 한 개인을 처치조건과 통

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

사전, 사후 시점 간 점수 변화에 대한 가정

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

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

서 평균적으로 동일하다는 가정을 나타낸다[3].

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

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

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

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

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

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

이러한 가정을 수식으로 나타내면 식 (22),

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

영가설 하에서 예측되는 결과 패턴

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

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

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

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

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

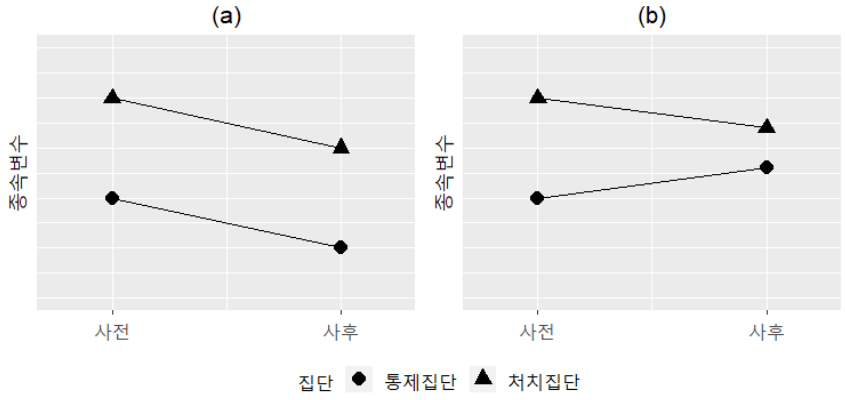

그림 2. 평균 처치효과가 존재하지 않을 때 기대되는 집단별 평균 점수 변화 양상

(30)과 같다.

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

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

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

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

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

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

Van Breukelen(2013)은 수식과 예시를 통해

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

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

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

적절한 분석 모형 선택을 위한 가이드라인

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

무선할당

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

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

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

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

비무선할당 1: 사전 점수에 기반한 집단 할당

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

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

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

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

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

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

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

비무선할당 2: 비동질적 집단 설계

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

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

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

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

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

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

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

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

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

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

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

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

비동질적 집단에 기반한 관련성 분석

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

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

표 1. 인과추론 분석 방법에 대한 가이드라인

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

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

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

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

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

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

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

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

결론 및 논의

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

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

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

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

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

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

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

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

이 앞으로 지속되어야 할 것이다.

참고문헌

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Youngsoo Lee **Hye Won Suk**

Department of Psychology, Sogang University

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

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