한국심리학회지: 일반
Korean Journal of Psychology: General
2025, Vol. 44, No. 3, 163-198
http://dx.doi.org/10.22257/kjp.2025.9.44.3.163

사전-사후 통제집단 설계에서 차이점수 분석과 공분산분석의 수행 비교: 시뮬레이션 연구*

이 영 수　　석 혜 원†

서강대학교 심리학과

차이점수 분석과 공분산분석은 사전-사후 통제집단 설계에서 처치효과 평가를 위해 널리 사용된다. 그러나, 두 방법은 항상 동일한 결과를 산출하지 않으며, 때로 부정확한 결과를 산출한다. 이에, 본 연구는 시뮬레이션을 통해 다양한 조건에서 두 방법이 처치효과 추론에 어떠한 차이를 나타내는지 살펴보았다. 구체적으로, 집단 할당 방식, 점수의 신뢰도, 점수의 안정성, 처치효과 크기, 표본 크기 등의 요인을 체계적으로 조작하여 다양한 조건에서 자료를 생성하고, 각 조건에서 두 방법의 수행을 추정의 편향, 제1종 오류율 및 검정력 측면에서 비교하였다. 그 결과, 무선할당 조건에서는 두 방법 모두 편향 없이 정확한 결과를 산출했으나, 공분산분석이 다소 높은 검정력을 보였다. 반면, 비무선할당 조건에서는 분석 방법에 따라 결과에 뚜렷한 차이가 관찰되었다. 비무선할당 중 사전점수 기반 할당에서는 공분산분석이, 비동질적 집단 설계에서는 차이점수 분석이 편향 없이 처치효과를 추정했고, 사전 진점수 기반 확률적 할당에서는 두 방법 모두 편향된 결과를 산출했다. 또한, 추정의 편향과 검정의 정확성은 각 집단 할당 방법 내에서 점수의 신뢰도와 안정성에 따라 달라졌다. 본 연구는 집단 할당 방법과 자료 특성을 고려한 분석 방법 선택의 중요성을 보여주며, 적절한 분석 방법 선택을 위한 실질적 지침을 제공한다.

주요어 : 차이점수, 공분산분석, 처치효과, Lord의 역설, 실험 설계

---

* 이 논문은 재단법인 止觀의 지원을 받아 작성되었음(202470090.01).

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

사전-사후 통제집단 설계(pretest-posttest control group design; Shadish et al., 2002; Campbell et al., 1963)는 심리학을 비롯한 다양한 연구 분야에서 처치효과(treatment effect; Maris, 1998) 혹은 인과효과(causal effect; Holland & Rubin, 1983)를 살펴보기 위해 가장 널리 사용되는 실험 설계 중 하나이다. 이 설계에서는 통제집단과 처치집단에 속한 참여자들을 사전, 사후 두 시점에 반복 측정하여 자료를 수집한다. 이때 처치집단에는 사전-사후 시점 사이에 처치가 주어지고, 통제집단에는 처치가 주어지지 않는다. 만약 처치에 효과가 있다면, 통제집단에서 나타나는 사전-사후 시점 간 변화에 비해 처치집단에서 그 변화가 더 크게 나타날 것이다.

사전-사후 통제집단 설계에서 이러한 변화의 집단차 분석을 위해 가장 널리 사용되는 방법은 공분산분석(Analysis of Covariance; ANCOVA)과 차이점수(difference score) 분석이다(Castro-Schilo & Grimm, 2018; Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013; Van Breukelen, 2006, 2013). 공분산분석은 사전점수를 공변량으로 통제하고 사후점수의 집단차를 분석하는 방법이다. 반면, 차이점수 분석은 사전-사후 측정치의 차이 즉, 변화량 자체를 종속변수로 사용하여 집단차를 분석한다.

그런데, 이 두 분석 방법은 항상 동일한 결론을 산출하지 않는다. Lord(1967)는 예시를 통해 동일한 자료에 대해 차이점수 분석과 공분산분석이 서로 다른 결과를 도출할 수 있음을 보인 바 있다. 그는 학생들의 체중 변화가 성별에 따라 다르게 나타나는지 검증하기 위해, 학년 초와 학년 말에 체중을 반복 측정하여 자료를 수집했다고 가정하였다. 이 가상의 자료에서, 학년 초 측정한 몸무게는 여학생에 비해 남학생이 평균적으로 높았지만, 남녀 두 집단 모두 학년 초에 측정한 평균 몸무게가 학년 말에 그대로 유지되었다. 차이점수 분석을 사용하여 이 자료를 분석한 결과 여학생과 남학생 모두 평균적으로 체중 변화가 없어 체중 변화에 성차가 없다는 결론이 도출되었다. 반면, 공분산분석을 사용하면 학년 초 체중을 통제했을 때 여학생에 비해 남학생이 학년 말에 더 높은 체중을 나타내어 성차가 존재한다는 결론이 도출되었다.

'Lord의 역설(Lord's paradox)'이라고 알려진 이러한 현상 즉, 차이점수 분석과 공분산분석이 서로 다른 결과를 도출하는 현상은 반복측정 자료에서 집단차를 분석할 때 방법론적 선택의 중요성을 환기시켰고, 이후 차이점수 분석과 공분산분석을 비교하는 다양한 이론적, 경험적 연구로 이어졌다. 이론적 연구들은 주로 Lord의 역설이 왜 혹은 언제 발생하는지를 설명하고(Allison, 1990; Gollwitzer et al., 2014; Holland & Rubin, 1983; Maris, 1998; Oaks & Feldman, 2001; Van Breukelen, 2006, 2013; Wainer, 1991), 차이점수 분석과 공분산분석이 각각 어떤 조건에서 정확하게 처치효과를 추론할 수 있는지 살펴보고자 하였다(Holland & Rubin, 1983; Maris, 1998). 경험적 연구들은 이론적 연구에 근거하여 차이점수 분석과 공분산분석이 서로 다른 결과를 산출할 것으로 예상되는 조건 하에서 예상된 차이가 실제로 발생하는지를 시뮬레이션에 기반하여 확인하고자 하였다(Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013; Petscher & Schatschneider, 2011; Wright, 2006).

이러한 선행 연구들은 두 분석 방법의 차이에 대한 기초적인 이해를 제공하며(Lee & Suk, 2024), 특히 시뮬레이션 연구들은 두 분

석 방법이 언제 어떻게 서로 다른 결과를 산출하는가에 대한 경험적이고 구체적인 근거를 제공한다. 그럼에도 불구하고, 기존 시뮬레이션 연구들은 각각 서로 다른 조건과 결과 지표에 초점을 맞추고 있어 그 결과를 통합적으로 이해하는 것이 쉽지 않다. 특히, 연구에 따라 시뮬레이션에서 조작한 요인들이 서로 다르거나, 동일한 요인을 고려하였더라도 그 정의나 조작 방식이 동일하지 않은 경우가 많다. 게다가, 자료 생성에 사용한 참모형(true model)이 서로 다르거나 참모형 자체를 정확히 기술하지 않은 연구도 있어, 연구 결과를 체계적으로 비교하는 것이 어렵다.

따라서, 본 연구에서는 이러한 기존 시뮬레이션 연구의 한계를 극복하고, 보다 체계적으로 차이점수 분석과 공분산분석의 결과를 비교할 수 있는 시뮬레이션 연구를 수행하고자 하였다.

본 논문의 구성은 다음과 같다. 우선, 차이점수 분석과 공분산분석을 간략하게 설명하고, 두 분석 방법을 비교하는 이론적 연구와 선행 시뮬레이션 연구들을 개관한다. 다음으로, 선행 시뮬레이션 연구의 한계를 제시하면서 본 연구의 목적을 소개하고, 이를 위한 시뮬레이션 절차와 조건을 설명한다. 이어, 시뮬레이션 결과를 제시하고, 마지막으로 결과에 대한 해석과 논의를 제시한다.

차이점수 분석과 공분산분석

차이점수 분석 모형은 식 (1)과 같이 나타낼 수 있다. 이 식은 동일한 측정 단위를 사용하여 얻은 사전점수(X_{1i})와 사후점수(X_{2i})의 차이값을 종속변수로 하고, 집단을 나타내는 더미변수(Z_i; 통제집단 $Z_i = 0$, 처치집단 $Z_i = 1$)를 독립변수로 하는 단순회귀모형이다. 이 모형에서 절편 γ_0는 $Z_i = 0$일 때 기대되는 차이점수 즉, 통제집단에서의 평균 차이점수를 나타내고, 기울기 γ_1은 Z_i가 1단위 증가할 때 기대되는 차이점수의 변화량 즉, 통제집단에 비해 처치집단의 차이점수가 평균적으로 얼마나 다른지를 나타낸다. e_{1i}는 모형의 잔차이다.

$$X_{2i} - X_{1i} = \gamma_0 + \gamma_1 Z_i + e_{1i} \quad (1)$$

이 식에서 집단변수의 기울기 γ_1이 바로 변화의 집단차를 나타낸다. 따라서, γ_1의 추정치가 곧 처치효과 추정치를 나타내고, 이 값의 유의성을 검증하면 처치효과가 유의한지 검증할 수 있다.

참고로, 본 논문에서는 식 (1)과 같이 회귀모형에 기반하여 차이점수 분석을 제시하였지만, 차이점수를 종속변수로 하여 독립집단 t-검정(혹은 ANOVA)을 실시하면 동일한 처치효과 유의성 검증 결과를 얻을 수 있다.

이와 달리, 공분산분석은 사후점수(X_{2i})를 종속변수로 하고, 사전점수(X_{1i})를 공변인, 집단(Z_i)을 독립변수로 하는 다중회귀모형에 기반하며, 이를 식 (2)와 같이 나타낼 수 있다. 여기서 β_0는 절편, β_1과 β_2는 각각 사전점수와 집단변수의 기울기, 그리고 e_{2i}는 잔차를 나타낸다. 공분산분석 모형에서는 β_2가 변화의 집단차 혹은 처치효과를 나타내며, β_2의 유의성을 검증하면 처치효과가 유의한지 검증할 수 있다. 이 때의 처치효과는 사전점수로 예측할 수 있는 것 이상으로 사후점수가 변화한 정도를 가리킨다.

$$X_{2i} = \beta_0 + \beta_1 X_{1i} + \beta_2 Z_i + e_{2i} \quad (2)$$

만약 동일한 자료를 식 (1)과 (2)를 사용하여 각각 분석했을 때, 식 (1)의 γ_1과 식 (2)의 β_2가 동일하게 추정되고 표준오차 또한 동일하게 추정된다면, 차이점수 분석과 공분산분석은 처치효과에 대한 동일한 결론을 산출하게 될 것이다. 그러나, 아래 제시된 것과 같이, 다양한 요인들에 따라 γ_1과 β_2 추정에 차이가 발생할 수 있다.

차이점수 분석과 공분산분석의 결과 차이를 가져오는 요인들

선행 연구에 따르면, 점수의 신뢰도, 사전-사후 점수 간 상관, 사전점수에서의 집단차 등에 따라 차이점수 분석과 공분산분석 간에 결과 차이가 발생할 수 있다(Castro-Schilo & Grimm, 2018; Gollwitzer et al., 2014; Van Breukelen, 2013). 이를 좀 더 명확히 이해하기 위해, 공분산분석 모형 식 (2)의 양 변에서 X_{1i}를 빼서 식 (3)과 같이 나타내보자.

$$\begin{aligned} & X_{2i} - X_{1i} \\ & = \beta_0 + (\beta_1 - 1) X_{1i} + \beta_2 Z_i + e_{2i} \end{aligned} \quad (3)$$

식 (1)에 제시된 차이점수 분석 모형과 식 (3)에 제시된 공분산분석 모형을 비교해보면, 두 모형은 차이점수가 종속변수 역할을 하고 집단변수가 예측변수 역할을 한다는 점에서 동일하지만, 공분산분석 모형에는 사전점수가 추가적인 예측변수로 포함되어 있다는 점에서 다르다고 할 수 있다(Castro-Schilo & Grimm, 2018).

또한, 다음 두 가지 조건 중 하나라도 만족하면 γ_1과 β_2가 동일한 값을 갖게 되어 Lord의 역설이 발생하지 않음을 알 수 있다(Castro-Schilo & Grimm, 2018). 첫째, 식 (3)에서 β_1이 1의 값을 갖는다면 식 (3)과 식 (1)은 동일해지고, Lord의 역설은 발생하지 않는다. 식 (3)에서 β_1은 사전-사후점수 간 안정성을 나타내는데(Kisbu-Sakarya et al., 2013), 안정성은 사전-사후 진점수 간 상관, 사전 및 사후점수의 신뢰도, 사전 및 사후점수의 집단차, 그리고 사전 및 사후점수 분산의 비율에 의해 결정된다[1]. 만약 사전-사후 점수 간에 완벽한

1) 식 (3)에서 $\beta_1 = r_{21.Z}\frac{sd_2}{sd_1}$와 같고, 여기서 $r_{21.Z}$는 Z의 영향을 제거하고 난 뒤 X_2와 X_1의 상관관계 즉, 편상관(partial correlation)을 나타내고, sd_2와 sd_1은 각각 사후점수와 사전점수의 표준편차를 가리킨다(Cohen et al., 2003). 즉, 점수의 안정성은 편상관과 사전, 사후점수의 표준편차 비율에 의해 결정됨을 알 수 있다. 이때 편상관은 다시 $r_{21.Z} = \frac{r_{21} - r_{2Z}r_{1Z}}{1 - r_{1Z}^2}$와 같이 나타낼 수 있고, 여기서 r_{21}은 사전-사후점수 간 상관, r_{2Z}와 r_{1Z}는 각각 사후점수와 집단 간 상관, 사전점수와 집단 간 상관을 나타낸다(Cohen et al., 2003). 즉, 편상관은 사전-사후점수 간 상관, 그리고 사전 및 사후점수에서의 집단차에 의해 결정된다. 그리고, 사전-사후점수 간 상관은 다시 $r_{21} = \rho_{21}\sqrt{rel(X_2)rel(X_1)}$와 같이 나타낼 수 있는데, ρ_{21}은 사전-사후 진점수 간 상관이고, $rel(X_2)$와 $rel(X_1)$은 각각 사후점수와 사전점수의 신뢰도를 나타낸다(Cohen et al., 2003). 따라서, 이를 종합하면, 사전-사후 점수 간 안정성 즉, β_1은 사전-사후 진점수 간 상관, 사전 및 사후점수의 신뢰도, 사전 및 사후점수에서의 집단차, 그리고 사전-사후 점수 간 표준편차 비율(또는 분산 비율)에 의해 결정됨을 알 수 있다. 이때,

안정성이 존재한다면 즉, 사전-사후 점수 진점수 간 상관이 1이고, 사전 및 사후점수에서 집단차가 동일하게 유지되고, 사전 및 사후점수의 분산이 동일하며, 사전 및 사후점수가 완벽하게 신뢰롭다면, $\beta_1 = 1$이 되어 Lord의 역설이 발생하지 않는다.

둘째, 만약 $\beta_1 \neq 1$이더라도(즉, 사전-사후 점수 간 완벽한 안정성이 존재하지 않더라도), 사전점수 X_{1i}와 집단변수 Z_i가 서로 상관을 갖지 않는 경우(즉, 두 집단 간 사전점수에 평균적으로 차이가 없는 경우)에는 Lord의 역설이 발생하지 않는다. 사전점수와 집단변수 간에 상관이 없으면, 집단변수 기울기 β_2는 사전점수가 모형에 포함되었는지의 여부에 따라 영향을 받지 않기 때문이다.

무선할당(random assignment) 설계의 경우, 처치집단과 통제집단에 무선적으로 참여자를 할당하므로, 사전점수에 체계적인 집단차가 존재하지 않으리라 기대할 수 있고 Lord의 역설이 발생하지 않을 것으로 예상할 수 있다. 선행 연구에 따르면, 실제로 무선할당 설계에서는 차이점수 분석과 공분산분석 모두 편향없이 처치효과를 정확히 추정함을 알 수 있다(Van Breukelen, 2006, 2013; Jennings & Cribbie, 2016). 다만, 공분산분석과 차이점수 분석은 서로 다른 잔차 분산과 자유도를 갖는데, 이로 인해 공분산분석이 차이점수 분석에 비해 더 높은 검정력을 나타내는 것으로 알려져 있다(Huck & McLean, 1975; Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013; Petscher & Schatschneider, 2011; Van Breukelen, 2013).

지금까지 Lord의 역설이 발생하지 않는 두 조건에 대해 살펴보았는데, 이 두 조건이 모두 만족되지 않는다면(즉, β_1이 1과 다르고, 사전점수와 집단변수 간 상관 또한 존재한다면), 차이점수 분석과 공분산분석이 상이한 처치효과 추정치를 산출하는 Lord의 역설이 발생한다(Gollwitzer et al., 2014). 실제 자료에서 점수가 완벽하게 안정적인 경우는 거의 없기 때문에, Lord의 역설이 드물지 않게 발생하리라고 예상할 수 있다. 또한, 윤리적, 현실적 이유로 무선할당이 가능하지 않아서 비무선할당(non-random assignment) 설계를 사용할 수밖에 없는 상황이 종종 발생하는데, 이 경우 사전점수에 체계적인 집단 차이가 존재할 수 있고, 이로 인해 사전점수와 집단변수 간 상관이 존재하여 Lord의 역설이 발생할 수 있다.

그렇다면, Lord의 역설이 발생했을 때 차이점수 분석과 공분산분석의 결과 중 어느 것이 정확한가? 이에 대한 답은 인과추론 맥락에서 두 방법을 비교한 연구들을 통해 찾을 수 있다. 이 연구들은 인과추론에 대해 상당히 방대한 내용을 다루고 있지만, 핵심은 비무선할당 중 구체적으로 어떤 집단 할당 방법을 사용했는가에 따라 인과추론의 정확성이 달라진다는 것이다(인과 추론에 대한 보다 자세한 논의는 Maris(1998) 그리고 Lee와 Suk (2024)에서 찾아볼 수 있다).

차이점수 분석과 공분산분석은 서로 다른 가정에 기반하여 인과추론 즉, 처치효과를 추정하며(Maris, 1998), 이로 인해 영가설이 참일 때(즉, 처치효과가 존재하지 않을 때) 예측하는 결과 패턴에 차이를 보인다[2](Castro-Schilo & Grimm, 2018; Van Breukelen, 2013). 차이점

$\rho_{21} = 1$이고, 사전, 사후점수가 완벽하게 신뢰로우며, $r_{2Z} = r_{1Z}$ 및 $sd_2 = sd_1$을 만족할 경우 $\beta_1 = 1$이 됨을 알 수 있다.

수 분석은 처치효과가 없다면, 사전점수에서 나타난 집단 간 차이가 사후점수에서도 그대로 유지될 것으로 예측한다. 반면, 공분산분석은 처치효과가 없다면, 사전점수에 존재하는 집단 차이가 사후에는 더 줄어들 것으로 예측한다. 예를 들어, 처치집단의 평균 사전점수가 통제집단보다 3점 더 높았다면, 차이점수 분석은 사후에도 두 집단 간 평균 차이가 동일하게 3점으로 유지될 것으로 예측한다. 반면, 공분산분석은 사후 시점에 이 차이가 3점보다는 줄어들 것으로 예측한다.

공분산분석이 예측하는 이러한 결과 패턴은 다음과 같은 가정을 반영한다. 공분산분석은 처치집단과 통제집단의 참여자들이 원래 하나의 동일한 모집단에 속하며, 집단 할당이 사전점수에 영향을 미치지 않는다고 가정한다. 즉, 모든 참여자가 원래 동일한 모집단에 속해 있었고, 어느 집단에 할당되었는가가 사전점수에 아무런 영향을 미치지 않았다면, 사전점수에서의 두 집단 간 표본 평균에 차이가 관찰되었을 때 이것이 모집단 평균 차이로 인해 발생한 것은 아니라고 할 수 있다. 따라서, 처치 효과가 없다면, 두 집단에 속한 개인들의 점수는 동일한 하나의 모집단 평균으로 회귀하게 되고, 이로 인해 사후 시점에는 두 집단의 평균 값이 서로 더 유사해지면서 차이가 줄어들게 된다(Castro-Schilo & Grimm, 2018; Van Breukelen, 2013).

이와 같은 맥락에서, Maris(1998)는 회귀불연속 설계를 사용할 경우 공분산분석이 처치효과를 정확히 추정함을 이론적으로 설명한 바 있다. 회귀불연속 설계는 집단 할당 이전에 사전점수를 측정하며, 관찰된 사전점수에 기반하여 집단을 할당한다. 따라서, 회귀불연속 설계는 두 집단에 속한 모든 참여자가 원래 동일한 하나의 모집단에 속하고(사전점수를 측정하기 전에는 참여자들을 두 집단으로 구분할 수 없으므로), 집단 할당이 사전점수에 영향을 미치지 않는다(집단 할당은 사전점수 측정 이후에만 가능하므로)는 공분산분석의 가정을 충족하고, 이로 인해 공분산분석이 정확하게 처치효과를 추정한다고 할 수 있다.

이와 달리, 차이점수 분석은 처치효과가 존재하지 않을 때 사전점수의 집단 간 평균 차이가 사후점수에도 그대로 유지될 것으로 예측한다(Castro-Schilo & Grimm, 2018). 이는 처치집단과 통제집단이 이질적인 모집단에 속한다는 가정을 내포하는데(Van Breukelen, 2013), 이러한 가정은 회귀불연속 설계에 부합하지 않는다. 따라서, 회귀불연속 설계에서 차이점수 분석은 처치효과를 정확하게 추정하지 못한다.

반면, 이미 존재하는 서로 다른 집단을 비교(예: 결혼의 효과를 검증하기 위해 기혼 집단과 미혼 집단을 비교)하는 비동질적 집단 설계(nonequivalent groups design)에서는 공분산분석의 가정을 만족시키기가 어렵다. 두 집단이 동일한 하나의 모집단에서 나온 것이 아

2) 공분산분석 및 차이점수 분석은 영가설이 참일 때(즉, 처치효과가 존재하지 않을 때)에는 일관되게 서로 다른 패턴의 결과를 예측하지만, 영가설이 참이 아닐 때에는 처치효과의 크기와 방향에 따라 비일관된 다양한 패턴의 결과를 예측한다. 따라서, 선행 연구(Castro-Schilo & Grimm, 2018; Van Breukelen, 2013)에서는 처치효과가 존재하지 않을 때를 기준으로 특정 실험 설계가 공분산분석 혹은 차이점수 분석의 가정에 부합하는지 살펴보았으며, 본 논문에서도 동일하게 처치효과가 존재하지 않을 때를 기준으로 두 분석 방법이 내포하고 있는 가정에 대해 살펴보았다.

니라 서로 다른 특징을 갖는 이질적인 모집단에서 나왔다면, 처치효과가 존재하지 않을 때 사후시점에 두 집단의 평균 점수가 동일한 모집단 평균으로 회귀할 이유가 전혀 없으며, 공분산분석이 예측하는 결과 패턴을 따르지 않게 된다. 따라서, 비동질적 집단 설계에서 공분산분석을 사용하여 처치효과를 추정하면 편향된 결과를 얻게 된다(Castro-Schilo & Grimm, 2018; Van Breukelen, 2013).

만약, 비동질적 집단 설계에서, 처치효과가 존재하지 않을 때 사후점수에서의 두 집단 간 평균 차이가 사전점수에서의 집단 간 평균 차이와 동일하게 유지된다면, 이는 차이점수 분석이 예측하는 결과 패턴에 부합하며, 이 경우 차이점수 분석을 사용하면 처치효과를 정확하게 추정할 수 있다(Castro-Schilo & Grimm, 2018). 그러나, 이 조건 또한 성립하지 않는다면, 두 분석 방법 모두 편향된 결과를 산출하게 된다.

정리하면, 비무선할당 설계에서는 Lord의 역설이 발생할 수 있으며, 비무선할당 중 구체적으로 어떤 집단 할당 방법을 사용했는가에 따라 각 분석 방법이 산출하는 처치효과 추정 결과의 정확성이 달라진다. 비무선할당 중 회귀불연속 설계를 사용한 경우 공분산분석의 결과는 정확하지만 차이점수 분석의 결과는 그렇지 않다. 반면, 비동질적 집단 설계를 사용한 경우에는 상황에 따라 차이점수 분석이 정확할 수도 있지만, 두 분석 방법 모두 부정확한 결과를 산출할 수도 있다.

선행 시뮬레이션 연구 개관

차이점수 분석과 공분산분석의 수행을 비교한 시뮬레이션 연구들은 Lord의 역설이 발생할 수 있는 다양한 조건 하에서 두 분석 방법의 수행을 비교하고자 하였다.

Wright(2006)는 비무선할당 설계에서 두 방법의 수행을 비교하기 위해, 사전점수에 기반하여 집단 할당한 경우(조건 1)와, 사전점수에 기반한 집단 할당은 아니지만 사전점수가 집단과 상관관계를 보이는 경우 즉, 집단 할당이 사전점수가 아닌 사전 진점수(true score)에 따라 확률적으로 이루어지는 상황(조건 2)을 고려하였다. Wright(2006)는 집단 할당 방법과 함께 평균 처치효과 크기와 점수의 신뢰도를 달리하여 다양한 조건을 생성하고, 각 조건에서 생성한 자료를 두 방법으로 분석하여 처치효과 추정 편향에서의 차이를 살펴보았다. 그 결과, 집단 할당 조건 1에서는 공분산분석이 편향되지 않은 추정치를 산출한 반면, 차이점수 분석은 신뢰도가 완벽한 경우에만 편향되지 않은 추정치를 산출하였고 신뢰도가 낮아질수록 더 크게 편향된 결과를 산출하였다. 이와 달리, 집단 할당 조건 2에서는 차이점수 분석이 편향되지 않은 추정치를 산출한 반면, 공분산분석은 편향된 추정치를 산출하였고 신뢰도가 낮아질수록 더 큰 편향을 나타냈다.

Petscher와 Schatschneider(2011)는 이와 달리 무선할당 설계만을 고려하여 시뮬레이션 연구를 수행하였다. 무선할당의 경우 두 방법 모두 편향되지 않은 결과를 산출하기 때문에, 편향은 고려하지 않고 제1종 오류율과 검정력에 초점을 두고 두 방법을 비교하였다. 보다 구체적으로, 이들은 자료의 안정성이 달라짐에 따라 두 분석 방법 결과에 어떠한 영향을 미치는지 살펴보았는데, 안정성의 정도를 조작하기 위해 사전-사후점수 간 상관 및 사전점수-변화 간 상관[3]을 체계적으로 변화시켰

다. 그리고, 추가적으로 자료가 편포된 정도를 조작하여, 자료가 정규분포를 따르지 않을 경우 검정력이 어떻게 달라지는지도 함께 살펴보았다. 그 결과, 제1종 오류율은 모든 조건에서 .05 수준을 유지하였고, 분석 방법 간 차이도 없었다. 검정력의 경우, 자료 분포에 따른 차이는 거의 없었고, 사전-사후점수 간 상관이 낮으면 공분산분석이 차이점수 분석에 비해 다소 높은 검정력을 보였으나, 상관이 높아질수록 두 방법 간 차이가 줄어들었다. 사전점수-변화 간 상관이 정적인 경우(확산 성장)에는 공분산분석이 차이점수 분석에 비해 더 높은 검정력을 보였으나, 사전점수-변화 간 상관이 0인 경우(평행 성장)와 부적인 경우(숙달 학습)에는 두 방법 간 검정력 차이가 거의 없었다.

Kisbu-Sakarya와 동료들(2013)은 점수의 안정성, 신뢰도, 사전점수의 집단차, 표본 크기, 그리고 처치효과 크기에 따라 두 방법[4]이 수행에 어떠한 차이를 보이는지 살펴보았다. 이때 처치효과에 대한 편향은 살펴보지 않았고, 제1종 오류율과 검정력에만 초점을 두었다. 이들은 공분산분석 모형을 참모형으로 하고, 사전점수가 사후점수를 예측할 때의 기울기를 조작함으로써 점수의 안정성을 변화시켰다. 무선할당, 사전점수에 의한 할당 등과 같은 집단 할당 메커니즘을 조작하지는 않았으며, 사전점수에 나타나는 집단 간 평균 차이의 크기만을 변화시켰다. 시뮬레이션 결과, 제1종 오류율에는 분석 방법 간 차이가 나타나지 않았다. 검정력의 경우, 점수가 완벽하게 안정적일 때는 두 방법 간 차이가 없었지만, 안정성이 낮은 조건에서는 공분산분석이 차이점수 분석에 비해 더 높은 검정력을 나타냈다. 안정성은 신뢰도와 상호작용을 나타냈는데, 공분산분석의 경우 신뢰도가 낮을 때는 안정성이 높아질수록 검정력이 낮아진 반면, 신뢰도가 높을 때는 안정성에 따른 일관된 패턴을 보이지 않았다. 이와 달리, 차이점수 분석은 신뢰도가 낮을 때 안정성에 따른 일관된 패턴을 보이지 않은 반면, 신뢰도가 높을 때는 안정성이 높아질수록 검정력이 높아지는 것으로 나타났다.

사전점수의 집단차에 따른 결과를 살펴보면, 신뢰도가 1인 경우에는 두 방법 간 차이가 나타나지 않았지만, 신뢰도가 1보다 낮은

3) 이들은 사전점수-변화 간 상관이 정적인 경우, 0인 경우, 부적인 경우를 고려하였다. 첫째, 사전점수-변화 간 상관이 정적인 경우는 사전점수가 높을수록 점수가 더 많이 상승하는 패턴을 나타낸다. 이 경우, 사전시점에 비해 사후시점에 점수가 더 많이 퍼져서 나타나므로, 사전점수에 비해 사후점수의 분산이 증가하는 경우에 해당되며, 이를 확산 성장(fan-spread growth)이라고 명명하였다. 둘째, 사전점수-변화 간 상관이 0인 경우는 모든 사람이 동일한 정도로 변화하는 패턴을 나타내고, 이를 평행 성장(parallel growth)이라 명명하였다. 셋째, 사전점수-변화 간 상관이 부적인 경우는 사전점수가 높을수록 점수가 더 적게 상승하는 패턴을 나타낸다. 이 경우, 사전시점에 비해 사후시점에 점수가 더 적게 퍼져서 나타나므로, 사전점수에 비해 사후점수의 분산이 감소하며, 이를 숙달 학습(mastery learning)이라고 명명하였다.

4) 이들은 차이점수 분석, 공분산분석, 그리고 잔차 차이점수(residual change score) 분석의 세 가지 방법을 비교하였다. 잔차 차이점수 분석은 사전점수로 사후점수를 예측했을 때의 잔차를 구한 후, 이 잔차에서의 집단차를 추정하는 방법이다. 잔차 차이점수 분석은 공분산분석과 수행에 큰 차이를 보이지 않았으며, 공분산분석에 비해 널리 사용되지 않기 때문에 본 연구에서는 고려하지 않았다.

경우에는 사전점수 집단차가 존재하는 조건에서 두 방법 간 차이가 관찰되었다. 공분산분석의 경우, 사전점수에 집단차가 존재할 때 처치효과가 없는 조건에서 제1종 오류율이 높아졌고, 처치효과가 존재하는 조건에서는 사전점수 집단차의 방향이 처치효과와 동일한 방향일 경우에는 검정력이 높아졌으나 집단차의 방향과 처치효과 방향이 반대일 경우에는 검정력이 낮아졌다. 차이점수 분석의 경우, 사전점수 집단차 크기나 방향에 따라 검정력이 달라지지 않았다.

Jennings와 Cribbie(2016)는 이전 연구들에서 고려되었던 신뢰도, 집단 할당 방법, 처치효과 크기, 표본 크기와 같은 다양한 요인들을 포괄적으로 고려하였고, 이에 더해 천장효과와 바닥효과의 유무에 따른 결과 차이도 살펴보았다. 또한, 편향, 검정력, 제1종 오류율과 같은 다양한 결과 지표를 모두 고려하였다. 시뮬레이션 결과, 전반적으로 천장효과와 바닥효과는 처치효과 추정에 편향을 가져오는 것으로 나타났다. 집단 할당 방법별로 살펴보면, 무선할당 조건에서는 분석 방법[5] 간 수행 차이가 크지 않았으나, 능력치에 따른 집단 할당에서는 공분산분석이, 능력치가 다른 이질적 집단 간 비교에서는 차이점수 분석이 편향되지 않은 결과를 산출하였다. 제1종 오류율과 검정력은 편향에 의해 영향을 받는 것으로 나타났다.

5) 이들도 Kisbu-Sakarya 등(2013)과 마찬가지로 차이점수 분석, 공분산분석, 그리고 잔차 차이점수(residual change score) 분석의 세 가지 방법을 비교했고, 잔차 차이점수 분석은 공분산분석과 수행에 큰 차이를 보이지 않는 것으로 나타났다.

선행 연구의 한계와 본 연구의 목적

이러한 선행 시뮬레이션 연구들을 검토한 결과, 다음과 같은 한계점을 찾을 수 있었다. 첫째, 앞서 언급한 선행 시뮬레이션 연구들의 특징을 정리한 표 1을 살펴보면, 연구마다 조작한 요인 및 각 요인의 수준을 조작한 방법에 차이가 있음을 알 수 있다. 특히, 연구에 따라 집단 할당 방법에 큰 차이를 보였는데, 무선할당만 고려했거나(Petscher & Schatschneider, 2011), 비무선할당만 고려한 경우(Wright, 2006)가 있었고, 무선할당과 비무선할당을 모두 고려했다고 하더라도 이를 시뮬레이션에서 구현한 방식에 차이가 있었다.

예를 들어, Wright(2006)는 비무선할당 방식으로 사전점수 기반 할당과 사전 진점수 기반 확률적 할당의 두 가지를 고려하였다. 사전점수 기반 할당은, 예컨대 자존감 향상 프로그램의 효과를 검증하는 연구에서, 자존감 사전점수가 기준 이하인 참여자는 프로그램이 보다 유용할 것으로 판단하여 처치집단에 배정하고, 기준 이상인 참여자는 이 프로그램이 상대적으로 덜 유용할 것으로 판단하여 통제집단에 배정하는 방식이다. 반면, 사전 진점수 기반 확률적 할당은, 예를 들어 우울치료 프로그램의 효과를 검증하는 연구에서 참여자들이 스스로 집단을 선택하도록 했을 때, 사전시점의 실제 우울 수준이 높을수록 처치집단을 선택할 확률이 높고, 낮을수록 통제집단을 선택할 확률이 높아지는 경우를 의미한다.

이와 달리, Kisbu-Sakarya 등(2013)은 시뮬레이션에서 비무선할당 중 비동질적 집단 할당, 즉 원래 평균이 서로 다른 이질적 모집단에서 처치집단과 통제집단의 참여자를 모집하

표 1. 선행 시뮬레이션 연구 비교

선행 연구	Wright(2006)	Petscher & Schatschneider (2011)	Kisbu-Sakarya et al. (2013)	Jennings & Cribbie (2016)
조작요인				
신뢰도	O (.2, .5, .8, 1)	X	O (.5, .8, 1)	O (.6, .9)
안정성	X	O 사전-사후점수 간 상관 (.2,.4,.6,.8) 사전점수-변화 간 상관 (-.3, 0, .3)	O 사전 점수 기울기 (.3, .5, .7, 1)	X 점수의 신뢰도에 따라 사전-사후점수 간 상관이 결정된다고 보았음
집단 할당 방법	O (사전점수 기반 할당, 사전 진점수 기반 확률적 할당)	X 무선할당만 고려	X	O (무선할당, 능력치 기반 할당, 능력치가 다른 기존 집단 비교)
사전 점수 집단차	X 집단 할당 방법에 따라 사전점수의 집단차가 발생할 수 있으나 직접 조작하지는 않음	X	O 두 집단 간 평균 차이(0, -.14, .14, -.59, .59)	O 능력치가 다른 기존 집단을 비교하는 집단 할당 방법에 대해서만 집단과 사전점수 간 상관 조작(.2, .4)
처치 효과 크기	X 집단 크기 50으로 고정	O 집단 크기 (20, 30, 50, 200, 500)	O 집단 크기 (50, 100, 200, 500)	O 집단 크기 (20, 50, 100)
기타 요인	X	O 왜도(0, -.5, .5, -1, 1)	X	O 천장 및 바닥효과(모두 없음, 천장효과만 있음, 바닥효과만 있음)
결과 지표				
편향	O	X	X	O
검정력	X	O	O	O
제1종 오류율	X	O	O	O

주. 표에서 O는 해당 요인(혹은 결과 지표)을 고려했음을, X는 고려하지 않았음을 나타낸다.
주. 괄호 안에 제시된 숫자 혹은 조건은 시뮬레이션에서 사용한 구체적인 수준을 나타낸다.

는 상황만을 고려하였으며, 두 집단 간 평균 차이의 크기를 여러 수준으로 조작하였다.

이처럼 기존 시뮬레이션 연구들에서 사용된 사전점수 기반 할당, 사전 진점수 기반 확률적 할당, 비동질적 집단 할당과 같은 비무선할당 방식은 결과적으로 사전점수에서 집단 간 평균 차이를 유발하게 된다. 예컨대, 자존감 향상 프로그램의 예처럼 사전점수 기반으로 집단을 나누는 경우, 사전점수가 낮은 참여자는 처치집단에, 높은 참여자는 통제집단에 배정되므로, 처치집단의 사전점수 평균은 통제집단보다 낮을 수밖에 없다. 마찬가지로, 우울 치료 프로그램의 예처럼 사전 진점수가 높을수록 처치집단을 선택할 확률이 높다면, 처치집단의 사전점수 평균이 통제집단보다 높아지게 된다. 비동질적 집단 설계의 경우에는 처음부터 두 집단이 평균 차이를 갖도록 데이터를 생성하기 때문에, 사전점수에 명확한 평균 차이가 존재하게 된다.

그러나, 사전점수에서 관찰되는 이러한 집단 간 평균 차이는 설령 그 크기가 완전히 동일할지라도 그것이 어떠한 집단 할당 메커니즘에 의해 발생했는가(예를 들어, 사전점수에 기반하여 집단을 할당한 결과로 발생한 차이인지, 원래 평균이 서로 다른 이질적 집단이기 때문에 나타난 차이인지)에 따라 어느 분석 방법의 결과가 정확한지가 달라지게 된다. 따라서, 다양한 집단 할당 방법을 보다 체계적으로 구현하고 그 영향을 비교할 필요가 있다.

둘째, 안정성을 체계적으로 조작한 연구가 많지 않다. 안정성은 앞서 언급하였듯이 사전-사후 진점수 간 상관, 사전 및 사후점수의 신뢰도, 사전 및 사후점수의 집단차, 그리고 사전-사후점수 분산 비율에 의해 달라진다. 사전 및 사후점수에서의 집단 차이는 집단 할당 메커니즘에 따라 영향을 받기 때문에, 집단 할당 메커니즘이 동일하다면 사전-사후 진점수 간 상관, 사전 및 사후점수의 신뢰도, 그리고 사전 및 사후 점수 분산의 비율에 따라 안정성이 달라진다고 할 수 있다.

그러나, Petscher와 Schatschneider(2011) 만이 사전, 사후점수 분산을 다른 요인과 구분하여 따로 조작하였고, 다른 연구들은 대체로 분산에 대한 명확한 언급을 하지 않았다. 또한, 사전-사후 진점수 간 상관 그리고 사전 및 사후점수의 신뢰도를 구분하지 않고, 사전-사후점수 상관을 조작한 경우가 많았다[6]. 예를 들어, Jennings와 Cribbie(2016)는 사전-사후점수의 상관이 신뢰도에 의해 완전히 결정되도록 시뮬레이션을 구현하였고, 사전-사후점수 간 상관과 신뢰도를 따로 고려한 선행 연구들도 모두 진점수가 아닌 관찰점수 상관에 기반하여 그 수준을 조작하였다(Kisbu-Sakarya et al., 2013; Petscher & Schatschneider, 2011).

셋째, 선행 연구에 따르면, 신뢰도, 사전-사후점수 간 상관, 집단 할당 방법에 따라 두 분석 방법 간 차이가 발생한다는 결과가 반복적으로 확인되었음에도 불구하고, 이 요인들을 통합적으로 고려하여 시뮬레이션을 수행한 연구는 없었다. 이러한 요인들이 복합적인 상호작용을 나타낼 가능성이 존재하지만,

6) 앞서 각주 1에서 설명했듯이 사전-사후점수 상관은 사전-사후 진점수 간 상관과 사전 및 사후점수 신뢰도에 의해 결정되며, 구체적으로 $r_{21} = \rho_{21}\sqrt{rel(X_2)rel(X_1)}$ 와 같은 관계를 갖는다. 선행 연구들에서는 진점수 상관 ρ_{12}와 신뢰도 $rel(X_2), rel(X_1)$ 수준을 따로 조작하지 않고, 관찰점수 상관 r_{21}의 수준만을 조작한 경우가 많았다.

선행 연구들은 이 요인들 중 일부만을 선택적으로 살펴보았기 때문에 이러한 상호작용을 살펴볼 수 없었다.

넷째, 연구마다 자료 생성에 사용한 참모형에 차이를 보였다. Wright(2006)는 고전검사이론 및 진점수에 대한 가정(처치가 없으면 진점수는 변화하지 않으며, 처치효과가 있는 경우 모든 사람들에게 동일한 크기로 발생한다)에 기반하여 점수를 생성한 반면, Kisbu-Sakarya 등(2013)은 공분산분석 모형을 참모형이라고 가정하고 점수를 생성하였다.

Petscher와 Schatschneider(2011) 그리고 Jennings와 Cribbie(2016)와 같이 참모형을 정확히 명세하지 않은 경우도 있었다. 참모형을 정확히 알 수 없으면 시뮬레이션에서 추정한 효과가 무엇인지 파악하기 어렵기 때문에, 연구 결과를 해석하고 일반화하는 데 제한이 따르며, 다른 연구자들이 해당 시뮬레이션 연구를 재현하거나 확장하는 것을 어렵게 만든다.

마지막으로, 연구에 따라 분석 방법의 수행을 평가하기 위해 사용한 결과 지표가 달랐다. Wright(2006)는 처치효과 추정의 편향을 살펴본 반면, Petscher와 Schatschneider(2011) 및 Kisbu-Sakarya 등(2013)은 편향은 살펴보지 않고 제1종 오류율과 검정력만을 살펴보았다. Jennings와 Cribbie(2016)는 편향, 제1종 오류율, 검정력을 모두 살펴보았는데, 이들이 지적한 바와 같이 편향이 발생하면 이것이 제1종 오류율과 검정력에도 영향을 미치기 때문에, 편향을 고려하지 않고 제1종 오류율과 검정력만을 살펴보게 되면 부정확한 결론을 도출할 우려가 있다.

이와 같은 선행 연구의 한계를 극복하기 위해 본 연구에서는 다음과 같이 시뮬레이션을 수행하고자 하였다. 첫째, 집단 할당 방식과 점수의 안정성을 보다 체계적으로 조작한다. 둘째, 신뢰도, 사전-사후점수 간 상관, 집단 할당 방식 등 다양한 요인들을 통합적으로 고려한다. 셋째, 자료 생성에 사용된 참모형을 명확히 제시하고, 이에 기반하여 결과를 평가한다. 넷째, 다양한 결과 지표를 종합적으로 고려한다. 이와 같이 포괄적이고 체계적으로 시뮬레이션을 수행함으로써, 차이점수 분석과 공분산분석을 비교하는 기존의 시뮬레이션 연구 결과들을 통합적으로 이해하고, 두 분석 방법 간의 차이 및 각 방법을 언제 사용하는 것이 적절한가에 대한 보다 심층적인 이해를 도모하고자 하였다.

연구 방법

자료 생성 모형

시뮬레이션 자료 생성을 위한 참모형은 Wright(2006)의 시뮬레이션을 참고하여 다음과 같이 설정하였다.

$$\text{사전점수: } X_{1i} = T_{1i} + m_1 E_{1i} \quad (4)$$

$$\text{사후점수: } X_{2i} = T_{2i} + dG_i + m_2 E_{2i} \quad (5)$$

식 (4)의 X_{1i}는 i번째 참여자의 사전점수를 의미하며, 고전검사이론(Lord et al., 1968)에 기반하여 진점수 T_{1i}와 측정 오차 $m_1 E_{1i}$의 합으로 구성된다고 보았다. 식 (5)의 X_{2i}는 i번째 참여자의 사후점수로, 마찬가지로 진점수 T_{2i}와 측정 오차 $m_2 E_{2i}$의 합으로 구성된다.

단, 사후점수의 경우, i번째 참여자가 처치집단($G_i = 1$)에 속할 때는 d 크기의 처치효과가 더해지고, 통제집단($G_i = 0$)에 속할 때는 처치효과가 더해지지 않도록 하였다.

사전, 사후 진점수와 측정오차는 다음과 같이 생성하였다. 먼저, T_{1i}, T_{2i}, E_{1i}, E_{2i}는 평균이 모두 0이고 분산은 모두 1인 다변량 정규분포를 따른다고 가정하였다. 이때 진점수와 측정오차 간에는 상관이 존재하지 않으며, 사전 및 사후 시점의 측정오차 간에도 상관이 존재하지 않는다고 가정하였다. 사전-사후 진점수 간 상관(r)은 조건에 따라 다르게 설정하였다. E_{1i}과 E_{2i}에 곱해진 상수 m은 측정오차의 표준편차와 동일하며, 이 값 또한 조건에 따라 다르게 설정하였다.

조작 요인

본 시뮬레이션에서는 선행 연구에서 주요하게 고려된 요인들에 기반하여, 점수의 신뢰도, 사전-사후 진점수 상관, 집단 할당 방법, 효과크기 및 표본크기의 다섯 개 요인을 다음과 같이 조작하였다.

신뢰도

집단 할당 방법이 동일할 때, 점수의 안정성은 사전-사후 진점수 간 상관, 사전-사후점수 분산 비율, 그리고 점수의 신뢰도에 따라 달라진다. 본 연구에서는 안정성에 영향을 미치는 다양한 요인들을 보다 체계적으로 조작하기 위해, 우선 두 집단에서 사전, 사후점수의 분산은 모두 동일하다고 가정하였다. 그리고, 사전-사후 진점수 상관과 신뢰도를 구분하여 각각 따로 그 수준을 조작하였다.

우선, 신뢰도는 0.6, 0.8, 1.0의 세 수준으로 설정하였다. 서로 다른 수준의 신뢰도를 나타내는 자료를 생성하기 위해, 측정오차의 표준편차(m) 값을 다음과 같이 설정하였다. 신뢰도는 관찰점수 분산 중 진점수 분산이 차지하는 비율로 정의된다(Lord et al., 1968). 또한 참모형에서 진점수와 측정오차 간 상관은 0으로 가정하였고, T_{1i}과 E_{1i}의 분산은 모두 1로 설정하기 때문에, 사전점수 신뢰도 $rel(X_1)$를 식 (6)과 같이 나타낼 수 있다.

$$rel(X_1) = \frac{Var(T_1)}{Var(T_1 + mE_1)} = \frac{Var(T_1)}{Var(T_1) + m^2 Var(E_1)} = \frac{1}{1+m^2} \quad (6)$$

이때 $Var(\cdot)$는 분산을 나타내며, 식 (6)을 측정오차 표준편차 m에 대해 풀면 식 (7)과 같다. 즉, 신뢰도 수준이 정해지면 이에 따라 측정오차 표준편차 m이 결정되고, 이렇게 구한 m을 식 (4)와 (5)에 대입하여 자료를 생성하였다.

$$m = \sqrt{\frac{1 - rel(X_1)}{rel(X_1)}} \quad (7)$$

이와 같은 방식으로 자료를 생성하면 사전점수가 사후점수와 동일한 신뢰도를 갖게 된다. 식 (8)은 사후점수의 신뢰도 $rel(X_2)$를 나타내는데, 집단 변수 G_i는 각 집단에 속한 모든 개인들은 동일한 값을 가지므로(통제집단 $G_i = 0$, 처치집단 $G_i = 1$), dG항은 상수가 되어 집단 내 점수 분산에 영향을 미치지 않는다. 진점수와 측정오차 간 상관은 0, T_{2i}과 E_{2i}의 분산은 모두 1로 설정하였으므로,

사후점수 신뢰도를 식 (8)과 같이 나타낼 수 있고, 이는 사전점수 신뢰도와 동일하다.

$$rel(X_2) = \frac{Var(T_2)}{Var(T_2 + dG + mE_2)} = \frac{Var(T_2)}{Var(T_2) + m^2 Var(E_2)} = \frac{1}{1+m^2} \quad (8)$$

즉, 본 연구에서는 사전, 사후점수의 신뢰도가 동일하며, 집단에 따라 신뢰도가 달라지지 않는다고 가정하고 자료를 생성하였다.

사전-사후 진점수 상관

사전-사후 진점수 상관은 0.6, 0.8, 1의 세 수준으로 설정하였다. 이때 사전-사후 진점수 상관은 통제집단과 처치집단에서 동일하다고 가정하였다.

사전-사후 진점수 상관이 1인 조건은 처치가 없을 때 모든 사람들의 사전 진점수가 사후 시점에 그대로 유지되는 상황에 해당되며, 사전-사후 진점수 상관이 1보다 낮은 조건은 사전-사후점수 간 변화가 개인마다 서로 다르게 발생하는 상황을 나타낸다. 예를 들어, 기분, 스트레스, 통증 등과 같은 속성들은 상황에 따라 쉽게 변하지만, 가치관, 삶의 의미, 사고 방식 등은 비교적 일관되게 유지된다. 상황에 따라 쉽게 변화하는 속성들은 아무런 처치가 없더라도 사전-사후 시점 간에 진점수가 변할 수 있고, 이 경우 사전-사후 진점수 상관이 1보다 낮아진다. 반면, 비교적 일관되게 유지되는 속성들의 경우 사전-사후 시점 간 진점수가 거의 변화하지 않고, 사전-사후 진점수 상관이 1에 가깝게 나타날 것이다.

즉, 본 시뮬레이션에서는 측정하고자 하는 속성이 안정적인 경우(처치가 없거나 처치효과가 존재하지 않으면 사전-사후 시점 간에 진점수가 변화하지 않음)와 측정하고자 하는 속성의 변동성이 상대적으로 큰 경우(처치가 없거나 처치효과가 존재하지 않더라도 사전-사후 시점 간에 진점수가 변화함)를 모두 고려하고자 하였다.

집단 할당 방법

집단 할당은 무선할당과 비무선할당 방법을 모두 고려하였고, 비무선할당 방법 중에서는 기존 연구들에서 다루어진 사전점수 기반 할당, 사전 진점수 기반 확률적 할당, 비동질적 집단 할당의 세 가지 방법을 모두 포함시켰다. 즉, 무선할당 및 세 가지 비무선할당 방법을 포함하여 총 네 가지 집단 할당 방법을 사용하였고, 이를 각각 다음과 같이 구현하였다.

첫째, 무선할당 조건에서는 식 (4)에 기반하여 개별 참여자들의 사전점수를 생성하고, 이들을 무선적으로 두 집단에 나누어 할당한 후, 식 (5)에 기반하여 사후점수를 생성하였다.

둘째, 사전점수 기반 할당 조건에서는 식 (4)에 기반하여 개별 참여자들의 사전점수를 생성한 후, 생성된 사전점수에 기반하여 표본평균보다 점수가 높은 경우 통제집단에, 낮은 경우 처치집단에 할당하였다. 이후 식 (5)에 기반하여 사후점수를 생성하였다.

셋째, 진점수 기반 확률적 할당 조건에서는 Wright(2006)과 동일하게 점수를 생성하였다. 구체적으로, 식 (4)에서 사전점수를 구성하는 진점수(T_{1i}) 값에 따라 $1/(1+\exp(-T_{1i}))$의 확률로 통제집단에 참여자를 할당하였다. 즉, 진점수가 높을수록 통제집단에 할당될 확률이 증가하도록 할당하였다. 이후 식 (4)와 (5)에 따라 사전, 사후점수를 생성하였다.

마지막으로, 비동질적 집단 설계 조건에서는 식 (4)에 기반하여 개별 참여자들의 사전 점수를 생성하고, 이들을 무선적으로 두 집단에 할당한 후, 식 (5)에 기반하여 사후점수를 생성하였다. 그 다음에, 통제집단에 속한 모든 참여자들의 사전, 사후점수에서는 0.1을 빼고, 처치집단에 속한 모든 참여자들의 사전, 사후점수에는 0.1을 더해줌으로써, 통제집단과 처치집단이 사전 시점에 평균 차이를 나타내도록 구현하였다. 이러한 방식으로 자료를 생성할 경우, 처치집단의 사전점수 평균이 통제집단에 비해 0.2만큼 높게 설정되며, 이는 두 집단이 평균이 서로 다른 이질적 집단으로부터 추출된 상황을 반영한 것이다. 또한, 이러한 자료 생성 방식을 사용하면, 통제집단(및 처치효과가 없을 때 처치집단)에서 사전점수의 집단 간 평균 차이가 사후시점에 그대로 유지된다.

효과 크기

처치효과의 크기 즉, 식 (5)에서 d의 값은 0, 0.5, 1의 세 수준으로 설정하였다. $d = 0$인 조건은 처치효과가 없는 경우이며, $d = 0.5$인 조건은 처치에 따른 평균 점수의 변화가 진점수 표준편차(=1)의 절반 크기에 해당하는 경우이고, $d = 1$인 조건은 처치에 따른 평균 점수의 변화가 진점수 표준편차 크기와 동일한 경우에 해당된다.

표본 크기

총 표본 크기는 100, 200, 400의 세 수준으로 조작하였다. 무선할당과 비동질적 집단 설계 조건에서는 두 집단의 참여자 수가 동일하도록 집단을 할당하였고(즉, 각 집단별 참여자 수 50, 100, 200), 사전점수 기반 할당과 사전 진점수에 의한 확률적 할당 조건에서는 두 집단의 크기가 생성한 자료마다 다소 달랐다.

이와 같이 총 다섯 개 요인(신뢰도, 사전-사후 진점수 상관, 집단 할당 방법, 효과 크기, 표본 크기)에 대해 그 수준을 조작하여 총 324(=3x3x4x3x3)가지 서로 다른 조건을 고려하였고, 각 조건 별로 1,000 세트씩 총 324,000 세트의 자료를 R 4.1.0(R Core Team, 2021)을 사용하여 생성하였다.

분석 방법 및 결과 지표

생성된 각 자료 세트를 차이점수 분석과 공분산분석으로 각각 분석하여 처치효과를 추정하고 유의성 검정을 실시하였다. 이때 추정의 정확성을 평가하기 위해 편향(bias)을, 검정의 정확성을 살펴보기 위해 제1종 오류율(type I error rate)과 검정력(power) 같은 결과 지표들을 사용하였고, 각 결과 지표는 다음과 같이 계산하였다.

첫째, 편향은 추정된 처치효과와 실제 처치효과 간의 평균적 차이를 나타낸다. 따라서, 편향이 작을수록 더 정확한 추정이 이루어졌음을 나타낸다. 처치효과의 참값이 0인 조건(즉, 처치효과가 존재하지 않는 조건)에서는 식 (9)와 같이 원편향(raw bias)을 계산하였고, 처치효과의 참값이 0이 아닌 조건(즉, 처치효과가 존재하는 조건)에서는 식 (10)과 같이 상대편향(relative bias)를 계산하였다. 식 (9)와 (10)에서 θ는 처치효과의 참값을 나타내고, $\hat{\theta}_k$는 k번째 자료 세트에서 얻은 처치효과 추정치를 가리킨다. 각 조건 별로 1,000개의 자료 세트를 생성하였으므로, 이 자료들로부터 구한 1,000개의 처치효과 추정치에 기반하여

원편향 및 상대편향을 계산하였다. 원편향의 경우 해석의 기준이 따로 존재하지 않지만, 상대편향의 경우 절대값이 0.1 이상(즉, 10% 이상)이면 수용 가능하지 않은 심각한 편향이라고 평가한다(Forero et al., 2009).

$$\text{원편향} = \frac{1}{1000}\sum_{k=1}^{1000}(\hat{\theta}_k - \theta) \tag{9}$$

$$\text{상대편향} = \frac{1}{1000}\sum_{k=1}^{1000}(\frac{\hat{\theta}_k - \theta}{\theta}) \tag{10}$$

둘째, 제1종 오류율은 처치효과가 존재하지 않을 때 유의한 결과를 얻을 확률을 나타내며, 식 (11)과 같이 각 조건 1,000개의 자료 세트 중 유의수준 0.05에서 영가설이 기각된 자료 세트의 비율로 계산하였다. 이 식에서 $I(p_k < 0.05)$는 k번째 자료에서 처치효과에 대한 유의확률(p_k)이 0.05보다 작으면 1, 그렇지 않으면 0의 값을 갖는다. 제1종 오류율은 그 값이 0.025~0.075일 때 수용 가능한 수준이라고 보며, 이 범위를 벗어나면 검정의 정확성이 낮은 것으로 본다(Cohen, 2016).

$$\text{제1종오류율} = \frac{\sum_{k=1}^{1000} I(p_k < 0.05)}{1000} \tag{11}$$

셋째, 검정력은 처치효과가 존재할 때 유의한 결과를 얻을 확률을 나타내며, 0.8 이상이어야 충분한 검정력을 확보했다고 본다(Cohen, 1988). 검정력은 식 (11)과 동일하게 계산하였다. 즉, 제1종 오류는 처치효과가 존재하지 않을 때 유의한 결과를 얻은 확률이고, 검정력은 처치효과가 존재할 때 유의한 결과를 얻을 확률로, 처치효과 유무 조건만 다를뿐 계산은 동일하다.

자료의 분석, 그리고 편향, 검정력, 제1종 오류율을 포함하는 결과 지표의 계산은 모두 R 4.1.0(R Core Team, 2021)을 사용하여 수행하였다.

결 과

시뮬레이션 결과, 처치효과 크기가 0.5인 조건과 1인 조건은 유사한 패턴을 나타냈기 때문에, 본문에는 효과 크기가 0.5인 조건에 대한 결과만을 제시하도록 하겠다. 또한, 표본크기의 경우, 작을수록 검정력이 낮아지고 클수록 검정력이 높아지는 일반적인 패턴이 나타났는데, 이 중 표본크기 200인 조건에서 다른 요인들의 영향을 가장 뚜렷하게 살펴볼 수 있었기 때문에, 본문에는 표본크기가 200인 조건에 대한 결과를 대표로 제시하도록 하겠다. 다른 효과크기 및 표본크기 조건에 대한 결과는 모두 부록으로 제시하였다.

편향

먼저, 표 2와 3은 편향에 대한 결과를 제시하고 있다. 표 2는 처치효과 크기가 0인 조건(즉, 처치효과가 존재하지 않는 조건)에 대한 결과이고, 표 3은 처치효과 크기가 0.5인 조건(즉, 처치효과가 존재하는 조건)에 대한 결과이다. 따라서, 표 2에는 원편향을, 표 3에는 상대편향을 제시하였다.

표 2와 표 3을 비교해보면 처치효과가 없을 때와 있을 때 결과 패턴이 상당히 유사한 것을 볼 수 있고, 특히 집단 할당 방법에 따

표 2. 차이점수 분석과 공분산분석의 원편향 (효과크기=0, 표본크기=200)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.001	-0.001	0.000	0.001	0.000
	0.8	-0.001	-0.002	-0.006	0.000	-0.003	-0.005
	0.6	0.006	0.003	0.004	0.003	0.003	0.006
사전 점수 기반 할당	1	0.000	0.358	0.814	0.000	0.005	-0.008
	0.8	0.316	0.643	1.062	-0.001	-0.004	-0.010
	0.6	0.633	0.932	1.314	-0.007	0.017	0.005
사전 진점수 기반 확률적 할당	1	0.000	0.005	-0.007	0.000	-0.188	-0.374
	0.8	0.163	0.164	0.176	-0.002	-0.156	-0.287
	0.6	0.336	0.330	0.326	0.006	-0.115	-0.222
비동질적 집단 할당	1	0.000	-0.001	0.001	0.000	0.039	0.082
	0.8	-0.006	-0.001	-0.009	0.035	0.072	0.095
	0.6	0.005	-0.003	-0.004	0.083	0.100	0.124

주. 음영 처리된 값은 원편향의 절대값이 0.05 이상임을 나타내며, 절대값 0.05라는 기준은 저자들이 임의로 설정하였다.

표 3. 차이점수 분석과 공분산분석의 상대편향 (효과크기=0.5, 표본크기=200)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.008	-0.001	0.000	0.005	0.000
	0.8	-0.005	-0.008	-0.007	-0.003	-0.011	0.005
	0.6	-0.002	-0.016	-0.013	-0.005	-0.010	-0.012
사전점수 기반 할당	1	0.000	0.707	1.648	0.000	-0.004	-0.017
	0.8	0.639	1.282	2.154	0.005	0.005	0.020
	0.6	1.281	1.862	2.626	0.003	0.017	-0.027
사전 진점수 기반 확률적 할당	1	0.000	0.016	-0.008	0.000	-0.365	-0.731
	0.8	0.326	0.335	0.334	-0.002	-0.301	-0.585
	0.6	0.663	0.671	0.675	0.004	-0.213	-0.441
비동질적 집단 할당	1	0.000	-0.008	-0.021	0.000	0.071	0.140
	0.8	-0.003	-0.002	0.008	0.078	0.137	0.218
	0.6	0.001	-0.002	0.013	0.161	0.203	0.264

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

라 편향이 크게 달라진다는 것을 확인할 수 있다. 무선할당의 경우 두 방법 모두 모든 신뢰도 및 사전-사후 진점수 상관 조건에서 편향 없이 처치효과를 추정하였으나, 나머지 세 개의 비무선할당 조건에서는 두 분석 방법 간 결과 차이가 뚜렷하게 관찰되었다.

비무선할당 조건들에서의 편향을 구체적으로 살펴보면 다음과 같다. 우선, 사전점수에 기반하여 집단을 할당한 경우, 공분산분석은 모든 신뢰도 및 사전-사후 진점수 상관 조건에서 편향 없이 처치효과를 추정하였지만, 차이점수 분석은 신뢰도가 1이고 사전-사후 진점수 상관이 1인 경우를 제외하면 편향된 처치효과 추정 결과를 산출하였다. 차이점수 분석의 경우, 신뢰도가 1보다 낮아지거나, 사전-사후 진점수 상관이 1보다 낮아질수록 처치효과 추정의 편향이 증가하였다.

사전 진점수 기반 확률적 할당 조건에서는 두 분석 모두 편향이 나타났으며, 차이점수 분석과 공분산분석 모두 편향된 결과를 산출할 수 있는 것으로 나타났다. 차이점수 분석은 사전-사후 진점수 상관이 1일 때에는 신뢰도와 무관하게 처치효과를 편향 없이 추정하였으나, 사전-사후 진점수 상관이 낮아질수록 양의 방향으로 편향이 증가하였다. 반면, 공분산분석은 신뢰도가 1인 경우에는 사전-사후 진점수 상관과 관계없이 처치효과를 정확하게 추정하였으나, 신뢰도가 낮아지면 처치효과가 음의 방향으로 편향되었고, 사전-사후 진점수 상관이 높을수록 더 크게 음의 방향의 편향이 나타났다.

비동질적 집단 할당의 경우, 차이점수 분석은 모든 신뢰도 및 사전-사후 진점수 상관 조건에서 처치효과를 편향 없이 추정하였으나, 공분산분석의 경우 신뢰도가 낮을수록, 사전-사후 진점수 상관이 낮을수록 처치효과를 더 편향되게 추정하는 것으로 나타났다.

또한, 공분산분석의 경우 사전 진점수 기반 확률적 할당과 비동질적 집단 할당에서 서로 반대 방향의 편향을 나타냈다. 즉, 10% 이상의 상대편향을 보인 조건만 살펴보았을 때, 사전 진점수 기반 확률적 할당에서는 처치효과가 과소 추정된 반면, 비동질적 집단 할당에서는 처치효과가 과대 추정되었다.

이렇게 서로 다른 방향의 편향이 발생한 이유는 두 집단 할당 조건에서 자료를 생성할 때 사전점수의 집단 간 평균 차이가 반대 방향으로 설정되었기 때문이다. 사전 진점수 기반 확률적 집단 할당에서는 사전 진점수가 높을수록 통제집단에 할당될 확률이 높도록 설정했기 때문에, 처치집단이 통제집단보다 사전점수의 평균이 더 낮았다. 반면, 비동질적 집단 할당에서는 처지집단의 사전점수 평균이 통제집단보다 더 높게 설정되었고, 이러한 차이가 다음과 같은 이유로 편향의 방향에 영향을 미쳤다고 할 수 있다.

Miyazaki와 동료들(2022)은 공분산분석에서 사전점수의 신뢰도가 낮고 집단 간 진점수 평균 차이가 존재할 경우, 처치가 점수를 증가시키는 방향으로 작동한다면 처치집단의 사전점수가 더 높은 경우 처치효과가 과대추정되고, 반대로 처치집단의 사전점수가 더 낮으면 과소추정이 발생함을 보였다. 즉, 공분산분석은 측정오차가 존재할 때 처치효과의 방향과 사전점수의 집단 차 방향이 동일할 경우에는 처치효과의 크기를 과대추정하고, 두 효과의 방향이 반대될 경우에는 처치효과의 크기를 과소추정한다. 이러한 공분산분석의 특성 때문에 사전 진점수 기반 확률적 할당과 비동질적 집단 할당에서 반대 방향의

편향이 발생했다고 할 수 있다.

제1종 오류율

표 4는 처치효과 크기가 0인 조건(즉, 처치효과가 존재하지 않는 조건)에 대해 제1종 오류율을 정리한 것이다. 참고로, 제1종 오류율은 편향에 의해 영향을 받는다. 처치효과가 과대 혹은 과소 추정될수록 처치효과 추정치는 참값인 0에서 더 크게 벗어나며, 이에 따라 검정 통계치의 절대값이 증가하면서 유의한 결과를 얻을 가능성이 높아진다. 때문에, 표 2에서 추정치가 상대적으로 크게 편향된 것으로 나타났던 모든 조건에서 제1종 오류율 또한 수용 가능한 범위를 벗어나는 것으로 나타났다. 그러나 추정에 편향이 거의 발생하지 않았던 다른 모든 조건들에서는 두 분석 방법 간 실질적인 차이가 관찰되지 않았고, 두 방법 모두 수용 가능한 범위의 제1종 오류율을 보였다.

검정력

표 5는 처치효과 크기가 0.5인 조건(즉, 처치효과가 존재하는 조건)에 대해 검정력 결과를 정리한 것이다. 검정력은 제1종 오류율과 마찬가지로 추정의 편향에 의해 영향을 받는다. 즉, 과대추정이 발생하면 처치효과가 참값인 0.5보다 더 크게 추정되므로 유의한 결과를 얻을 가능성이 높아지고 검정력도 높아

표 4. 차이점수 분석과 공분산분석의 제1종 오류율 (효과크기=0, 표본크기=200)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.054	0.052	0.049	0.050	0.056	0.051
	0.8	0.054	0.046	0.058	0.053	0.053	0.064
	0.6	0.034	0.043	0.048	0.045	0.042	0.045
사전점수 기반 할당	1	0.053	0.957	1.000	0.056	0.037	0.062
	0.8	0.952	0.999	1.000	0.047	0.060	0.035
	0.6	1.000	1.000	1.000	0.049	0.049	0.055
사전 진점수 기반 확률적 할당	1	0.046	0.057	0.054	0.038	0.431	0.687
	0.8	0.453	0.225	0.148	0.060	0.221	0.400
	0.6	0.776	0.533	0.354	0.036	0.111	0.233
비동질적 집단 할당	1	0.038	0.051	0.043	0.039	0.065	0.080
	0.8	0.049	0.048	0.052	0.069	0.085	0.092
	0.6	0.066	0.059	0.036	0.118	0.116	0.110

주. 음영 처리된 값은 수용 가능한 범위(0.025~0.075)를 벗어났음을 나타낸다.

표 5. 차이점수 분석과 공분산분석의 검정력 (효과크기=0.5, 표본크기=200)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	1.000	0.868	1.000	1.000	0.928
	0.8	1.000	0.965	0.763	1.000	0.988	0.870
	0.6	0.976	0.835	0.665	0.990	0.942	0.815
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	0.890	0.530
	0.8	1.000+	1.000+	1.000+	0.932	0.691	0.481
	0.6	1.000+	1.000+	1.000+	0.752	0.576	0.414
사전 진점수 기반 확률적 할당	1	1.000	0.999	0.856	1.000	0.867-	0.143-
	0.8	1.000+	1.000+	0.950+	1.000	0.749-	0.233-
	0.6	1.000+	0.999+	0.983+	0.973	0.725-	0.341-
비동질적 집단 할당	1	1.000	0.999	0.841	1.000	1.000	0.978+
	0.8	1.000	0.961	0.784	1.000	0.994+	0.965+
	0.6	0.972	0.872	0.678	0.999+	0.991+	0.952+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.

진다. 반대로 과소추정이 발생하면 처치효과가 0.5보다 작아지면서 0에 보다 가깝게 추정되므로 유의한 결과를 얻을 가능성이 낮아지고 검정력도 떨어진다. 표 5에서 +로 표시된 칸은 표 3에서 상대편향 10% 이상의 과대추정이 발생했던 조건을, -로 표시된 칸은 표 3에서 상대편향 10% 이상의 과소추정이 발생했던 조건을 나타낸다. 과대추정이 발생한 모든 조건에서는 검정력이 0.95 이상으로 매우 높게 나타난 것을 볼 수 있고, 과소추정이 발생한 조건에서는 검정력이 상대적으로 낮게 나타난 것을 확인할 수 있다.

편향이 없었던(즉, 표 3에서 상대편향의 절대값이 10%를 넘지 않았던) 조건들에서 검정력을 살펴보면 다음과 같다. 무선할당의 경우, 차이점수 분석과 공분산분석 모두 편향되지 않은 추정치를 산출하였으나, 두 방법 모두 신뢰도가 낮아질수록, 그리고 사전-사후 진점수 상관이 낮아질수록 검정력이 낮아졌다. 또한 모든 조건에서 공분산분석의 검정력은 차이점수 분석의 검정력과 같거나 더 높은 것으로 나타났다.

다음으로, 사전점수에 기반한 집단 할당의 경우, 공분산분석만 편향되지 않은 결과를 산출하였는데, 이 경우 점수의 신뢰도가 낮을수록, 그리고 사전-사후 진점수 상관이 낮을수록 검정력이 낮아지는 것으로 나타났다.

사전 진점수에 기반한 확률적 할당의 경우,

차이점수 분석은 사전-사후 진점수 상관이 1인 경우에만 편향되지 않은 결과를 산출하였는데, 이때 신뢰도가 낮아질수록 검정력도 낮아졌다. 반면, 공분산분석은 점수의 신뢰도가 1인 경우에만 편향되지 않은 결과를 산출하였고, 이때 사전-사후 진점수 상관이 낮아질수록 검정력도 감소하였다.

마지막으로, 비동질적 집단 할당의 경우, 차이점수 분석은 모든 조건에서 추정에 편향이 없었으나, 사전-사후 진점수 상관이 낮아질수록 그리고 점수의 신뢰도가 낮아질수록 검정력이 감소하였다. 공분산분석은 사전-사후 진점수 상관이 높고 점수의 신뢰도도 높은 조건들에서 편향이 없었고, 해당 조건들에서는 검정력도 매우 높게 나타났다.

정리하면, 편향이 발생하지 않은 조건들에서 차이점수 분석과 공분산분석 모두 사전-사후 진점수 상관이 낮을수록(즉, 처치효과가 없을 때 진점수의 변동성이 클수록), 그리고 점수의 신뢰도가 낮을수록 더 낮은 검정력을 보이는 것으로 나타났다.

논 의

본 연구는 사전-사후 통제집단 설계에서 처치효과 추론을 위해 가장 널리 사용되는 두 분석 방법인 차이점수 분석과 공분산분석의 수행을 다양한 조건에서 비교해보고자 하였다. 이를 위해 시뮬레이션 연구를 수행하였고, 집단 할당 방법, 점수의 신뢰도, 진점수 상관, 효과 *크기*, 표본 *크기*와 같은 다양한 요인을 체계적으로 조작하여, 두 분석 방법이 각각의 조건에서 처치효과에 대해 얼마나 정확한 추정 및 검정 결과를 산출하는지 확인하였다.

본 연구의 주요한 결과 및 함의는 다음과 같다. 첫째, 시뮬레이션을 통해 두 분석 결과가 다르게 나타나는 조건 즉, Lord의 역설이 발생하는 조건을 경험적으로 확인할 수 있었다. 앞서 선행 연구들을 통해서도 살펴보았듯, 점수가 완벽하게 안정적인 경우(본 시뮬레이션에서는 사전-사후 진점수 상관이 1이고 신뢰도도 1인 경우)에는 Lord의 역설이 발생하지 않는다. 또한, 점수가 완벽하게 안정적이지 않더라도 사전점수에 집단 차가 존재하지 않으면 Lord의 역설은 발생하지 않는다. 표 2~5를 살펴보면, 사전-사후 진점수 상관이 1이고 신뢰도가 1인 모든 조건에서 두 방법 간에 편향, 제1종 오류율 및 검정력에 실질적인 차이가 없음을 확인할 수 있다. 또한, 사전 점수에 체계적인 집단 차가 존재하지 않는 무선할당 조건의 경우, 사전-사후 진점수 상관이나 신뢰도와 관계 없이 편향이나 제1종 오류율에서 두 방법 간 차이가 관찰되지 않았다. 다만, 검정력에는 다소 차이가 있었는데, 선행 연구들에서와 같이 공분산분석이 차이점수 분석과 동일하거나 더 높은 검정력을 보이는 것으로 나타났다(Huck & McLean, 1975; Jennings & Cribbie, 2016; Kisbu-Sakarya et al., 2013; Petscher & Schatschneider, 2011; Van Breukelen, 2013).

이를 제외한 다른 모든 조건 즉, 비무선할당 설계를 사용했고, 사전-사후 진점수 상관이 1이 아니거나 신뢰도가 1이 아닌 경우에는 차이점수 분석과 공분산분석 간 결과에 차이가 발생한다는 것을 확인할 수 있었다.

둘째, 비무선할당 설계를 살펴보면, 집단 할당 방법에 따라 두 분석 방법의 수행에 뚜렷한 차이가 관찰되었고, 점수의 안정성과 신뢰도는 집단 할당 방법에 따라 두 방법의

수행에 서로 다른 영향을 미치는 상호작용을 나타냈다.

우선, 사전점수에 기반하여 집단을 할당한 경우, 공분산분석은 사전-사후 진점수 상관이나 신뢰도에 관계없이 모든 조건에서 편향 없이 정확하게 처치효과를 추정하였다. 반면, 차이점수 분석은 신뢰도와 사전-사후 진점수 상관이 모두 1인 경우를 제외하면 심각한 편향을 보였다. 이러한 결과는 Wright(2006)의 시뮬레이션 결과와 동일하며, 회귀불연속 설계에서 공분산분석이 항상 편향 없이 처치효과를 추정할 수 있다는 Maris(1998)의 이론적 설명을 경험적으로 뒷받침하는 것이라 할 수 있다.

반대로, 비동질적 집단 설계를 사용한 경우에는 차이점수 분석이 점수의 사전-사후 진점수 상관이나 신뢰도와 관계없이 편향되지 않은 정확한 추정 결과를 산출하였고, 공분산분석은 사전-사후 진점수 상관이나 신뢰도가 낮아짐에 따라 편향된 결과를 산출하였다. 이러한 결과는 Jennings와 Cribbie(2016)의 시뮬레이션 중 능력치가 다른 기존 집단을 비교하는 조건(moderate pre-test group differences in ability 조건)에서의 결과와 일치한다.

본 시뮬레이션에서 비동질적 집단 할당을 구현한 방식을 살펴보면, 앞서 언급했듯 처치효과가 없는 경우에 사전점수의 집단 간 평균 차이가 사후시점에 그대로 유지되도록 자료를 생성하였다. 이는 차이점수 모형이 예상하는 결과 패턴과 일치한다고 할 수 있으며, 이로 인해 차이점수 분석이 정확한 결과를 산출한 것이라고 이해할 수 있다. 만약, 비동질적 집단 설계를 사용했더라도 차이점수 모형의 가정이 성립하지 않는다면(예를 들어, 사전점수에서의 집단 간 평균 차이가 사후시점에 더 크게 벌어진다면), 차이점수 분석 또한 편향된 결과를 산출할 가능성이 있다.

다음으로, 진점수에 기반한 확률적 집단 할당의 경우, 사전-사후 진점수 상관이나 신뢰도가 1이 아닌 조건에서는 두 방법 모두 편향된 결과를 산출하였다. 즉, 사전-사후 진점수가 완벽한 상관을 보이는 동시에 관찰된 사전, 사후점수가 완벽하게 신뢰로운 경우가 아니라면, 두 분석 방법 모두 처치효과를 정확하게 추론하지 못하였다.

Wright(2006)의 시뮬레이션 연구에서도 본 연구에서와 동일하게 진점수에 기반한 확률적 집단 할당 조건이 고려되었다. 이 연구에서도 ANCOVA는 신뢰도가 1인 경우에만 편향 없이 처치효과를 추정하였고, 신뢰도가 낮아질수록 더 편향되게 처치효과를 추정하였다. 반면, 차이점수 분석은 본 연구에서와는 달리 항상 편향 없이 처치효과를 추정하는 것으로 나타났다. 이러한 차이가 발생한 이유는 Wright(2006)의 시뮬레이션에서는 사전-사후 진점수 간 상관이 1인 경우만을 고려했기 때문이라고 할 수 있다. Wright는 자료를 생성할 때 사전, 사후 진점수가 동일하다고 가정했는데, 이 경우 사전-사후 진점수 상관은 항상 1이 된다. 때문에 그의 연구에서는 차이점수 분석이 항상 편향없이 처치효과를 추정했던 것이다. 이와 달리, 본 연구에서는 사전-사후 진점수 상관이 1인 조건 뿐만 아니라 0.8, 0.6인 조건을 추가적으로 살펴보았고, 사전-사후 진점수 간 상관이 낮아질수록 차이점수 분석이 더 편향된 처치효과 추정치를 산출함을 확인할 수 있었다.

셋째, 편향이 발생하지 않은 조건(즉, 무선할당 설계를 사용했거나, 사전점수 기반 할당에서 공분산분석을 사용했거나, 비동질적 집

단 설계에서 차이점수 분석을 사용한 경우)에 한해 살펴보면, 사전-사후 진점수 상관과 신뢰도는 검정력에 일관되게 영향을 미치는 것으로 나타났다. 즉, 사전-사후 진점수 상관과 신뢰도가 낮아질수록 검정력이 감소함을 확인할 수 있었다.

사전-사후 진점수 상관이 낮은 경우란, 측정하고자 하는 속성이 순간 순간 혹은 매일 매일에 따라 큰 변동성을 보이는 경우를 가리킨다. 예를 들어, 기분, 스트레스, 통증 등과 같은 속성들은 가치관, 삶의 의미, 사고방식 등과 같은 속성들에 비해 변동성이 더 크다고 할 수 있다. 본 시뮬레이션 결과에 따르면, 처치효과의 크기가 동일하더라도, 이와 같이 변동성이 큰(즉, 안정성이 낮은) 속성들에 대한 처치효과를 검증할 때에는 검정력을 확보하는 것이 더 어렵기 때문에, 정확한 검증 결과를 얻기 위해서는 상대적으로 더 큰 표본을 사용할 필요가 있음을 알 수 있다. 마찬가지로, 신뢰도가 더 낮은 측정 도구를 사용하여 처지효과를 검증하기 위해서는 상대적으로 더 큰 표본 크기가 필요하다.

종합적으로, 본 연구 결과는 사전-사후 통제집단 설계에 기반하여 처치효과를 검증하고자 할 때, 분석 방법의 선택이 단순히 선호의 문제가 아님을 분명하게 보여준다. 처치집단과 통제집단에 참여자를 어떻게 할당했는가, 측정하고자 하는 속성이 얼마나 안정적인가, 그리고 측정 도구가 얼마나 신뢰로운가에 따라 차이점수 분석과 공분산분석이 서로 다른 결과를 산출할 수 있으며, 이 중 어느 방법이 산출한 결과가 정확한지 또한 달라지기 때문이다.

따라서, 본 연구 결과에 기반한여, 사전-사후 통제집단 설계 연구에서 차이점수 분석 혹은 공분산분석을 사용하여 처치효과를 검증하고자 하는 연구자들에게 다음과 같은 분석 지침을 제공할 수 있다. 첫째, 무선할당 연구의 경우, 두 분석 방법 모두 처치효과를 편향 없이 정확하게 추정하지만 검정력의 측면에서 이점이 있는 공분산분석을 사용할 것이 권장된다. 둘째, 회귀불연속 설계 연구에서는 차이점수 분석과 공분산분석 중에서 공분산분석을 사용할 것이 권장된다. 셋째, 사전 진점수에 기반한 확률적 할당에 해당되는 경우(예를 들어, 참여자들이 스스로 필요에 의해 집단을 선택하는 경우), 측정하고자 하는 속성이 매우 안정적이라면 차이점수 분석을 사용할 수 있고, 측정 도구가 매우 신뢰롭다면 공분산분석을 사용할 수 있다. 측정하고자 하는 속성의 변동성이 크거나 측정 도구의 신뢰도가 매우 높지 않다면, 두 분석 방법 모두 권장되지 않는다. 이 경우에는 성향점수 분석이나 잠재변수에 기반한 분석과 같은 다른 분석 방법들을 고려해볼 필요가 있다(Lee & Suk, 2024). 넷째, 비동질적 집단 설계 연구에서는 차이점수 분석이 더 권장되나, 경우에 따라 부정확한 결과를 도출할 수 있음을 주지할 필요가 있다. 다섯째, 두 분석 방법 중 어느 방법을 사용하건, 개입하고자 하는 속성의 변동성이 클수록, 그리고 측정 도구의 신뢰도가 낮을수록 정확한 처치효과 검증을 위해서는 상대적으로 더 많은 참여자를 확보할 필요가 있다.

본 연구는 기존 시뮬레이션 연구들의 한계를 보완하여, 안정성과 집단 할당 방식을 보다 체계적으로 조작하고, 신뢰도, 사전-사후 진점수 상관 및 집단 할당 방식의 상호작용을 포괄적으로 고려함으로써, 연구자들이 보다 정교하게 두 분석 방법의 차이를 이해할

수 있도록 방법론적 통찰을 제공하고, 적절한 분석 전략을 수립하기 위한 실질적 가이드라인을 제공했다는 의의를 갖는다.

그럼에도 불구하고, 본 연구에는 다음과 같은 제한점이 있다. 첫째, 본 연구의 시뮬레이션에서 사용한 참모형은 자료의 정규성과 등분산성을 가정하고 있으며, 점수의 신뢰도가 시점과 집단에 따라 달라지지 않고, 사전-사후 진점수 간 상관 또한 집단에 따라 달라지지 않는다는 가정에 기반하고 있다. 그러나, 이러한 가정들이 성립하지 않을 경우, 두 분석 방법이 산출하는 결과가 달라질 수 있다. 실제로, 자료의 왜도를 조작했거나, 천장/바닥 효과 등을 고려한 선행 연구들에서 정규성이 깨질 때 처치효과 추정에 편향이 발생한다는 결과가 보고된 바 있다(Jenning & Cribbie, 2016; Petscher & Schatschneider, 2011). 따라서, 이러한 가정이 성립하지 않는 보다 현실적인 조건들을 추가적으로 함께 살펴보고, 본 연구의 결과가 이러한 상황에도 일반화될 수 있는지 확인할 필요가 있다.

둘째, 본 연구의 시뮬레이션에서는 모든 개인에게 처치효과가 동일한 크기(d)로 나타난다는 가정 하에 자료를 생성하였다. 그러나, 처치에 대한 반응성은 개인에 따라 다르게 나타날 수 있으며(Gollwitzer et al., 2014), 개인차의 정도에 따라 두 방법이 산출하는 결과의 정확성 또한 달라질 가능성이 있다. 따라서, 향후 연구에서는 처치효과의 개인차를 무선효과로 모형에 포함하여 시뮬레이션을 수행하고, 이러한 무선효과의 특성이 두 방법의 수행에 어떠한 영향을 미치는지 추가적으로 살펴볼 필요가 있다.

셋째, 본 연구는 특정 집단 할당 메커니즘(무선 할당, 사전점수에 기반한 할당, 사전 진점수 기반 확률적 할당, 비동질적 집단 할당)이 사용된 조건 하에서 두 분석 방법의 수행을 평가하였다. 그러나 실제 연구에서는 집단 할당 메커니즘이 불분명하거나, 본 연구에서 고려하지 않은 방식으로 집단이 할당될 수도 있다. 본 연구의 결과는 이러한 상황에 모두 일반화될 수 없다는 한계를 갖는다.

넷째, 본 연구는 비무선할당 조건들에서 집단 간 사전점수 차이는 고려하였으나, 연령, 성별, 사회경제적 지위 등 사전점수 외의 공변인들에 대해서는 집단 간 차이가 없다고 가정하였다. 그러나 실제 상황에서는 이러한 공변인들에서 집단 차이가 존재할 수 있으며, 이는 분석 결과에 영향을 미칠 수 있다. Wright(2020)는 비무선할당 설계에서 처치집단에 속할 확률을 결정하는 잠재변수인 성향(propensity)이 다양한 공변인의 영향을 받는다고 전제하였다. 그는 성향이 사전점수와 어떠한 관계를 갖는가에 따라 다양한 조건에서 데이터를 생성한 후, 차이점수 분석, 공분산 분석, 성향점수 분석(propensity score analysis)을 적용하여 각각의 방법이 처치효과를 얼마나 정확하게 추정하는지를 비교하였다. 그 결과, 자료 생성 조건에 따라 분석 방법 별 처치효과 추정의 정확도가 따라 현저히 달라질 수 있음이 확인되었다. 이처럼 공변인이 집단 및 사전점수와 갖는 관련성은 처치효과 추정에 중요한 영향을 미칠 수 있음에도, 본 연구에서는 공변인의 영향을 별도로 고려하지 않았다. 본 연구에서 수행한 시뮬레이션은 비무선할당 조건에서 처치집단과 통제집단 간의 차이가 오직 사전점수에만 존재한다는 전제를 기반으로 하였기 때문에, 공변인에서 집단 차이가 존재하는 경우로 결과를 일반화하기 어렵다는 중요한 한계를 갖는다. 따라서, 향후

연구에서는 비무선할당 설계에서 집단 간 공변인 차이가 존재할 경우, 각 분석 방법의 수행에 어떠한 영향을 미치는지를 체계적으로 검토할 필요가 있다.

마지막으로, 본 연구에서는 제1종 오류율과 검정력을 해석할 때, 편향에 의해 제1종 오류율이나 검정력이 부정확하게 산출되었을 가능성을 언급하였다. 그러나, 제1종 오류율이나 검정력은 편향 뿐만 아니라 표준오차(standard error) 추정의 정확성에 의해서도 영향을 받는다. 따라서, 이후 연구에서는 추정치 편향과 표준오차 추정의 정확성을 분리하여 살펴볼 수 있도록 시뮬레이션을 설계하고, 어떤 조건에서 추정치 편향에 의해, 어떤 조건에서는 표준오차 추정의 부정확성으로 인해 제1종 오류율이나 검정력이 부정확하게 산출되는지 살펴볼 필요가 있다.

참고문헌

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological methodology*, *20*, 93-114. https://doi.org/10.2307/271083

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Houghton, Mifflin and Company.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203771587

Cohen, J. (2016). Things I have learned (so far). In A. E. Kazdin (Ed.) *Methodological issues and strategies in clinical research* (4th ed., pp 265-276). American Psychological Association. https://doi.org/10.1037/14805-017

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

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural equation modeling*, *16*(4), 625-641. https://doi.org/10.1080/10705510903203573

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

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). Erlbaum.

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*(4), 511-518. https://doi.org/10.1037/h0076767

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

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

Lee, Y. S., & Suk, H. W. (2024). How to analyze group difference in change: Comparing difference score model and analysis of covariance model. *Korean Journal of Psychology: General. 43*(3), 231-260. https://doi.org/10.22257/kjp.2024.9.43.3.231

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 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*(3), 309-327. https://doi.org/10.1037/1082-989X.3.3.309

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

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, *25*(1), 3-28. https://doi.org/10.1177/0193841X0102500101

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

R Core Team. (2021). *R: A language and environment for statistical computing*(Version 4.1.0). R Foundation for Statistical Computing. https://www.r-project.org/

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

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

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

Wright, D. B. (2020). Gain scores, ANCOVA, and propensity matching procedures for evaluating treatments in education. *Open Education Studies*, *2*(1), 45-65.
https://dx.doi.org/10.1515/edu-2020-0107

1차원고접수 : 2025. 04. 12
2차원고접수 : 2025. 06. 09
최종게재결정 : 2025. 06. 24

한국심리학회지: 일반
Korean Journal of Psychology: General
2025, Vol. 44, No. 3, 163-198
http://dx.doi.org/10.22257/kjp.2025.9.44.3.163

Comparing Difference Score Analysis and ANCOVA in Pretest-Posttest Control Group Designs: A Simulation Study

Youngsoo Lee **Hye Won Suk**

Department of Psychology, Sogang University

Difference score analysis and analysis of covariance (ANCOVA) are commonly used to evaluate treatment effects in pretest-posttest control group designs. However, the two methods do not always yield equivalent results and may produce biased estimates under certain conditions. This study conducted a simulation to compare their performance across various conditions. Factors such as group assignment method, score reliability, score stability, treatment effect size, and sample size were systematically manipulated. The two methods were compared in terms of estimation bias, Type I error rate, and statistical power. Under random assignment, both methods produced unbiased estimates, with ANCOVA showing slightly greater power. Under non-random assignment conditions, the results varied depending on the assignment mechanism. ANCOVA performed better under pretest-score-based assignment, whereas difference score analysis was more accurate under nonequivalent group designs. When assignment was based on true pretest scores probabilistically, both methods produced biased estimates. Additionally, score reliability and stability influenced the results within each assignment condition. These findings highlight the importance of considering group assignment and measurement characteristics when selecting an appropriate analysis method.

Key words : difference score, ANCOVA, treatment effect, Lord's paradox, experimental design

부 록

부록 A. 시뮬레이션 결과

본문에 제시되지 않은 모든 시뮬레이션 조건에 대한 결과를 부록 A에 제시하였다. 본문에 제시된 표 2~5와 부록 A에 제시된 표 A-1부터 A-14에 제시된 결과를 비교하여 살펴보면, 동일한 효과크기 조건에서는 표본크기와 관계없이 유사한 정도의 편향이 발생한 것을 알 수 있다. 표본크기가 동일할 때에는 효과크기가 0.5에서 1로 증가할수록 상대편향이 약 1/2로 감소하는 것을 확인할 수 있었다. 상대편향을 구할 때 원편향을 효과크기로 나눈다는 점을 고려하면, 상대편향에 효과크기를 곱하여 원편향을 계산해볼 수 있다. 이와 같이 원편향을 비교해보면, 효과크기가 0.5에서 1로 증가하더라도 원편향은 유사한 정도로 발생한다는 것을 알 수 있다.

제1종 오류율의 경우, 표본크기가 증가할수록 제1종 오류율이 수용가능한 범위를 벗어나는 조건의 경우에는 제1종 오류율이 더 커지는 것으로 나타났고, 제1종 오류율이 수용가능한 범위를 벗어나지 않는 조건에서는 큰 차이를 보이지 않았다.

검정력은 표본크기가 동일할 때에는 효과크기가 더 클수록, 그리고 효과크기가 동일할 때에는 표본크기가 더 클수록 일관되게 높아지는 것으로 나타났다.

표 A-1. 차이점수 분석과 공분산분석의 원편향 (효과크기=0, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선 할당	1	0.000	0.001	0.001	0.000	0.002	0.004
	0.8	-0.001	0.007	0.020	0.000	0.004	0.025
	0.6	-0.001	-0.003	0.004	0.000	-0.003	0.003
사전 점수 기반 할당	1	0.000	0.367	0.834	0.000	0.012	0.005
	0.8	0.315	0.646	1.074	-0.002	0.008	-0.003
	0.6	0.643	0.922	1.306	0.001	-0.004	-0.015
사전 진점수 기반 확률적 할당	1	0.000	-0.001	-0.001	0.000	-0.195	-0.372
	0.8	0.167	0.164	0.173	0.003	-0.147	-0.290
	0.6	0.318	0.336	0.327	-0.010	-0.111	-0.221
비동질적 집단 할당	1	0.000	0.003	0.007	0.000	0.040	0.088
	0.8	0.006	-0.004	-0.003	0.046	0.072	0.103
	0.6	0.006	0.007	-0.011	0.089	0.107	0.120

주. 음영 처리된 값은 원편향의 절대값이 0.05 이상임을 나타내며, 절대값 0.05라는 기준은 저자들이 임의로 설정하였다.

표 A-2. 차이점수 분석과 공분산분석의 상대편향 (효과크기=0.5, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	-0.010	-0.011	0.000	-0.010	-0.006
	0.8	-0.003	-0.005	-0.005	0.000	-0.012	-0.004
	0.6	0.001	0.024	-0.017	-0.002	0.020	-0.005
사전 점수 기반 할당	1	0.000	0.725	1.653	0.000	0.013	-0.006
	0.8	0.640	1.299	2.149	0.004	0.001	0.006
	0.6	1.273	1.859	2.635	0.017	-0.005	-0.005
사전 진점수 기반 확률적 할당	1	0.000	-0.003	-0.020	0.000	-0.391	-0.751
	0.8	0.319	0.316	0.310	-0.012	-0.316	-0.590
	0.6	0.648	0.666	0.674	-0.009	-0.225	-0.422
비동질적 집단 할당	1	0.000	-0.007	0.013	0.000	0.075	0.164
	0.8	0.016	-0.012	0.012	0.097	0.134	0.219
	0.6	-0.006	-0.013	0.024	0.155	0.200	0.274

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

표 A-3. 차이점수 분석과 공분산분석의 상대편향 (효과크기=1, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.004	0.000	0.000	0.003	-0.002
	0.8	-0.002	-0.002	-0.018	-0.002	0.001	-0.012
	0.6	0.006	0.007	-0.007	0.008	-0.001	-0.002
사전 점수 기반 할당	1	0.000	0.356	0.824	0.000	0.007	-0.012
	0.8	0.318	0.645	1.065	-0.002	0.004	0.005
	0.6	0.643	0.923	1.308	-0.013	-0.001	0.010
사전 진점수 기반 확률적 할당	1	0.000	-0.004	-0.012	0.000	-0.197	-0.372
	0.8	0.170	0.167	0.163	0.004	-0.153	-0.292
	0.6	0.331	0.320	0.325	0.004	-0.123	-0.218
비동질적 집단 할당	1	0.000	0.000	-0.010	0.000	0.039	0.067
	0.8	-0.005	0.004	0.009	0.037	0.076	0.111
	0.6	-0.008	-0.005	0.014	0.069	0.100	0.138

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

표 A-4. 차이점수 분석과 공분산분석의 제1종 오류율 (효과크기=0, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.039	0.058	0.041	0.046	0.055	0.045
	0.8	0.044	0.049	0.060	0.053	0.056	0.055
	0.6	0.062	0.036	0.054	0.050	0.046	0.050
사전 점수 기반 할당	1	0.043	0.760	0.966	0.047	0.049	0.061
	0.8	0.716	0.952	0.993	0.042	0.054	0.046
	0.6	0.969	0.991	0.999	0.044	0.056	0.050
사전 진점수 기반 확률적 할당	1	0.044	0.055	0.059	0.048	0.270	0.410
	0.8	0.257	0.154	0.097	0.050	0.117	0.224
	0.6	0.426	0.316	0.197	0.052	0.081	0.136
비동질적 집단 할당	1	0.050	0.052	0.040	0.049	0.062	0.072
	0.8	0.054	0.041	0.052	0.065	0.056	0.079
	0.6	0.052	0.038	0.050	0.082	0.085	0.079

주. 음영 처리된 값은 수용 가능한 범위(0.025~0.075)를 벗어났음을 나타낸다.

표 A-5. 차이점수 분석과 공분산분석의 검정력 (효과크기=0.5, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	0.944	0.551	1.000	0.955	0.656
	0.8	0.979	0.723	0.468	0.988	0.803	0.566
	0.6	0.794	0.607	0.393	0.867	0.739	0.516
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	0.601	0.287
	0.8	1.000+	1.000+	1.000+	0.697	0.408	0.245
	0.6	1.000+	1.000+	1.000+	0.470	0.317	0.229
사전 진점수 기반 확률적 할당	1	1.000	0.934	0.539	1.000	0.550-	0.080-
	0.8	0.999+	0.931+	0.701+	0.953	0.443-	0.133-
	0.6	0.995+	0.951+	0.815+	0.794	0.443-	0.208-
비동질적 집단 할당	1	1.000	0.931	0.585	1.000	0.972	0.794+
	0.8	0.984	0.741	0.480	0.996	0.892	0.757+
	0.6	0.778	0.573	0.409	0.948+	0.865+	0.747+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.

표 A-6. 차이점수 분석과 공분산분석의 검정력 (효과크기=1, 표본크기=100)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	1.000	0.991	1.000	1.000	0.997
	0.8	1.000	1.000	0.954	1.000	1.000	0.990
	0.6	1.000	0.990	0.912	1.000	0.999	0.982
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	0.996	0.805
	0.8	1.000+	1.000+	1.000+	0.999	0.921	0.734
	0.6	1.000+	1.000+	1.000+	0.958	0.859	0.703
사전 진점수 기반 확률적 할당	1	1.000	1.000	0.987	1.000	1.000-	0.817-
	0.8	1.000+	1.000+	0.988+	1.000	0.992-	0.848-
	0.6	1.000+	1.000+	0.999+	1.000	0.976-	0.860-
비동질적 집단 할당	1	1.000	1.000	0.988	1.000	1.000	1.000
	0.8	1.000	0.999	0.970	1.000	1.000	0.998+
	0.6	1.000	0.991	0.932	1.000	0.999+	0.997+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.

표 A-7. 차이점수 분석과 공분산분석의 상대편향 (효과크기=1, 표본크기=200)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.003	-0.001	0.000	0.002	-0.004
	0.8	0.004	0.003	0.007	0.004	0.002	-0.001
	0.6	0.004	-0.005	0.006	0.002	-0.004	0.000
사전 점수 기반 할당	1	0.000	0.358	0.831	0.000	-0.001	0.014
	0.8	0.323	0.646	1.069	-0.001	0.002	-0.011
	0.6	0.641	0.921	1.322	0.005	-0.014	0.002
사전 진점수 기반 확률적 할당	1	0.000	0.003	-0.003	0.000	-0.190	-0.373
	0.8	0.168	0.169	0.162	0.001	-0.149	-0.295
	0.6	0.329	0.327	0.328	-0.004	-0.116	-0.220
비동질적 집단 할당	1	0.000	-0.003	0.010	0.000	0.038	0.088
	0.8	-0.004	0.004	0.005	0.037	0.075	0.108
	0.6	-0.001	0.005	-0.001	0.080	0.107	0.128

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

표A-8. 차이점수 분석과 공분산분석의 검정력 (효과크기=1, 표본크기=200)

집단 할당 방법	진점수상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	1.000	1.000	1.000	1.000
	0.6	1.000	1.000	0.999	1.000	1.000	1.000
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	1.000	0.984
	0.8	1.000+	1.000+	1.000+	1.000	0.999	0.957
	0.6	1.000+	1.000+	1.000+	1.000	0.988	0.931
사전 진점수 기반 확률적 할당	1	1.000	1.000	1.000	1.000	1.000-	0.989-
	0.8	1.000+	1.000+	1.000+	1.000	1.000-	0.988-
	0.6	1.000+	1.000+	1.000+	1.000	0.999-	0.992-
비동질적 집단 할당	1	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	0.999	1.000	1.000	1.000+
	0.6	1.000	1.000	0.995	1.000	1.000+	1.000+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.

표 A-9. 차이점수 분석과 공분산분석의 원편향 (효과크기=0, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.001	-0.002	0.000	0.001	-0.003
	0.8	0.001	0.004	0.006	0.001	0.003	0.004
	0.6	0.001	-0.004	-0.002	-0.001	-0.003	-0.004
사전 점수 기반 할당	1	0.000	0.356	0.826	0.000	-0.003	0.007
	0.8	0.319	0.643	1.068	0.000	-0.002	-0.008
	0.6	0.636	0.929	1.318	-0.008	-0.008	-0.003
사전 진점수 기반 확률적 할당	1	0.000	0.003	0.000	0.000	-0.187	-0.368
	0.8	0.168	0.169	0.166	0.004	-0.147	-0.296
	0.6	0.331	0.333	0.327	-0.001	-0.114	-0.221
비동질적 집단 할당	1	0.000	0.002	0.006	0.000	0.043	0.085
	0.8	0.003	0.000	-0.001	0.040	0.073	0.103
	0.6	-0.001	-0.002	-0.001	0.078	0.100	0.127

주. 음영 처리된 값은 원편향의 절대값이 0.05 이상임을 나타내며, 절대값 0.05라는 기준은 저자들이 임의로 설정하였다.

표 A-10. 차이점수 분석과 공분산분석의 상대편향 (효과크기=0.5, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.006	0.007	0.000	0.005	0.010
	0.8	0.007	0.005	-0.002	0.007	0.004	-0.008
	0.6	0.003	0.001	-0.009	0.004	0.000	-0.003
사전 점수 기반 할당	1	0.000	0.716	1.650	0.000	0.012	-0.002
	0.8	0.638	1.275	2.144	0.000	-0.003	0.002
	0.6	1.272	1.854	2.642	-0.009	0.011	0.013
사전 진점수 기반 확률적 할당	1	0.000	0.002	-0.010	0.000	-0.379	-0.748
	0.8	0.330	0.326	0.338	-0.002	-0.312	-0.586
	0.6	0.662	0.664	0.652	0.002	-0.232	-0.455
비동질적 집단 할당	1	0.000	-0.012	0.008	0.000	0.069	0.167
	0.8	-0.004	0.000	0.007	0.075	0.148	0.213
	0.6	0.010	-0.005	-0.013	0.168	0.205	0.247

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

표 A-11. 차이점수 분석과 공분산분석의 상대편향 (효과크기=1, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.000	0.004	0.000	0.000	0.003	-0.002
	0.8	-0.002	-0.002	-0.018	-0.002	0.001	-0.012
	0.6	0.006	0.007	-0.007	0.008	-0.001	-0.002
사전 점수 기반 할당	1	0.000	0.356	0.824	0.000	0.007	-0.012
	0.8	0.318	0.645	1.065	-0.002	0.004	0.005
	0.6	0.643	0.923	1.308	-0.013	-0.001	0.010
사전 진점수 기반 확률적 할당	1	0.000	-0.004	-0.012	0.000	-0.197	-0.372
	0.8	0.170	0.167	0.163	0.004	-0.153	-0.292
	0.6	0.331	0.320	0.325	0.004	-0.123	-0.218
비동질적 집단 할당	1	0.000	0.000	-0.010	0.000	0.039	0.067
	0.8	-0.005	0.004	0.009	0.037	0.076	0.111
	0.6	-0.008	-0.005	0.014	0.069	0.100	0.138

주. 음영 처리된 값은 상대편향이 10% 이상임을 나타낸다.

표 A-12. 차이점수 분석과 공분산분석의 제1종 오류율 (효과크기=0, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	0.040	0.045	0.044	0.039	0.052	0.036
	0.8	0.041	0.043	0.049	0.048	0.050	0.046
	0.6	0.046	0.063	0.053	0.046	0.051	0.056
사전 점수 기반 할당	1	0.045	1.000	1.000	0.049	0.050	0.049
	0.8	1.000	1.000	1.000	0.045	0.048	0.044
	0.6	1.000	1.000	1.000	0.053	0.043	0.045
사전 진점수 기반 확률적 할당	1	0.049	0.056	0.050	0.052	0.728	0.923
	0.8	0.763	0.427	0.257	0.034	0.365	0.697
	0.6	0.971	0.823	0.606	0.047	0.196	0.413
비동질적 집단 할당	1	0.046	0.047	0.054	0.047	0.092	0.134
	0.8	0.052	0.055	0.052	0.108	0.142	0.151
	0.6	0.045	0.044	0.044	0.163	0.162	0.179

주. 음영 처리된 값은 수용 가능한 범위(0.025~0.075)를 벗어났음을 나타낸다.

표 A-13. 차이점수 분석과 공분산분석의 검정력 (효과크기=0.5, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	1.000	0.990	1.000	1.000	0.997
	0.8	1.000	1.000	0.964	1.000	1.000	0.993
	0.6	1.000	0.994	0.916	1.000	0.998	0.985
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	0.994	0.824
	0.8	1.000+	1.000+	1.000+	1.000	0.930	0.746
	0.6	1.000+	1.000+	1.000+	0.959	0.858	0.707
사전 진점수 기반 확률적 할당	1	1.000	1.000	0.984	1.000	0.985-	0.221-
	0.8	1.000+	1.000+	1.000+	1.000	0.958-	0.398-
	0.6	1.000+	1.000+	1.000+	1.000	0.960-	0.569-
비동질적 집단 할당	1	1.000	1.000	0.992	1.000	1.000	1.000+
	0.8	1.000	1.000	0.963	1.000	1.000+	0.999+
	0.6	0.999	0.991	0.926	1.000+	1.000+	1.000+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.

표 A-14. 차이점수 분석과 공분산분석의 검정력 (효과크기=1, 표본크기=400)

집단 할당 방법	진점수 상관	차이점수 분석			공분산분석		
		신뢰도			신뢰도		
		1	0.8	0.6	1	0.8	0.6
무선할당	1	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	1.000	1.000	1.000	1.000
	0.6	1.000	1.000	1.000	1.000	1.000	1.000
사전점수 기반 할당	1	1.000	1.000+	1.000+	1.000	1.000	1.000
	0.8	1.000+	1.000+	1.000+	1.000	1.000	1.000
	0.6	1.000+	1.000+	1.000+	1.000	1.000	0.999
사전 진점수 기반 확률적 할당	1	1.000	1.000	1.000	1.000	1.000-	1.000-
	0.8	1.000+	1.000+	1.000+	1.000	1.000-	1.000-
	0.6	1.000+	1.000+	1.000+	1.000	1.000-	1.000-
비동질적 집단 할당	1	1.000	1.000	1.000	1.000	1.000	1.000
	0.8	1.000	1.000	1.000	1.000	1.000	1.000+
	0.6	1.000	1.000	1.000	1.000	1.000+	1.000+

주. 음영 처리된 값은 검정력이 0.8 미만임을 나타내며, + 표시는 해당 조건에서 상대편향 10% 이상의 과대추정이 발생했음을, - 표시는 해당 조건에서 상대편향 10% 이상의 과소추정이 발생했음을 나타낸다.