아시아교육연구 19권 2호
Asian Journal of Education
2018, Vol. 19, No. 2, pp. 477-496.
https://doi.org/10.15753/aje.2018.06.19.2.477

부트스트랩 방법으로 추정한 동등화 표준오차에 의한 검사동등화 방법의 비교

김화영(金和泳)*
김현철(金顯哲)**

논문 요약

본 연구의 목적은 공통문항 비동등집단 설계에서 모수와 비모수 부트스트랩 방법을 이용하여 동등화 표준오차를 추정하고 검사동등화 방법들을 비교하는 것에 있다. 비교의 조건으로는 검사 난이도 수준, 검사간 난이도 차이, 피험자 능력 차이, 모수모형 조건 등이 고려되었으며, 동등화 방법으로는 빈도추정 동백분위동등화, Tucker 동등화, Braun-Holland 동등화 등이 비교되었다.

연구의 결과는 이 연구의 모든 실험조건에서 빈도추정 동백분위동등화 방법이 안정적으로 나타났고 특히, 특정 모수모형 조건을 사용하는 모수 부트스트랩에 의한 빈도추정 동백분위동등화 방법이 가장 안정적으로 나타났다. Tucker 동등화 방법과 Braun-Holland 동등화 방법은 대부분 비모수 부트스트랩 방법을 사용하여 동등화 표준오차를 추정하는 것이 더 정확한 것으로 나타났으나, 검사의 난이도가 어려운 경우, 구검사가 신검사보다 어려운 경우, 피험자 능력 분포가 평균 중심으로 밀집되어 있는 경우에는 특정한 모수모형 조건의 모수 부트스트랩을 사용하는 것이 더 안정적이었다.

주요어 : 공통문항 비동등집단 설계, 부트스트랩, 로그선형모형, 검사동등화

* 제1저자: 한국행동과학연구소 책임연구원
** 교신저자: 성균관대학교 교육학과 교수

Ⅰ. 서론

회차별 검사상황에서는 동일한 구인(construct)을 측정하는 검사들이 사용된다. 그런데 이들 검사의 난이도와 변별도를 동일하게 제작하는 것은 현실적으로 불가능하므로 여러 검사형에서 나온 점수들이 상호 교환되어 사용될 수 있도록 점수들을 조정하는 검사동등화(test equating) 과정이 필요하다. 검사 점수는 선발, 평가의 목적으로 사용할 수 있고 동등화에 따른 이익이나 불이익이 발생할 수 있기 때문에 검사동등화를 통해 점수를 조정하는 것은 매우 민감한 일이다. 따라서 동등화 관계를 추정하기 위한 검사동등화 방법은 신중하게 선택되어야 한다.

검사를 동등화하는 과정에서 오차는 무선오차(random error)와 체계적 오차(systematic error)로 구분된다(Kolen & Brennan, 2004). 무선오차는 모집단으로부터 표집된 피험자들의 점수로 동등화 관계를 추정할 때 나타나게 되고 체계적 오차는 동등화 관계를 추정하는 데 있어서 추정 방법이 야기하는 편이(bias)에 의해서 발생할 수 있다. 무선오차는 동등화의 표준오차(standard error of equating; SEE)로 나타내며 동등화 결과가 신뢰할만한지 평가해주는 기준이 된다(Kolen & Brennan, 2004). 따라서 무선오차를 확인하고 감소시킬 수 있는 적절한 동등화 표준오차 추정 방법을 사용하는 것이 바람직하다.

무선오차는 부트스트랩(bootstrap)과 같은 재표집에 기반한 방법에 의해 추정될 수 있다(Efron, 1982; Efron & Tibshirani, 1993). 부트스트랩 방법은 모수 부트스트랩(parametric bootstrap)과 비모수 부트스트랩(nonparametric bootstrap)의 두 가지 방법이 있는데, 부트스트랩 방법에 의하여 표준오차를 추정한 선행연구들에서는 대부분 비모수 부트스트랩 방법이 사용되었다(Hanson, Zeng & Kolen, 1993; Kolen, 1985; Kolen & Brennan, 2004; Zeng, 1991). 두 가지 부트스트랩 방법을 사용한 연구로는 무선집단 설계(Cui & Kolen, 2008)와 공통문항 동등집단 설계(Wang & Zhang, 2009), 공통문항 비동등집단 설계(Wang, 2011)에서 빈도추정 동백분위동등화 방법의 동등화 오차를 비교한 연구들이 있는데 Kolen & Brennan(2014)은 두 가지 방법을 비교하는 더 많은 연구의 필요성을 지적하였다.

이에 본 연구에서는 Wang(2011)이 수행한 연구를 확장하여 공통문항 비동등집단 설계에서 모수와 비모수 부트스트랩 방법에 의하여 동등화 표준오차를 추정하는 것을 비교하며, 검사동등화 방법들을 추가하여 각 조건에 따른 바람직한 부트스트랩 방법과 검사동등화 방법을 탐색한다. Wang(2011)의 연구조건과 달리, 검사 난이도 수준, 검사간 난이도 차이, 피험자 능력 평균 차이, 피험자 능력 표준편차 차이 등의 조건을 추가하고, 빈도추정 동백분위동등화, Tucker 동등화, Braun-Holland 동등화 등의 동등화 방법을 다양화한다. 그리고 각 조건별로 모수와 비모수 부트스트랩 방법을 이용하여 추정된 동등화 표준오차를 통하여 각 검사동등화 방법을 비교한다. 이에 대한 구체적인 연구문제는 다음과 같다.

첫째, 동등화 표준오차의 추정을 위한 모수와 비모수 부트스트랩 방법에 따라 어떤 동등화 방법이 가장 안정적인가?

둘째, 검사 난이도 수준과 검사간 난이도 조건별로 모수와 비모수 부트스트랩 방법에 따라 어떤 동등화 방법이 가장 안정적인가?

셋째, 피험자 능력의 조건에 따라 모수와 비모수 부트스트랩 방법을 이용한 동등화 방법 중 어떤 동등화 방법이 가장 안정적인가?

Ⅱ. 이론적 배경

1. 부트스트랩의 정의와 동등화에서의 적용

모수 추정치의 표준오차를 추정하는 과정이 수학적으로 복잡하더라도 부트스트랩을 활용하면 쉽게 표준오차의 추정이 가능하다(Efron, 1982; Efron and Tibshirani, 1993). 부트스트랩은 복원추출을 허용하여 재표집한 부트스트랩 표본(bootstrap sample)으로부터 표준오차를 계산하는 방법으로, 비교적 간단하고 유연하기 때문에 다양한 자료수집설계에서 사용될 수 있다(Chernick, 1999; Davison & Hinkley, 1997; Efron and Tibshirani, 1993).

본 연구에서 사용되는 모수 부트스트랩과 비모수 부트스트랩 방법에 의한 동등화 표준오차의 추정은 다음과 같은 과정을 거친다. 우선, K개 문항으로 구성된 X형과 Y형의 두 검사가 X형은 N_X 피험자들에게, Y형은 N_Y 피험자들에게 시행되었을 때 비모수 부트스트랩에 의한 원점수 i의 동등화 표준오차는 다음의 단계로 추정된다.

첫째, N_X 피험자 표본으로부터 무작위 복원추출로 N_X 피험자 표본을 만든다.

둘째, N_Y 피험자 표본으로부터 무작위 복원추출로 N_Y 피험자 표본을 만든다.

셋째, 첫째와 둘째 단계에서 얻은 자료로부터 X형 검사의 원점수 x_i의 Y형 척도에 동등화된 점수 $eq^*_{Y1}(x_i)$가 각 동등화 방법에 의해 계산된다. 동등화점수 $eq^*_{Y1}(x_i)$에서 Y는 Y형 척도를, 1은 첫 번째 부트스트랩 표본을, i는 원점수를 나타내며, *는 이러한 결과들이 원표본이 아닌 부트스트랩 표본으로부터 산출되었음을 표현한다.

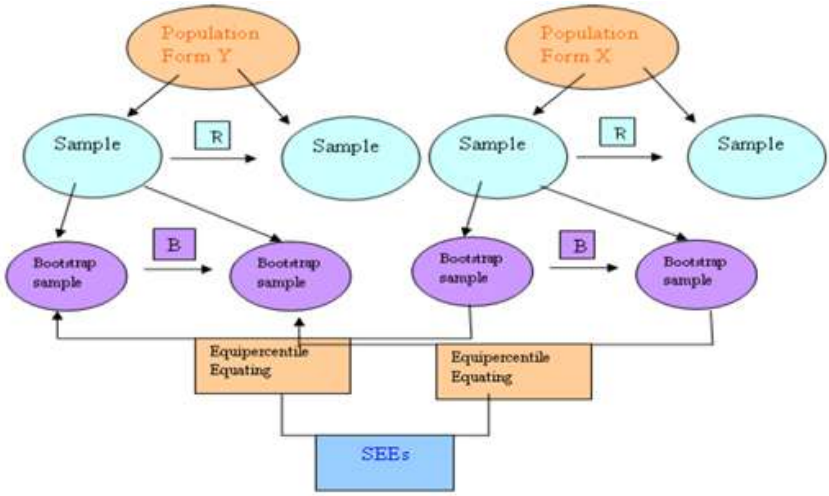
넷째, $eq^*_{Y1}(x_i)...eq^*_{YB}(x_i)$를 산출하기 위해서 첫째 단계에서 셋째 단계를 B번 반복한다. B는 부트스트랩 표본추출 수인데 본 연구에서는 1,000번 반복 수행되었다.

다섯째, 특정 동등화 점수 x_i의 동등화 표준오차를 다음과 같이 추정한다.

$$se^*[eq_Y(x_i)] = \sqrt{\frac{\sum_{b=1}^{B}[eq_{Yb}^*(x_i) - \overline{eq_Y^*}(x_i)]^2}{B-1}} \quad (1)$$

여기에서 $\overline{eq_Y^*}(x_i) = \sum_{b=1}^{B} eq_{Yb}^*(x_i)/B$ 이다.

여섯째, $se_1^*[eq_Y(x_i)]....se_R^*[eq_Y(x_i)]$을 계산하기 위해 첫째 단계에서 다섯째 단계를 R번 반복한다. R은 모집단 표본추출 수인데 본 연구에서는 300개의 표본이 모집단으로부터 추출되었다. 이는 [그림 1]의 Wang(2011)의 연구와 같은 방법으로 본 논문에서는 두 가지 검사동등화 방법이 추가되었다. R은 평가지수인 동등화 표준오차의 Bias, SE(Standard error), RMSE(Root mean squared error)의 계산에서도 사용된다.



[그림 1] 동등화 표준오차들을 추정하는 부트스트랩 절차(Wang(2011), p134)

한편, 모수 부트스트랩 방법은 모수모형에 의해 각각 X형과 Y형에 대한 경험적인 분포들을 대응시키고, 부트스트랩 표본이 이들 대응된 분포로부터 추출된다는 점을 제외하고는 비모수 부트스트랩과 동일하다.

2. 모수 부트스트랩: 로그선형모형(Log-linear model)

모수 부트스트랩 사용을 위한 모수모형 중 로그선형모형은 일변량과 이변량 빈도분포 두 가지 모두에 적합한데(Holland & Thayer, 2000), 본 연구는 비동등집단 공통문항 설계이므로 총점과 공통문항 점수의 두 가지 변수가 포함되어 이변량 로그선형모형이 사용된다. 공통문항 V 세트를 가지는 검사 X는 이변량 분포에 대한 로그선형모형의 일반적인 형태로 다음과 같이 표현된다.

$$\log(p_{ik}) = \alpha + u_{ik} + \sum_{c_x=1}^{C_X} \beta_x (x_i)^{c_x} + \sum_{c_v=1}^{C_V} \beta_v (v_k)^{c_v} + \sum_{c_{ix}=1}^{C_{IX}} \sum_{c_{iv}=1}^{C_{IV}} \beta_{xv} (x_i)^{c_{ix}} (v_k)^{c_{iv}} \quad (2)$$

여기에서 p_{ik}는 총점수 x_i와 공통점수 v_k에 대한 결합점수 확률이며, β들은 최대 우도추정 방법에 의해서 모형에서 추정될 모수들이다. 식에서 C_X와 C_V의 값은 대응분포들에서 나타나는 경험적 자료 주변분포의 적률(moment) 수를 나타낸다. 예를 들어, $C_X = 4$ 와 $C_V = 2$ 이면 총점의 4개 적률(평균, 분산, 왜도, 첨도)이 경험적 총점의 주변분포와 동일하고, 공통문항 점수의 2개 적률(평균, 분산)이 경험적 공통문항 점수의 주변분포와 동일하다는 것을 의미한다(Wang, 2011). 그리고 $C_X = 5$, $C_X = 6$은 평균, 분산, 왜도, 첨도 외에도 최대값, 최소값 등을 포함한다.

본 연구에서는 대응되는 모형이 이변량 자료세트이므로 총점과 공통문항 점수들 사이의 관계를 나타내는 벡터적률(cross product moment, $C_{IX} \leq C_X$, $C_{IV} \leq C_V$)의 숫자를 고려하는 이변량분포를 생성한다. 예를 들어, $C_{IX} = C_X = 0$ 일 때 선택된 이변량 모형은 총문항 점수와 공통문항 점수들 사이에 공분산을 계산하지 못하고, $C_{IX} = C_X = 1$ 일 때 총문항 점수와 공통문항 점수 사이에 공분산은 대응모형에서 고려되며, $C_{IX} = C_X = 2$ 일 때에는 총문항 점수와 공통문항 점수 사이에 더 높은 벡터적률들이 고려된다(Wang, 2011). 검사동등화를 할 때 이변량 로그선형모형을 사용하는 이전 연구에서는 대부분 하나의 벡터적률을 적용하였는데(Holland & Thayer, 2000; Hou 2007; Livingston, 1993) 본 연구에서는 Wang(2011)의 연구와 같이 두 가지 벡터적률들을 고려한다.

3. 검사동등화 방법

공통문항 비동등 집단설계에서 전통적 동등화 방법(남현우, 2001)을 사용하기 위해 세 가지

동등화 방법을 설정하였다. 먼저, 빈도추정 동백분위동등화 방법은 공통문항 결과를 바탕으로 합집단의 분포를 만들어 사용하게 되고, 공통 문항의 점수가 주어질 때 검사 X의 점수와 검사 Y 점수의 조건 분포가 집단 1과 집단 2에서 같다는 통계적 가정이 필요하다. 그리고 조건 분포와 결합 분포의 관계에 따라 공식들을 도출하고 합집단의 빈도 분포를 계산하여 결합 모집단의 백분위 함수식을 만든다.

Tucker 동등화 방법은 공통문항 점수의 평균과 분산의 차이를 통해 두 모집단 능력의 차이를 교정하는 방법으로, 검사를 통해 추정이 되지 않는 모수들을 추정하기 위해 두 가지 가정을 한다. 첫 번째는 검사 X의 점수와 공통문항 V의 점수간의 관계와 검사 Y의 점수와 공통문항 V의 점수간의 관계는 집단 1과 집단 2에서 동일한 선형 회귀식을 갖는다는 것이고, 두 번째는 공통문항 V의 어떤 점수 하에서 검사 X의 조건부 분산은 집단 1과 집단 2에서 동일하며, 검사 Y의 조건부 분산도 동일하다는 것이다(남현우, 2001). 이를 통해 결합 집단의 평균과 분산을 계산하는 공식을 도출한다.

그리고 Braun-Holland 동등화 방법은 빈도추정 동백분위동등화와 같은 가정에 의해 도출한 평균과 표준편차를 이용하는 선형동등화 방법을 사용하여 동등화 점수를 산출하게 된다. 이는 Tucker 선형동등화 방법과 관련이 있으며, 전체검사의 회귀식에서 공통문항이 선형이 아닌 경우의 Tucker 방법의 일반화라고 할 수 있다.

Ⅲ. 연구 방법

Wang(2011)은 모수와 비모수 부트스트랩 방법에 의한 빈도추정 동백분위동등화 방법을 사용하는 모의실험 연구를 하여 공통문항 비동등집단 자료수집 설계에서 피험자의 능력차이(두 집단간 평균차이가 표준정규분포에서 0, 0.1, 0.25일 경우), 피험자의 수(300명, 1,000명, 3,000명), 문항의 수(검사 1은 36문항, 검사 2와 검사 3은 75문항), 공통문항의 비율(전체 문항수에 대한 비로 1:3, 1:5)을 조건으로 설정하였다. Wang(2011)은 다양한 검사동등화 방법, 검사 난이도와 관련된 조건, 피험자 능력의 표준편차 차이와 구검사와 신검사 사이의 조건 등을 후속 연구로 제안한 바 있다.

이에 본 연구는 빈도추정 동백분위동등화 외에 Tucker 동등화 방법, Braun-Holland 동등화 방법을 포함한다. 또한, 각 검사의 난이도 관련 조건들을 추가한다. 각 검사 난이도 조건은 검사 난이도 수준과 검사간 난이도 차이를 포함하는데, 검사간 난이도 차이에는 이전의 연구들과 달리 구검사가 신검사보다 어려운 경우도 고려한다. 그리고 피험자의 능력조건을 다양화한다. Wang(2011)은 피험자의 능력 차이에서 평균만을 고려하였는데 이 연구에서는 피험자 능력 차이

에 각 검사 응시집단의 표준편차 차이를 추가하고 이전의 연구들과는 달리 구검사 피험자가 신검사 피험자보다 능력이 좋은 경우의 조건도 포함한다.

1. 모의실험 설계

본 연구에서는 검사 난이도 수준, 검사간 난이도 차이, 피험자 능력 평균 차이, 피험자 능력 표준편차 차이 등 4가지의 조건들이 고려되었다. 본 연구에서 사용된 조건들을 정리하면 다음과 같다.

첫째, 검사의 난이도가 쉽거나('하', N(-1.6, 1)) 중간이거나('중', N(0, 1)) 어려울 경우('상', N(1.6, 1))를 고려한다. 둘째, 검사간 난이도 차이는 같은 경우(0.0), 조금 나는 경우(0.3), 많이 나는 경우(0.5)를 연구조건으로 하는데 이때 구검사가 신검사보다 어려운 경우도 포함하여 설정한다. 이와 같은 난이도와 관련된 조건들은 반재천과 김선(2015)의 기준을 참고한 것이다. 셋째, 피험자 능력 평균 차이 조건은 같은 경우(0.0), 조금 나는 경우(0.1), 많이 나는 경우(0.25)이다. 검사동등화에서 일반적으로 평균차이 0.1은 '차이가 있음', 0.25는 '매우 큰 차이'로 간주된다(Wang & Brennan, 2009). 이때 구검사를 기준으로 새로운 검사를 보는 피험자 능력이 더 낮은 경우도 고려한다. 넷째, 피험자 능력 표준편차 차이는 이현숙과 김성훈(2010)의 기준을 참고하여 같은 경우(1.0), 평균을 중심으로 어느 정도 퍼져 있는 경우('보통', 0.64), 평균을 중심으로 매우 제한된 범위에 밀집되어 있는 경우('좁음', 0.25)로 설정한다.

피험자 수와 문항수는 Wang(2011)의 연구결과를 참고하여 고정하여 사용한다. 피험자 수가 3,000명 이상이면 모수와 비모수 부트스트랩의 결과가 유사하게 나타났으므로 피험자 수는 1,000명으로 한다. 문항수는 36문항보다는 75문항일 경우에 피험자 능력에 따라 모수와 비모수 부트스트랩의 결과가 차이가 있게 나타나므로 75문항으로 한다. 공통문항 비율은 1:3 또는 1:5의 비율을 사용하여 공통문항 수에 차이를 두는 것이 모수와 비모수 부트스트랩 방법을 사용하는 것에 뚜렷한 영향을 보이지 않았으므로 전체 문항의 20% 비율인 15문항을 사용한다. 이는 공통문항이 전체 문항수의 20% 정도일 때 내용 대표성을 확보할 수 있기 때문이다(남현우, 2001, 재인용). 공통문항은 각 조건에서 Y검사의 문항 모수와 같도록 설정하고, Y검사 고유문항 60문항과 공통문항 15문항을 생성한다.

<표 1> 모의실험 조건

검사 난이도	검사간 난이도 차이		피험자 능력 차이	
	Y 검사 및 공통문항	X 검사	Y 검사	X 검사
하	N(−1.6, 1)	N(−1.6, 1) N(−1.3, 1) N(−1.1, 1) N(−1.9, 1) N(−2.1, 1)	N(0.0, 1)	N(0.0, 1) N(0.1, 1) N(0.1, 0.64) N(0.1, 0.25) N(0.25, 1) N(0.25, 0.64) N(0.25, 0.25) N(−0.1, 1) N(−0.1, 0.64) N(−0.1, 0.25) N(−0.25, 1) N(−0.25, 0.64) N(−0.25, 0.25)
중	N(0, 1)	N(0, 1) N(0.3, 1) N(0.5, 1) N(−0.3, 1) N(−0.5, 1)		
상	N(1.6, 1)	N(1.6, 1) N(1.9, 1) N(2.1, 1) N(1.3, 1) N(1.1, 1)		

또한, 이들 조건에 의해서 생성된 자료를 통한 동등화 표준오차의 계산은 다음절차로 수행된다. 첫째, 부트스트랩 방법은 모수 부트스트랩과 비모수 부트스트랩 방법이 적용된다. 둘째, 모수 부트스트랩 방법을 사용할 때의 모수추정은 이변량 로그선형모형을 사용한다. Wang(2011)은 동등화 표준오차의 추정에서 C_X와 C_V의 값을 2, 3, 4, 5, 6 등의 5개로 연구하였는데, 그 결과 4와 6 사이가 동등화 표준오차의 추정에 비교적 정확한 결과로 나타났다. 따라서 본 연구에서도 4와 6 사이의 값을 사용한다. 이때 Wang(2011)은 연구 조건 수가 많아져 수행하는 것이 복잡해지지 않도록 $C_{IX} = C_{IV} = 2$ 일 때 벡터적률은 주변 다항식 값 C_X와 C_V가 짝수인 경우에만 함께 조건으로 포함시켰는데, 본 연구에서도 C_X와 C_V가 4와 6일 때에만 함께 조건으로 고려한다. Y형 검사도 이와 같은 조건을 가진다. 이에 모수 부트스트랩 방법의 이변량 로그선형모형의 조건을 <표 2>와 같이 정리하고, 모수모형을 간이하게 구분하기 위해 C_X와 C_V의 숫자와 C_{IX}와 C_{IV}의 숫자를 이용하여 모수모형 조건 번호를 다음과 같이 설정한다.

<표 2> 모수 부트스트랩 방법에서 이변량 로그선형모형 조건

모수 모형	X검사				Y 검사			
	C_X	C_V	C_{IX}	C_{IV}	C_Y	C_V	C_{IY}	C_{IV}
41	4	4	1	1	4	4	1	1
42	4	4	2	2	4	4	2	2
51	5	5	1	1	5	5	1	1
61	6	6	1	1	6	6	1	1
62	6	6	2	2	6	6	2	2

2. 모의자료 생성 및 동등화

이 연구의 조건은 검사 난이도 수준(3), 검사간 난이도 차이(5), 피험자 능력 평균 차이(4), 피험자 능력 표준편차 차이(3)로 총 195가지($3\times5\times\{1+(4\times3)\}$)이다. Wang(2011)의 연구에서 검사 1은 실제 자료이고 이 실제 자료를 기반으로 문항모수를 추정하여 검사 2와 검사 3의 모의검사들을 구성하고자 했는데, 실제 자료의 형태가 3모수 IRT(Item Response Theory) 모형을 사용하여 문항모수를 추정하는 것에 적합하여 그에 따라 자료를 생성하였다. 이 연구에서도 동일한 방법으로 모의실험을 위한 자료를 생성한다. 모든 문항의 IRT 변별도 모수 a는 Log(0, 0.3), 추측도 모수 c는 U(0.05, 0.35)로 동일하도록 하고, 문항난이도 모수는 이 연구에서 설정한 난이도 조건에 따라 정규분포에서 무선 생성하여 100,000명의 모의자료 세트를 구성하고 각 조건별 자료 세트에서 1,000명을 표집하여 빈도추정 동백분위동등화, Tucker 동등화, Braun-Holland 동등화 방법으로 동등화 표준오차를 산출한다. 이때, 모수 부트스트랩과 비모수 부트스트랩 방법의 비교를 위해 Wang(2011)의 연구와 같이 $B=1,000$의 시도로 동등화 표준오차를 추정하기 위한 부트스트랩 표본쌍들이 추출되고, $R=300$으로 Wang(2011)보다 100번이 더 해진 수이고, 각 조건에 따라 300회씩 반복 실시하여 평균적인 경향을 파악하려 한다. 여기에서 $R=300$은 실험의 신뢰를 높이기 위해 결정된 것으로 이현숙과 김성훈(2010)의 연구에서도 300회를 반복하여 실시하였고, $B=1,000$은 이전 연구들(Cui & Kolen, 2008; Kolen & Brennan, 2004; Wang, 2011)의 조건을 참고한 것이다. 위와 같은 모의자료 생성, 동등화 및 모수와 비모수 부트스트랩 방법을 이용한 동등화 표준오차의 추정에는 R(3.3.2)이 사용된다.

3. 평가준거

각 검사동등화 방법의 동등화 오차는 RMSE(Root mean squared error)를 평가준거로 사용한다. RMSE는 동등화 표준오차의 bias, 동등화 표준오차의 SE(Standard error)를 이용하여 계산되

고, 이는 동등화된 점수 각각의 동등화 오차이다. 이 연구에서는 모든 점수를 포함하는 종합적인 오차를 나타내기 위해서 동등화 표준오차의 WRMSE(Weighted root mean squared error)로 결과를 해석한다. 이와 관련된 식은 다음과 같다.

$$RMSE(x_i) = \sqrt{BIAS^2(x_i) + SE^2(x_i)} \tag{3}$$

RMSE의 계산을 위한 Bias와 SE의 식은 다음과 같다.

$$Bias(x_i) = \overline{se^*}[eq_Y(x_i)] - \sigma[eq_Y(x_i)] \tag{4}$$

여기에서 $\overline{se^*}[eq_Y(x_i)] = \frac{1}{300}\sum_{r=1}^{300} se_r^*[eq_Y(x_i)]$으로 이는 무선표본에 대한 300회 반복에서의 평균 동등화 표준오차이다. $se_r^*[eq_Y(x_i)]$는 $B=1{,}000$ 부트스트랩 표본들에 대한 r^{th} 무선표본에서의 동등화 표준오차이다. 그리고 $\sigma[eq_Y(x_i)]$는 모집단 문항모수로 산출한 점수 i에서의 동등화 표준오차이다.

$$SE(x_i) = \sqrt{\frac{\sum_{r=1}^{300}\left\{se_r^*[eq_Y(x_i)] - \overline{se^*}[eq_Y(x_i)]\right\}^2}{300-1}} \tag{5}$$

여기에서 $se_r^*[eq_Y(x_i)]$와 $\overline{se^*}[eq_Y(x_i)]$는 위 식(4)에서 계산된다.

그리고 모든 점수대에 대한 종합적인 오차를 나타내기 위한 동등화 표준오차의 WRMSE는 다음의 식으로 계산된다.

$$WRMSE(x) = \sum_{i=0}^{K} \frac{f_1(x_i) + f_2(x_i)}{2} \times RMSE(x_i) \tag{6}$$

여기에서 $f_1(x_i)$는 점수 i에서 신검사 모집단 점수 분포의 상대적 빈도이고, $f_2(x_i)$는 점수 i에서 구검사 모집단 점수 분포의 상대적 빈도이다. K는 검사 문항수로 이 연구에서는 75이고, $RMSE(x_i)$는 (3)의 식에 의해 정의된다.

Ⅳ. 연구 결과

본 연구에서 산출된 WRMSE 결과를 표로 제시하였다. 연구결과는 Cui & Kolen(2008), Wang(2011) 등을 참고하여 각 조건에 대한 주요효과 결과를 중심으로 정리되었다. 우선, <표 3>에 의하면 모든 조건들에서 산출된 WRMSE의 평균으로 살펴보았을 때, 각 동등화 방법들에 대한 WRMSE는 빈도추정 동백분위동등화 방법이 다른 방법들보다 작은 값을 가졌으며, 모수 부트스트랩을 사용한 경우에 더 작았다. Tucker 동등화 방법과 Braun-Holland 동등화 방법들은 모수 부트스트랩을 이용한 것보다 비모수 부트스트랩 방법으로 동등화 표준오차를 구하는 경우의 WRMSE가 더 작았다.

<표 3> 부트스트랩 방법별 각 검사동등화 방법의 WRMSE

검사동등화 방법	부트스트랩 방법	
	모수 부트스트랩	비모수 부트스트랩
빈도추정 동백분위	0.438	1.142
Tucker	0.821	0.730
Braun-Holland	0.817	0.729

모수 부트스트랩의 이변량 로그선형모형에서 C 값의 조건에 따른 각 검사동등화 방법의 WRMSE 결과인 <표 4>에 의하면, 모수모형 조건에 따른 WRMSE는 모든 조건에서 빈도추정 동백분위동등화 방법이 가장 작았다. 그 중 빈도추정 동백분위동등화 방법으로 C_X와 C_V, C_Y와 C_V 값이 4, C_{IX}와 C_{IV}, C_{IY}와 C_{IV}의 값이 2에서 모수 부트스트랩 방법을 사용할 때 WRMSE가 가장 작았다. 모수모형 조건62의 빈도추정 동백분위동등화 방법에서 모수 부트스트랩을 사용하는 경우에도 비교적 작은 WRMSE를 가졌다.

<표 4> 이변량 로그선형모형 조건별 각 검사동등화 방법의 WRMSE

부트스트랩 방법		검사동등화 방법		
		빈도추정 동백분위	Tucker	Braun-Holland
모수	41	0.470	0.810	0.808
	42	0.318	0.922	0.906
	51	0.491	0.810	0.808
	61	0.499	0.778	0.775
	62	0.412	0.787	0.788
비모수		1.142	0.730	0.729

<표 5>에는 검사 난이도 수준 조건별 WRMSE가 제시되었다. 이 표에 의하면 모든 검사 난이도에서 모수 부트스트랩을 사용하는 빈도추정 동백분위동등화 방법이 작은 WRMSE를 가졌고, WRMSE가 가장 작은 경우는 모수모형 조건42였는데 검사가 보통이거나 어려운 경우에 WRMSE가 더 작았다. 한편, 검사 난이도가 낮거나 보통일 경우에는 Tucker와 Braun-Holland 동등화 방법은 비모수 부트스트랩을 사용하는 것이 비교적 작은 WRMSE를, 검사가 어려운 경우에는 모수모형 조건42의 모수 부트스트랩을 사용하는 것이 가장 작은 WRMSE를 가졌다.

<표 5> 검사 난이도 수준에 따른 WRMSE

검사 난이도	부트스트랩 방법		검사동등화 방법		
			빈도추정 동백분위	Tucker	Braun-Holland
하	모수	41	0.654	0.913	0.906
		42	0.370	1.382	1.340
		51	0.661	0.913	0.906
		61	0.658	0.921	0.916
		62	0.471	0.900	0.898
	비모수		1.042	0.860	0.870
중	모수	41	0.379	0.756	0.759
		42	0.293	0.809	0.790
		51	0.379	0.756	0.759
		61	0.381	0.693	0.696
		62	0.338	0.686	0.689
	비모수		1.175	0.658	0.645
상	모수	41	0.379	0.760	0.760
		42	0.291	0.576	0.588
		51	0.432	0.760	0.760
		61	0.459	0.721	0.715
		62	0.426	0.774	0.778
	비모수		1.210	0.671	0.673

다음 <표 6>에는 검사간 난이도 차이에 대한 결과가 제시되었다. 이 표에 의하면 검사간 난이도 차이에 상관없이 빈도추정 동백분위동등화 방법으로 모수 부트스트랩을 사용하는 것이 안정적인 결과를 보여주었으며 모수모형 조건42를 사용한 경우의 WRMSE가 가장 작았다. 그리고 검사간 난이도 차이가 같거나 조금 차이가 있거나 많이 차이가 있는 경우, 신검사가 구검사보다 어려운 경우에는 비모수 부트스트랩에 따른 Tucker 동등화 방법과 Braun-Holland 동등화 방법을 사용하는 것이 더 낮은 WRMSE를 가졌다. 한편, 구검사가 신검사보다 어려운 경우에는 <표

6>에서 볼 수 있듯이 모수모형 조건61의 모수 부트스트랩 방법 사용을 사용하는 것이 Tucker 동등화 방법과 Braun-Holland 동등화 방법의 WRMSE를 낮게 하였다.

<표 6> 검사간 난이도 차이에 따른 WRMSE

검사간 난이도 차이	부트스트랩 방법		검사동등화 방법		
			빈도추정 동백분위	Tucker	Braun-Holland
같음	모수	41	0.482	0.878	0.876
		42	0.312	0.930	0.908
		51	0.503	0.878	0.876
		61	0.510	0.790	0.787
		62	0.419	0.732	0.734
	비모수		1.214	0.720	0.702
조금	모수	41	0.469	0.795	0.795
		42	0.318	0.939	0.922
		51	0.489	0.795	0.795
		61	0.498	0.760	0.757
		62	0.413	0.794	0.795
	비모수		1.128	0.731	0.733
많이	모수	41	0.466	0.791	0.788
		42	0.321	0.902	0.889
		51	0.486	0.791	0.788
		61	0.495	0.790	0.788
		62	0.407	0.806	0.808
	비모수		1.121	0.733	0.739
Y<X	모수	41	0.463	0.817	0.812
		42	0.311	0.939	0.924
		51	0.494	0.817	0.812
		61	0.502	0.797	0.796
		62	0.420	0.769	0.769
	비모수		1.111	0.707	0.719
Y>X	모수	41	0.472	0.769	0.770
		42	0.328	0.902	0.887
		51	0.481	0.769	0.770
		61	0.491	0.753	0.749
		62	0.400	0.832	0.834
	비모수		1.138	0.757	0.754

다음 <표 7>, <표 8>에는 각각 피험자 능력 평균 차이, 피험자 능력 표준편차 차이의 조건별

WRMSE가 제시되었다. 이들 표에 의하면 모두 조건42를 사용하는 모수 부트스트랩 방법에 따른 빈도추정 동백분위동등화 방법으로 동등화 표준오차의 WRMSE를 산출하는 것이 가장 효과적이었다. 피험자 능력 차이와 관련된 조건들에서 Tucker 동등화 방법과 Braun-Holland 동등화 방법을 사용하는 것에는 대부분 비모수 부트스트랩 방법을 사용하는 것이 비교적 낮은 WRMSE를 가졌다. 그러나 <표 8>에서 볼 수 있듯이 피험자 능력의 표준편차 차이가 '좁음'으로 분포가 평균 중심으로 밀집되어 있는 경우에는 조건62의 모수 부트스트랩 방법으로 Tucker 동등화 방법과 Braun-Holland 동등화 방법을 사용하는 것이 더 낮은 WRMSE를 나타내었다.

<표 7> 피험자 능력 평균 차이에 따른 WRMSE

피험자 능력 차이 평균	부트스트랩 방법		검사동등화 방법		
			빈도추정 동백분위	Tucker	Braun-Holland
같음	모수	41	0.469	0.893	0.900
		42	0.305	1.010	0.995
		51	0.485	0.893	0.900
		61	0.495	0.806	0.809
		62	0.444	0.857	0.863
	비모수		1.325	0.731	0.738
조금	모수	41	0.474	0.815	0.813
		42	0.318	0.920	0.903
		51	0.495	0.815	0.813
		61	0.500	0.772	0.768
		62	0.405	0.763	0.761
	비모수		1.131	0.727	0.729
많이	모수	41	0.467	0.791	0.788
		42	0.320	0.910	0.894
		51	0.487	0.791	0.788
		61	0.499	0.780	0.777
		62	0.414	0.798	0.803
	비모수		1.123	0.732	0.728
Y<X	모수	41	0.479	0.789	0.788
		42	0.320	0.895	0.878
		51	0.493	0.789	0.788
		61	0.497	0.776	0.771
		62	0.403	0.761	0.760
	비모수		1.113	0.730	0.736
Y>X	모수	41	0.463	0.817	0.813
		42	0.318	0.935	0.919
		51	0.489	0.817	0.813
		61	0.503	0.775	0.774
		62	0.415	0.800	0.803
	비모수		1.140	0.729	0.722

<표 8> 피험자 능력 표준편차 차이에 따른 WRMSE

피험자 능력 차이 표준편차	부트스트랩 방법		검사동등화 방법		
			빈도추정 동백분위	Tucker	Braun-Holland
같음	모수	41	0.461	0.799	0.799
		42	0.311	1.023	1.007
		51	0.479	0.799	0.799
		61	0.488	0.784	0.784
		62	0.409	0.833	0.835
	비모수		1.189	0.725	0.724
보통	모수	41	0.475	0.820	0.819
		42	0.318	0.899	0.884
		51	0.494	0.820	0.819
		61	0.497	0.768	0.764
		62	0.409	0.765	0.765
	비모수		1.145	0.714	0.713
좁음	모수	41	0.478	0.814	0.809
		42	0.326	0.820	0.801
		51	0.501	0.814	0.809
		61	0.515	0.780	0.777
		62	0.418	0.750	0.751
	비모수		1.081	0.751	0.752

V. 논의 및 결론

검사동등화를 실시하여 나타나게 되는 동등화 오차는 검사동등화를 위한 자료 수집 방법, 검사동등화 방법의 선택, 검사동등화 결과의 신뢰성 등을 결정할 수 있도록 하는 매우 중요한 부분이다. 검사는 그 특성과 상황이 매우 다양하게 나타나므로 각각의 경우에 대해 동등화 오차를 줄일 수 있는 방법을 선택하여 적합한 검사동등화 방법을 탐구하는 것이 필요하다. 따라서 본 연구는 검사동등화의 표준오차를 구하는 방법 중 모수와 비모수 부트스트랩 방법을 사용하여 비동등집단 공통문항 자료 설계의 여러 가지 조건에 따른 각 검사동등화 방법의 비교를 통해 각 조건별로 더 안정적인 검사동등화 방법을 탐구하였다. Wang(2011)의 연구를 확장하여 연구의 조건은 검사 난이도 수준, 검사간 난이도 차이, 피험자 능력 평균 차이, 피험자 능력 표준편차 차이 등으로 설정하였고, 부트스트랩 방법과 이변량 로그선형모형의 조건을 달리하며 각 조건에서 빈도추정 동백분위동등화, Tucker 동등화, Braun-Holland 동등화 방법을 적용하였다.

먼저, 세 가지 검사동등화 방법을 비교한 결과 모수 부트스트랩 방법을 사용한 빈도추정 동백분위동등화 방법이 가장 낮은 WRMSE를 가졌다. Wang(2011)의 연구결과에서도 빈도추정 동백분위동등화 방법을 사용할 때 비모수보다는 모수 부트스트랩을 사용하는 것이 더 안정적이었다. 그리고 동백분위 방법이나 빈도추정 동백분위동등화 방법을 사용하여 동등화 표준오차의 추정을 위한 모수와 비모수 부트스트랩 방법들을 비교하는 이전 연구에서도 대부분의 조건들에서 비모수 부트스트랩 방법보다 모수 부트스트랩 방법이 더 정확하게 동등화 표준오차를 추정하였던 것과 유사한 결과이다(Cui & Kolen, 2007; Wang & Zhang 2009). 그러나 Tucker 동등화 방법과 Braun-Holland 동등화 방법은 비모수 부트스트랩 모형을 사용할 경우에 더 낮은 WRMSE를 보여 검사동등화 방법에 따라 부트스트랩 방법을 다르게 사용해야 동등화 표준오차를 낮게 만들 수 있다는 점을 보여주었다. 모수 부트스트랩 방법의 모수모형은 이변량 로그선형모형이 사용되었는데 모형의 조건 중에는 조건42의 WRMSE가 가장 낮게 나타났다. 이는 Wang(2011)의 연구 결과와 유사한 것으로 C_X와 C_V의 값이 클수록 완곡화가 덜 이루어진 분포를 가지기 때문에 더 큰 SE를 가지지만, C_X와 C_V의 값이 작아지면 더 완곡화되어 bias가 커지게 되므로 C_X와 C_V의 값이 4인 경우가 적정하여 WRMSE가 낮게 나타난 것으로 볼 수 있을 것이다. 이에 이러한 연구의 결과는 일반적으로 모수 부트스트랩을 사용하는 모수모형의 조건을 결정하는 것에 활용될 수 있다. 하지만 Tucker 동등화 방법과 Braun-Holland 동등화 방법은 오히려 대부분의 조건에 대해 모수모형 조건42에서 WRMSE가 크게 나타나고 비모수나 61, 62의 조건을 가지는 모수 부트스트랩 방법을 사용할 때 더 낮은 WRMSE를 가지는 것을 볼 수 있으므로 검사동등화 방법에 따라 부트스트랩 방법 및 모수모형을 선택하는 것에 참고할 수 있다. 또한, Kolen과 Brennan(2014)은 Braun-Holland 동등화 방법은 전체 검사의 관계식에서 공통문항이 선형이면 Tucker 동등화 방법과 유사한 결과가 도출된다고 하였는데 본 연구에서도 그와 같이 두 동등화 방법이 같은 경향의 결과를 보이고 있다.

검사 난이도별 조건에 따른 WRMSE는 검사 난이도 수준이 상, 중, 하의 경우, 검사간 난이도 차이가 없거나 있는 경우 등을 살펴보았을 때 모두 빈도추정 동백분위동등화 방법을 사용하는 모수모형 조건42로 모수 부트스트랩을 사용한 경우에 낮게 나타났다. 한편, Tucker 동등화 방법과 Braun-Holland 동등화 방법에서는 대부분 비모수 부트스트랩을 사용하는 것이 모수 부트스트랩을 사용하는 것보다 낮은 WRMSE를 가졌는데, 검사의 난이도가 어려운 경우에는 모수모형 조건42를 사용하여 모수 부트스트랩으로 WRMSE를 구하는 것이 가장 낮은 값을 가지는 것으로 나타났고, 구검사가 신검사보다 어려운 경우에는 모수 부트스트랩 방법의 모수모형 조건61을 사용하는 것이 가장 낮은 WRMSE를 가졌다.

피험자 능력의 차이에서는 피험자 능력의 평균과 표준편차가 같거나 차이가 있는 모든 조건

에서 모수모형 조건42의 모수 부트스트랩을 사용하는 빈도추정 동백분위동등화 방법에서 가장 낮은 동등화 표준오차의 WRMSE 값을 나타내었다. 그리고 Tucker 동등화 방법과 Braun-Holland 동등화 방법에서는 대부분 비모수 부트스트랩을 사용하는 것이 모수 부트스트랩을 사용하는 것보다 나은 결과를 보여주었으나, 피험자 능력의 표준편차 분포가 평균 중심으로 밀집되어 있는 경우에는 모수 부트스트랩 방법을 사용하여 모수모형 조건62로 WRMSE를 산출하는 것이 가장 낮은 WRMSE 값을 가졌다. 한편, 피험자 능력 차의 평균과 표준편차가 같은 경우에는 빈도추정 동백분위동등화 방법을 사용할 때 비모수 부트스트랩으로 산출된 WRMSE가 매우 높아지는 것을 확인할 수 있어서 사용이 권장되지 않는다.

본 연구는 Wang(2011)의 연구를 참고하여 조건을 추가하거나 다양화하여 진행하였다. Wang(2011)의 연구결과와 유사한 결과로 모수모형 조건42의 모수 부트스트랩을 이용하여 동등화 표준오차를 추정하는 빈도추정 동백분위동등화 방법을 사용하는 것이 모든 조건에서 안정적이므로 일반적으로 사용할 수 있음이 확인되었다. 또한, Tucker와 Braun-Holland 동등화 방법은 비모수 부트스트랩 방법을 사용하는 것이 대부분 적절하다고 할 수 있겠으나 각 조건에 따라 더 낮은 WRMSE을 가지는 모수 부트스트랩 방법이 존재하는 경우가 있었다.

본 연구는 검사동등화 방법 면에서 Wang(2011)의 연구보다 두 가지를 더 포함하여 연구를 진행하였으나, IRT 검사동등화 방법들과 같은 다른 동등화 방법들이 포함되지 않았다. 따라서 더욱 다양한 검사동등화 방법을 사용하여 동등화 표준오차 산출과 관련된 연구를 진행할 필요가 있다. 그리고 공통문항을 사용하여 검사동등화가 이루어질 때 공통문항은 중요한 역할을 하게 되므로 공통문항이 전체 문항을 대표하는 특성 정도 등의 공통문항과 관련된 조건들을 다양화하여 적절한 검사동등화 방법 및 부트스트랩 방법에 대한 연구가 필요하다.

참고문헌

남현우(2001). **검사동등화 방법.** 서울; 교육과학사.

반재천, 김선(2015). 소표본 동등화를 위한 명목 가중 평균동등화 방법의 동등화 오차 연구. **교육평가연구, 28**(4). 1049-1075.

이현숙, 김성훈(2010). 외적 가교검사의 통계적 구성 조건 완화가 검사동등화 결과에 미치는 영향. **교육평가연구, 23**(2), 417-439.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. Wiley. New York.

Cui, Z., & Kolen, M. J. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement, 32,* 334-347.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* New York: Cambridge University Press.

Efron. B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Number 38 in CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM

Efron. B., & Tibshirani. R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17,* 225-237.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics 25,* 133-183.

Hou, J. (2007). *Effectiveness of the hybrid Levine equipercentile and modified frequency estimation equating methods under the common-item nonequivalent groups design.* Unpublished Doctoral Dissertation, The University of Iowa.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9,* 209-223.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York: Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking: Methods and practices*

(3nd ed.). New York: Springer.

Wang, C. (2011). *An Investigation of bootstrap methods for estimating the standard error of equating under the common-item nonequivalent groups design.* Unpublished Doctoral Dissertation, The University of Iowa.

Wang, T. & Brennan, R.L. (2009). A modified frequency estimation equating method for the common-item nonequivalent groups design estimating random error in equipercentile equating. *Applied Psychological Measurement, 33,* 118-132.

Wang, C., & Zhang, S. (2009). *Bootstrapping to estimate standard errors of equating: Parametric or nonparametric?* Paper presented at annual meeting of National Council of Measurement in Education, April, San Diego, CA.

Zeng, L. (1991). *Standard errors of linear equating for the single-group design.* (Research Report 91-4) Iowa City, IA: ACT.

*논문접수 2018년 2월 2일 / 1차 심사 2018년 3월 9일 / 2차 심사 2018년 4월 10일 / 게재승인 2018년 6월 14일

* 김화영: 성균관대학교 사범대학 교육학과를 졸업하고, 동대학원 교육학과에서 박사수료를 하였다. 현재 한국행동과학연구소에서 책임연구원으로 재직 중이다.
* E-mail: alpsgirl@hanmail.net

* 김현철: 성균관대학교 통계학과를 졸업하고, 미국 University of Florida에서 교육학 박사학위를 취득하였다. 현재 성균관대학교 사범대학 교육학과 교수로 재직 중이다.
* E-mail: hkim@skku.edu

Abstract

Comparison of Test Equating Methods by estimating the Standard Error of Equating using the Bootstrap Methods

Kim, Hwa Young*
Kim, Hyunchul**

The purpose of this study is to compare the three test equating methods- frequency estimation equating, Tucker equating, and Braun-Holland equating- in common-item nonequivalent groups design using the parametric and nonparametric bootstrap for estimating the standard error of equating. The conditions- The levels of test difficulty, the difference in difficulty between forms, the group ability- are considered.

The results show that the frequency estimation equating with parametric bootstrap in specific parametric model is more stable than other equating methods under all conditions. Tucker and Braun-Holland equating have more reliable results using nonparametric bootstrap under almost all conditions, but are relatively more stable using parametric bootstrap in specific parametric models depending on certain conditions.

Key words: Common-item nonequivalent groups design, bootstrap, log-linear, test equating

* First author, Head Researcher, Korea Institute for Research in the Behavioral Sciences
** Corresponding author, Professor, Sungkyunkwan University