아시아교육연구 20권 3호
Asian Journal of Education
2019, Vol. 20, No. 3, pp. 797-829.
https://doi.org/10.15753/aje.2019.09.20.3.797

소규모 집단에 대한 동등화 방법 비교: 모의실험 연구를 중심으로*

김화영(金和泳)**
김현철(金顯哲)***

논문 요약

본 연구는 피험자 수가 적은 검사를 동등화하는 상황에서 공통문항 비동등집단 설계의 다양한 조건에 대해 동등화 방법들의 비교를 통해 각 조건에서 효과적인 동등화 방법을 탐색하고자 하는 모의실험 연구이다. 비교의 조건은 문항 수, 공통문항 비율, 피험자 수, 검사난이도 수준, 검사간 난이도 차이, 피험자 능력 평균 차이, 피험자 능력 표준편차 차이 등으로 설정되었고, 소규모 집단 동등화의 경우에 효과적이라고 연구되어 온 동일동등화, 평균동등화, 명목 가중 평균동등화, Tucker 선형동등화, 연쇄 선형동등화, 원호동등화, 합성동등화 방법 등 7가지 동등화 방법이 적용되었다.

연구결과 대부분의 조건에서 평균동등화와 원호동등화 방법이 대체로 안정적인 결과를 보였고 조건에 따라 두 방법 중 조금 더 효과적인 방법이 있었다. 또한, 두 검사가 어렵고 검사간 난이도 차이가 없는 경우, 두 검사간 난이도 차이가 없고 피험자 능력 평균 차이도 없는 경우, 검사간 난이도 차이가 조금 있고 피험자 능력 평균 차이가 많이 나는 경우에는 동일동등화가 가장 낮은 동등화 오차를 나타냈다.

주요어 : 소규모 집단, 소표본 동등화, 검사 동등화, 모의실험

* 본 연구는 성균관대학교 일반대학원 '소규모 집단에 대한 동등화 방법 비교(2019)' 박사학위 논문을 기반으로 수정·보완한 것임.
** 제1저자: 한국행동과학연구소 연구위원
*** 교신저자: 성균관대학교 교육학과 교수

Ⅰ. 서론

같은 능력을 측정하는 유사한 형식과 통계적 특성을 가지는 검사를 동형검사(equivalent form)라고 한다. 동형검사로 제작된 검사들이 같은 구인을 재고 일반적으로 같은 검사명세표 또는 검사청사진에 의해 만들어졌다 할지라도, 한 검사의 다른 판(edition)이나 형(form)은 통계적 특성들이 다소 다르게 된다(Brennan, 2015). 동형검사의 난이도와 변별도를 동일하게 제작하는 것은 현실적으로 불가능하다는 의미이다. 따라서 한 검사의 여러 형에서 나온 점수들이 각 검사 형의 난이도 및 다른 통계적 특성들에 영향을 받지 않고 상호 교환되어 사용될 수 있도록 하기 위해서는 그 점수들을 조정하는 통계적 과정인 검사 동등화(test equating)가 필요하다(Kolen & Brennan, 2004). 검사 동등화는 다른 형의 검사를 치르는 피험자들에게 공정함을 주기 위해 필요하며, 어떤 형의 검사를 치르는지에 관계없이 같은 의미의 점수를 제공할 수 있게 한다(Angoff, 1971; Kolen & Brennan, 2004). 검사 동등화를 수행할 때에는 각 검사의 특성 및 상황을 고려하여 바람직한 동등화 방법을 선택하고 신중하게 사용하며, 점수전환 시 오류를 최소화하도록 하는 노력이 요구된다.

일반적으로 검사 동등화를 위해서는 동등화 결과의 안정성과 정확성 보장을 이유로 많은 사례수가 필요하다(Kolen & Brennan, 2004). 선형동등화(linear equating) 방법을 사용하기 위해서는 적어도 400명, 문항반응이론(item response theory; IRT)의 3모수 모형을 적용하거나 동백분위동등화(equipercentile equating) 방법을 위해서는 적어도 1,500명 이상을 권장하고 있다(Kolen & Brennan, 2004). 그리고 동등화를 하려고 할 때 각 검사 형당 동등화 표본의 수가 100명 이하인 경우 동등화를 하지 않는 것이 바람직하다는 것이 그동안의 정설로 여겨졌다(Hanson, Zeng & Colton, 1994; Kolen & Brennan, 2004). 하지만 대규모로 실시되는 검사가 아니더라도 검사 동등화는 필요하다. 예를 들어, 어떤 검사가 막 개발되어 시행 초기의 상황일 때에는 아직 피험자 수가 많지 않을 것이고, 대중화되지는 않았지만 특수한 자격을 위한 검사의 경우에도 피험자가 많지 않을 수 있다. 현재는 폐지가 된 우리나라 국가영어능력평가시험은 토익, 토플과 같이 회차별로 동형검사를 실시하였다. 이 같은 경우에는 검사 점수의 보고에서 검사 동등화가 적용된 조정점수가 제공되었어야 하지만 시행 초기로 인한 피험자 수 불충분의 문제로 적합한 동등화 방법을 찾지 못하여 원점수와 백분위 점수가 공지되었다. 이렇게 불가피하게 피험자 수가 많지 않게 되는 새로 개발된 검사에 대해서 바람직한 동등화 방법 적용에 대한 충분한 연구가 있다면 신생 검사 시행에 많은 도움이 될 것이다. 또한, 일반적인 학교나 교육 현장에서도 학생의 능력을 측정하기 위한 시험이 실시되므로 검사 동등화가 이루어져야 하는 상황이 분명히 있는데 그 때의 피험자 수는 50명 이하일 경우가 대부분이다. 따라서 규모나 중요도에 상관없이 정확한 측정 결과를 주기 위해 피험자 수가 충분하지 않은 경우 검사 동등화를 해야 하는 상황에 대한 적합한 동등화 방법을 모색해 볼 필요가 있다.

이에 본 연구에서는 검사의 피험자 수가 적은 경우의 검사 동등화 상황에 관심을 두고 여러 가지

조건을 가진 검사에서 어떤 동등화 방법의 적용이 피험자들에게 더 정확한 점수를 제공해줄 수 있을지에 대해 살펴보고자 하였다. 따라서 본 연구의 목적은 소규모 집단을 위해 적절하다고 제안되는 동일동등화(identity equating), 평균동등화(mean equating), 명목 가중 평균동등화(nominal weights mean equating), Tucker 선형동등화, 연쇄(chained) 선형동등화, 원호동등화(circle-arc equating), 합성동등화(synthetic equating) 방법 등 7가지 동등화 방법들을 문항 수, 공통문항 비율, 피험자 수, 검사난이도 수준, 검사간 난이도 차이, 피험자 능력의 평균 및 표준편차 차이 등의 다양한 조건 하에서 적용하여 동등화 오차(error of equating)를 비교하고 각각의 조건에 대해 보다 효과적인 동등화 방법을 탐색하는 것에 있다. 이를 위해 모의실험을 통한 공통문항 비동등집단 설계를 구성하여 연구를 진행하였다.

Ⅱ. 이론적 배경

1. 선행 연구

소규모 집단에 대한 검사 동등화 연구는 피험자 집단의 크기에 따라 효과적인 동등화 방법에 대해 연구하는 것으로 소개되었다. Hanson, Zeng과 Colton(1994), Livingston(1993)은 공통문항 또는 무선집단 설계로 25명부터 3000명까지 사전완곡화(presmoothing), 사후완곡화(postsmoothing) 또는 완곡화하지 않은 동백분위동등화 방법을 사용하였다. Parshall, Houghton과 Kromrey(1995)는 15~100명의 피험자 수로 선형동등화 방법을 사용할 수 있음을 보이며 50~100명일 때 Tucker 선형동등화 방법이 사용될 수 있다고 하였는데, 피험자 수가 작을 때 평균과 떨어진 점수에 대해서 동등화를 하는 것에 대해서는 주의가 필요하다고 밝혔다. Skaggs(2005)는 25명에서 200명의 표본으로 동백분위동등화 방법 외에 선형동등화, 평균동등화, 동일동등화 방법도 사용하여 비교하였는데, 25명일 때에는 동일동등화가 우수하였다. 50명 이상에서는 원점수의 위치에 따라 다른 동등화 방법이 효과적이었는데, 원점수가 평균보다 아래로 해당되는 빈도가 높지 않으면 평균동등화가 비교적 나은 수행을 보였다. 합성동등화 방법도 소개되었는데, 10명, 25명의 피험자 수에 대해서 동일동등화가 더 정확하고, 100명 이상일 때에는 합성동등화의 오차가 작았으며, 대체로 피험자 수가 50~100명 정도일 경우에 사용할 수 있었다(Kim, von Davier, & Haberman, 2008). 또한, 원호동등화 방법을 적용하는 것이 동등화하지 않는 것보다 더 정확한 결과를 제공한다고 제안되기도 하였다(Kim & Livingston 2010; Livingston & Kim, 2008, 2009, 2010). 원호동등화 방법은 평균동등화, Tucker 선형동등화 등의 다른 동등화 방법들보다 더 나은 수행을 보였는데, 특히 10% 아래의 점수와 90% 이상의 점수에서 더 효과적이었다. Babcock, Albano와 Raymond(2012)는 소규모 집단에

대한 동등화를 위하여 명목 가중 평균동등화 방법을 소개하였는데 연구 결과, 명목 가중 평균동등화는 일반적으로 효과적이었으며 어떤 조건에서도 가장 높은 오차를 산출하지 않았으며 원호동등화도 좋은 수행을 가졌다. Kurtz와 Dwyer(2013)은 공통문항 비동등집단 설계로 대규모 국가자격시험에서 자료를 재표집하였는데 연구결과, 평균을 중심으로 멀리 떨어져 있는 준거 점수를 가진 준거참조검사에서 50명 이하 피험자의 경우에 원호동등화와 합성동등화가 비교적 좋은 수행을 보였다. 그리고 Aşiret과 Sünbül(2016)은 무선집단 설계로 피험자 수, 검사간 난이도 차이, 추측도 모수 등을 조건으로 하였는데 연구결과, 피험자 수 50명, 검사간 난이도 차이가 0.4 나면 동등화하지 않는 것보다 동등화하는 것이 효과적이었다. 또한, 대부분의 조건에서 원호동등화와 평균동등화 방법이 비교적 낮은 동등화 오차를 가졌다.

국내에서 원호동등화 방법은 안수현(2016), 우중호(2016), 임의진(2011), 임의진과 이규민(2017) 외에 의해 연구되었다. 명목 가중 평균동등화 방법은 반재천과 김선(2015), 임의진과 이규민(2017)이 소개하였다. 그리고 합성동등화, 평균동등화의 경우 이규민(2005), 임의진과 이규민(2017)의 사용 외에 국내에서 연구된 사례가 거의 없고, Tucker 선형동등화 방법은 여러 동등화 연구에서 사용된 경우는 많으나 소규모 집단을 위한 연구 사례는 거의 없다. 반재천과 김선(2015)은 소표본 동등화를 위한 명목 가중 평균동등화 방법의 동등화 오차의 크기와 차이를 다양한 조건에서 탐구하였는데, 검사난이도가 쉬울 때, 동등화 집단간 능력차이가 없을 때, 표본 크기가 커질 때 동등화 오차가 줄어들었다. 임의진(2011)은 혼합형 검사에서 표본 크기가 작을 때 동등화 방법을 비교하였는데, 모든 점수 구간이 동등하게 여겨진다면 Levine 선형동등화 방법, 구성형 문항으로만 구성된 가교검사를 이용할 때에는 원호동등화 방법, 혼합형 가교검사를 사용할 때에는 Levine 평균 동등화 방법의 동등화 오차가 낮아지는 것으로 나타났다. 또한, 임의진과 이규민(2017)은 표본 크기가 작을 때 혼합형 검사에 대한 더 많은 동등화 방법을 비교하였는데, 대부분 동등화를 하지 않았을 때 동등화 오차가 가장 작았고, 동등화를 실시한 경우에는 합성동등화, 원호동등화, 평균동등화 방법이 보다 정확한 동등화 결과를 산출하는 것으로 나타났다.

이러한 연구 결과를 참고하여 문항 수, 공통문항 비율, 난이도 수준 및 차이 등 검사의 구성 특성들을 달리하거나 피험자 수, 피험자의 능력 차이 등의 피험자와 관련된 조건들을 다르게 하여 연구가 이루어진다면 다양한 조건 하에서 검사 동등화를 시행할 때 그 방법의 선택에 도움이 될 것이다. 또한, 대부분의 연구(반재천, 김선, 2015; 우중호, 2016; 임의진, 이규민, 2017; Kurtz & Dwyer, 2013)에서 동등화의 기준이 되는 기준검사의 피험자 수를 동등화하고자 하는 검사인 변환검사의 피험자 수보다 더 많도록 고정하여 연구를 진행하고 있는데 변환검사의 피험자 수와 유사하거나 다르게 설정하는 연구도 필요하다. 개발이 된지 얼마 되지 않은 시행 초기의 검사의 경우에는 기준이 되는 검사의 피험자 수도 충분하지 않거나 유동적일 수 있기 때문이다. 피험자 능력 차이에 대한 조건도 피험자 능력의 평균 차이에 대한 연구(반재천, 김선, 2015; 우중호, 2016; Aşiret & Sünbül,

2016; Babcock et al., 2012)는 있으나 피험자 능력의 표준편차 차이에 대한 연구는 많지 않다.

2. 소규모 집단에 적합한 동등화 방법

소규모 집단에 비교적 효과적이라고 연구가 되어 온 동등화 방법은 동일동등화, 평균동등화, 명목 가중 평균동등화, Tucker 선형동등화, 연쇄 선형동등화, 원호동등화, 합성동등화 등의 방법이다. 먼저, 동일동등화 방법은 동등화를 하려는 검사들이 완벽히 같을 때 소규모 집단에 대해 동등화하기 위해 사용되는 방법이다(Kolen & Brennan, 2014). 동일동등화 방법은 어떤 동등화도 수행하지 않는 것과 같다. 수식으로 표현하면 다음과 같으며, 여기에서 x는 직접적인 선형 방법으로 기준검사의 원점수 척도와 대치되는 변환검사의 원점수이다.

$$ID_Y(x) = x \tag{1}$$

평균동등화는 각 검사의 평균으로부터 동일한 거리에 있는 점수를 동등한 점수로 간주하는 방식으로 수행된다. 만일 표집 크기가 100 이하로 작아진다면 선형동등화의 표준오차는 매우 커질 것인데, 그러한 경우에 평균동등화가 고려될 수 있다(Kolen & Brennan, 2014). Tucker나 Levine 관찰점수 동등화 방법에서 평균동등화에 대한 동등화 식은 $\sigma_s(Y)/\sigma_s(X) = 1$의 식이 주어질 때 다음과 같이 산출된다. 이때 s는 모집단(population) 1과 2에서 표집된 표본의 크기 비중에 따라 결합한 모집단(synthetic population)을 의미한다.

$$m_{Y_s}(x) = [x - \mu_s(X)] + \mu_s(Y) \tag{2}$$

명목 가중 평균동등화는 Babcock 외(2012)가 소규모 집단을 위한 동등화 방법으로 제안한 것으로 Tucker 선형동등화 방법을 단순화한 것이다. 표집 크기가 작기 때문에 표준편차에 대한 추정치가 정확하지 않을 것이므로 두 검사 형의 표준편차가 합성집단(synthetic group)에서 동일하다고 가정하고, 분산과 공분산 또한 소표본일 경우 잘 추정되지 않을 것이므로 회귀계수 대신 공통문항 수와 전체문항 수의 비율을 사용할 것을 제안했다(Babcock et al., 2012). 따라서 명목 가중 평균동등화의 최종적인 동등화 함수 nwm_{Y_s}는 다음과 같다. 여기에서 1은 X형을 치른 집단, 2는 Y형을 치른 집단, V는 공통문항을 의미하며, K_X는 X형의 문항 수, K_Y는 Y형의 문항 수, K_V는 공통문항의 수이다.

$$nwm_{Y_s} = x - \mu_1(X) + \mu_2(Y) + [\frac{(N_X)(K_Y) + (N_Y)(K_X)}{(N_X + N_Y)(K_V)}][\mu_1(V) - \mu_2(V)] \quad (3)$$

Tucker 선형동등화는 공통문항 점수의 평균과 분산의 차이를 통해 두 모집단 능력의 차이를 교정하는 방법으로, 공통문항 점수를 v라고 하고 동등화할 검사들의 점수를 각각 x와 y라고 할 때, v에 대한 x의 선형 회귀와 v에 대한 y의 선형 회귀가 검사 X와 검사 Y를 치른 모든 피험자에게 같아야 한다는 가정을 필요로 한다(남현우, 2001). 공통문항 비동등집단 설계에서 선형동등화를 위해서는 우선 합성집단을 만들고 합성집단이 두 검사 형을 실행했을 때의 동등화 함수를 구해야 한다. 합성집단에서 Tucker 선형동등화 함수는 다음과 같다.

$$l_{Y_s} = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y) \quad (4)$$

연쇄 선형동등화 방법은 Tucker와 Levine 선형동등화 방법보다 단순한 편으로, 1단계에서는 X를 V 척도에 연계하고($l_V(x)$), 2단계에서 V를 Y 척도에 연계한다($l_Y(V)$). 그리고 3단계에서 $l_Y(x)$로 동등화한 Y를 산출한다($l_Y[l_V(x)]$). 공통문항 비동등집단 설계에서는 1단계가 오직 모집단 1을 사용하여 수행되고, 2단계는 모집단 2를 사용하여 수행되기 때문에 3단계에서 사용할 모집단을 결정해야 한다는 문제가 있다(Kolen & Brennan, 2014). Tucker와 Levine 관찰 점수 절차와 오직 γ에 대한 것이 다르며, 그것은 선형 관찰점수 동등화 기본 식의 모수이다. 다음과 같은 함수를 가진다.

$$l_Y(x) = \{\mu_2(Y) + \gamma_2[\mu_1(V) - \mu_2(V)] - (\gamma_2/\gamma_1)[\mu_1(X)]\} \quad (5)$$

원호동등화 방법은 검사의 세 가지 점수를 사용하여 연결할 때 직선이 나오지 않는 경우 원호를 사용하여 동등화 식을 생성하는 것이다. 동등화 함수는 직선 요소와 곡선 요소로 분해할 수 있고, 곡선 요소를 위한 방법으로 원호 방법을 사용한다(von Davier, Holland, & Thayer, 2004). x_1, y_1은 추측하여 받은 점수를 제외한 실제 획득한 최저 점수, x_2, y_2는 중간점, x_3, y_3은 획득한 최고 점수라고 했을 때, 어떤 y에서 동등화 직선인 $L_y(x)$까지 차이가 나게 되는데, 각 y에서 이 차이를 빼면 새로운 점이 생긴다. 이 점을 y^*라고 하고 다음과 같이 구한다.

$$y^* = y - L_y(x) \quad (6)$$

동등화 함수는 원호를 적용하여 산출된 값과 선형함수의 값을 합쳐서 생성한다.

$$y = y^{*} + L_{y}(x) \tag{7}$$

합성동등화 방법은 표본 동등화 방법들과 동일동등화를 사용하여 동등화하지 않는 방법 사이를 특별한 가중 체제를 이용하여 결합시키는 것이다(Kolen & Brennan, 2014). 동일동등화와 표본 동등화 방법들에 같은 가중이 주어지면 동등화 오차가 반으로 줄어들게 된다. w는 0과 1 사이 가중, x는 변환검사의 원점수, e_{Y}는 동등화 방법, ID_{Y}는 동일동등화라고 할 때 다음과 같이 산출식을 나타낼 수 있다.

$$Syn_{Y}(x) = w \times e_{Y}(x) + (1-w) \times ID_{Y}(x) \tag{8}$$

이와 같이 소규모 집단을 위한 동등화 방법에 대해 사용에 적절한 조건이나 고려해야 할 점 등을 다음과 같이 정리하였다.

〈표 1〉 소규모 집단을 위한 동등화 방법 비교

동등화 방법	장점, 효과적인 조건	단점, 주의할 점
동일 동등화	피험자가 25명 이하일 때, 검사 형이 유사할 때, 난이도 차이가 없을 때 사용, 무선 동등화 오차가 없음	검사 명세표 반드시 필요, 난이도, 내용 모두 거의 같은 경우에 유용
평균 동등화	원점수가 평균 이하일 때, 피험자 200명 이하일 때 사용, 형태가 간단함	기준검사와 변환검사가 모든 점수대에서 일정한 난이도 차이를 가지고 있을 때 유용
명목 가중 평균 동등화	피험자가 80명 이하일 때 다양한 상황에서 사용, 난이도가 쉽고 짐단간 능력 차이가 없을 때 사용	공통문항과 전체문항에서의 평균 문항반응이 유사할 것이라는 가정 필요
Tucker 선형 동등화	피험자 50~100명일 때, 원점수가 평균 근처일 때, 피험자 능력 차이가 크지 않을 때 효과적, 난이도 차이가 있을 때 동일동등화보다 적절함	공통문항에 대한 두 검사 점수의 선형 회귀가 두 검사를 치른 피험자들에게 같아야 한다는 가정 필요, 전체문항과 공통문항의 상관이 떨어지면 오차 커짐
연쇄 선형 동등화	표본 크기가 작을 때, 총점이 낮을 때 사용, 다른 선형동등화 방법들보다 단순함	연쇄 단계에서 모집단 선정의 문제

원호 동등화	표본 크기가 작을 때, 난이도 차이가 있을 때, 집단간 능력 차이가 있을 때, 원점수가 중앙값 이상일 때, 공통문항이 구성형일 때 사용, 양극단 점수에서 효과적, 선형 관계를 가정할 때의 오차를 없앨 수 있음	최저점수 선정의 문제
합성 동등화	표본 크기가 작을 때, 난이도 차이가 크지 않을 때 사용, 동등화 오차 줄일 수 있음	검사 명세표, 자료 수집 설계, 표본 크기, 피험자 집단, 가교 정보 등 다각적인 면을 고려하는 가중치 결정의 문제

Ⅲ. 연구 방법

1. 연구조건

연구의 조건으로는 문항 수, 공통문항 비율, 피험자 수, 검사 난이도 조건(난이도 수준, 검사간 난이도 차이), 피험자 능력 조건(피험자 능력 평균 차이, 피험자 능력 표준편차 차이) 등으로 7가지이다. 먼저, 문항 수는 20문항, 40문항, 60문항의 3가지 선다형 문항의 조건으로 설정하였다. 일반적으로 학교, 강의에서 사용되는 평가 시험의 경우에 20문항에서 40문항 정도가 사용되고, 국가영어능력평가시험의 듣기, 읽기 과목의 경우에는 40문항 정도였으며, 토익이나 텝스 등의 한 영역에서는 60문항 내외, 의사국가시험은 과목당 20개, 60개 등으로 문항이 구성되므로 이와 같은 다양한 문항 수에 대한 부분을 반영하여 조건을 설정하였다.

공통문항이 포함된 검사 동등화의 설계에서 바람직한 동등화 결과를 얻기 위해서는 공통문항의 역할이 매우 중요하다. 여러 가지 적절하다고 추천되는 공통문항의 조건들이 있는데 그 중 공통문항의 수는 전체문항 수의 20% 이하가 되었을 때 내용 대표성을 확보하지 못하여 동등화 오차가 커지게 된다(남현우, 2001). 또한, 공통문항의 수가 확대되면 일반적으로 동등화 오차를 크게 감소시키지만, 30% 이상으로 증가시키는 것은 큰 영향력이 없다고 보고되었다(김연정, 2009). 이에 본 연구에서는 공통문항 비율을 전체문항의 20%와 25%의 2가지 조건으로 설정하여 연구를 진행하였다.

피험자 수 조건으로는 동등화하는 변환검사의 피험자 수는 25명부터 시작하여 25명, 50명, 75명, 100명, 125명, 150명으로 25명씩 차이를 두어 6가지로 나누었다. 최소값을 25명으로 한 이유는 학교나 기관에서 한 학급 또는 반 인원이 대략 25명 전후이기 때문이고, 피험자 수를 25명씩 추가하는 것은 피험자 수를 세분하고 다양하게 하여 피험자 수에 따른 동등화 오차 크기의 변화를 살펴보고 적절한 동등화 방법을 탐구하기 위해서이다. 그리고 동등화의 기준이 되는 기준검사의 피험자 수는 100명으로 고정하였다. 선행 연구들에서는 기준검사는 이미 치러진 검사이므로 피험자 수가 더 많

다고 가정하여 피험자 수를 변환검사보다 더 많게 고정하여 연구하였다(반재천, 김선, 2015; 우중호, 2016; 임의진, 2011). 그러나 본 연구에서는 기준검사의 피험자 수를 100명으로 설정하고 변환검사보다 더 많은 경우, 같은 경우, 더 적은 경우 등의 다양한 조건으로 조합하였다.

기준검사와 변환검사의 검사 난이도와 관련하여 검사 난이도 수준은 어려운 경우('상'), 중간인 경우('중'), 쉬운 경우('하')의 3가지로 나누어 가정하였다. 검사들 중에는 중등교사 임용시험이나 공무원 선발시험, 각종 자격증시험과 같이 난이도 수준이 높은 경우가 있고, 의사, 약사, 한의사 면허시험과 같이 난이도가 쉬운 경우도 있다. 이러한 검사들의 검사 동등화 실행 상황을 고려하여 난이도 수준의 조건에 따라 안정적인 동등화 방법에 차이가 있는지 살펴보고자 하였다. 또한, 기준검사와 변환검사 두 검사 간에 난이도 차이가 있는 경우도 고려하여, 난이도 차이가 없는 경우(난이도가 같음), 조금 있는 경우, 많이 있는 경우로 설정하였다. 이를 위해 문항반응이론에 따른 문항난이도 모수 b는 정규분포에서 무선 생성하되, 검사의 난이도에 따라 난이도가 상인 경우에는 1.3, 중인 경우에 0.0, 하인 경우에는 −1.3으로 하였다. 두 검사 간에 나타날 수 있는 난이도 차이에 대한 조건은 차이가 없는 경우는 두 검사의 평균 차이가 0.0, 차이가 조금 있는 경우는 0.3, 차이가 많은 경우는 0.5만큼 나도록 하여 변환검사의 문항난이도 모수 b를 기준검사와는 다르게 설정하였다. 이 차이는 반재천과 김선(2015)의 조건을 참고하였고, 연구결과 해석의 편의를 위해 변환검사는 기준검사보다 어려운 경우만을 포함하였다. 이때 공통문항의 모수 b는 기준검사와 동일한 분포에서 무선 생성하였다.

기준검사의 피험자와 변환검사의 피험자 능력 차이에 대한 고려가 필요한데 이를 위해 피험자 능력 분포의 평균과 표준편차의 정도 값으로 조건을 설정하였다. 이전의 연구들에서는 피험자 능력의 차이를 대부분 평균의 차이로 설정하였으나, 본 연구에서는 피험자 능력의 표준편차 차이도 포함하였다. 문항반응이론에 따른 피험자 능력 모수 θ는 피험자 능력 분포의 평균 차이에서 피험자의 능력 차이가 없는 경우 0.0, 조금 있는 경우는 0.1, 많이 있는 경우는 0.25의 평균 차이가 나도록 하였다. 피험자 능력 평균 차이에서 일반적으로 평균 차이 0.1은 '차이가 있음', 0.25는 '매우 큰 차이'로 간주된다(Wang & Brennan, 2009). 피험자 능력 분포의 표준편차 차이에서는 이현숙과 김성훈(2010)의 기준을 참고하여 같은 경우 0.0('없음'), 평균을 중심으로 어느 정도 퍼져 있는 경우 0.64('보통'), 평균을 중심으로 제한된 범위에 밀집되어 있는 경우 0.25('좁음')으로 표준편차 차이가 나도록 설정하였다.

이와 같이 본 연구에서 구성한 변수는 검사 동등화의 수행에 직접 적용되어 중요한 역할을 하는 조건들이고, 동등화 방법의 결정에 신중히 고려되어야 하는 사항들이다. 또한, 본 연구에서 설정한 7개의 동등화 방법들은 선행연구에서 소규모의 피험자 집단일 경우 비교적 효과적이라고 나타난 방법들로 소규모 집단을 위한 적절한 동등화 방법을 탐구하고자 할 때 조건 변수를 다르게 하여 적용하고 비교해 볼 필요가 있는 방법들이다. 이에 소규모 집단에 대한 동등화 방법의 탐구와 관련된

선행연구들에서 구성되지 않았던 문항 수, 공통문항 비율, 기준검사와 변환검사 피험자 수의 조건, 피험자 능력의 표준편차 차이 등의 조건 변수들과 소규모 집단에게 효과적인 동등화 방법을 추가 포함시켜 소규모 집단에 대한 동등화 방법의 연구를 확장하고자 하였다. 따라서 본 연구의 조건은 〈표 2〉와 같이 문항 수 3가지, 공통문항 비율 2가지, 서로 다른 변환검사의 피험자 수 6가지와 검사 난이도 조건에서 난이도 수준 3가지와 기준검사와 변환검사 간의 차이 3가지, 피험자 능력 조건에서 피험자 능력 평균 차이 3가지와 피험자 능력 표준편차 차이 3가지로 총 2,268가지 ({3×2×6×3×3×(1+2×3)})의 경우이고, 이에 대해 7가지 동등화 방법을 적용시켜 연구를 진행하였다. 기술통계 결과 제시의 편이를 위해 난이도 조건과 피험자 능력 조건을 조건 1부터 조건 63까지 정리한 표는 〈표 3〉과 같다.

〈표 2〉 모의실험 조건

문항 수(개)	공통문항 비율(%)	피험자 수(명)	검사 난이도		피험자 능력 차이	
			기준검사(Y) 및 공통문항	변환검사 (X)	기준검사 (Y)	변환검사 (X)
20 40 60	20 25	25 50 75 100 125 150	N(1.3, 1)	N(1.3, 1) N(1.6, 1) N(1.8, 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)	N(0, 1) N(0.3, 1) N(0.5, 1)		
			N(−1.3, 1)	N(−1.3, 1) N(−1.0, 1) N(−0.8, 1)		

〈표 3〉 검사 난이도 조건 및 피험자 능력 차이 조건

조건명	검사 난이도				피험자 능력			
	기준검사		변환검사		기준검사		변환검사	
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
조건1	−1.3	1	−1.3	1	0	1	0	1
조건2	−1.3	1	−1.3	1	0	1	0.1	1
조건3	−1.3	1	−1.3	1	0	1	0.1	0.64
조건4	−1.3	1	−1.3	1	0	1	0.1	0.25
조건5	−1.3	1	−1.3	1	0	1	0.25	1
조건6	−1.3	1	−1.3	1	0	1	0.25	0.64
조건7	−1.3	1	−1.3	1	0	1	0.25	0.25
조건8	−1.3	1	−1	1	0	1	0	1
조건9	−1.3	1	−1	1	0	1	0.1	1

조건10	−1.3	1	−1	1	0	1	0.1	0.64
조건11	−1.3	1	−1	1	0	1	0.1	0.25
조건12	−1.3	1	−1	1	0	1	0.25	1
조건13	−1.3	1	−1	1	0	1	0.25	0.64
조건14	−1.3	1	−1	1	0	1	0.25	0.25
조건15	−1.3	1	−0.8	1	0	1	0	1
조건16	−1.3	1	−0.8	1	0	1	0.1	1
조건17	−1.3	1	−0.8	1	0	1	0.1	0.64
조건18	−1.3	1	−0.8	1	0	1	0.1	0.25
조건19	−1.3	1	−0.8	1	0	1	0.25	1
조건20	−1.3	1	−0.8	1	0	1	0.25	0.64
조건21	−1.3	1	−0.8	1	0	1	0.25	0.25
조건22	0	1	0	1	0	1	0	1
조건23	0	1	0	1	0	1	0.1	1
조건24	0	1	0	1	0	1	0.1	0.64
조건25	0	1	0	1	0	1	0.1	0.25
조건26	0	1	0	1	0	1	0.25	1
조건27	0	1	0	1	0	1	0.25	0.64
조건28	0	1	0	1	0	1	0.25	0.25
조건29	0	1	0.3	1	0	1	0	1
조건30	0	1	0.3	1	0	1	0.1	1
조건31	0	1	0.3	1	0	1	0.1	0.64
조건32	0	1	0.3	1	0	1	0.1	0.25
조건33	0	1	0.3	1	0	1	0.25	1
조건34	0	1	0.3	1	0	1	0.25	0.64
조건35	0	1	0.3	1	0	1	0.25	0.25
조건36	0	1	0.5	1	0	1	0	1
조건37	0	1	0.5	1	0	1	0.1	1
조건38	0	1	0.5	1	0	1	0.1	0.64
조건39	0	1	0.5	1	0	1	0.1	0.25
조건40	0	1	0.5	1	0	1	0.25	1
조건41	0	1	0.5	1	0	1	0.25	0.64
조건42	0	1	0.5	1	0	1	0.25	0.25
조건43	1.3	1	1.3	1	0	1	0	1
조건44	1.3	1	1.3	1	0	1	0.1	1
조건45	1.3	1	1.3	1	0	1	0.1	0.64
조건46	1.3	1	1.3	1	0	1	0.1	0.25
조건47	1.3	1	1.3	1	0	1	0.25	1
조건48	1.3	1	1.3	1	0	1	0.25	0.64

조건49	1.3	1	1.3	1	0	1	0.25	0.25
조건50	1.3	1	1.6	1	0	1	0	1
조건51	1.3	1	1.6	1	0	1	0.1	1
조건52	1.3	1	1.6	1	0	1	0.1	0.64
조건53	1.3	1	1.6	1	0	1	0.1	0.25
조건54	1.3	1	1.6	1	0	1	0.25	1
조건55	1.3	1	1.6	1	0	1	0.25	0.64
조건56	1.3	1	1.6	1	0	1	0.25	0.25
조건57	1.3	1	1.8	1	0	1	0	1
조건58	1.3	1	1.8	1	0	1	0.1	1
조건59	1.3	1	1.8	1	0	1	0.1	0.64
조건60	1.3	1	1.8	1	0	1	0.1	0.25
조건61	1.3	1	1.8	1	0	1	0.25	1
조건62	1.3	1	1.8	1	0	1	0.25	0.64
조건63	1.3	1	1.8	1	0	1	0.25	0.25

2. 연구절차

본 연구를 위해 연구의 조건 7가지에서 3모수 문항반응이론에 따라 변별도 모수 a는 Lognormal(0, 0.3), 추측도 모수 c는 Uniform(0.05, 0.35)로 동일하도록 하고, 문항난이도 모수는 이 연구에서 설정한 난이도 조건에 따라 정규분포에서 10,000명의 모의인을 무선 생성하여 각 조건에서 피험자 수를 대치적 표본방법으로 재표집(resampling)하였다. 재표집은 각 피험자 수마다 500회를 반복 표집하여 데이터 세트를 만들었다. 재표집은 실험의 신뢰를 높이기 위하여 실시되는데 선행연구에서는 100회(안수현, 2017; Aşiret & Sünbül, 2016)부터 300회(이현숙, 김성훈, 2010), 1,000회(반재천, 김선, 2015; 임의진, 이규민, 2017) 등으로 실시하였다. 하지만 1,000회를 실시했던 연구에서는 동등화 방법이 한 가지만 적용되었거나 문항 수, 공통문항 비율, 검사난이도와 피험자 조건 등의 연구조건이 다소 간단하였다.

데이터 세트 생성 후에는 동일동등화 방법, 평균동등화 방법, 명목 가중 평균동등화 방법, Tucker 선형동등화 방법, 연쇄 선형동등화 방법, 원호동등화 방법, 합성동등화 방법 등의 7가지 검사 동등화를 수행하였다. 동일동등화는 동등화하지 않은 것과 같은 값을 나타내는데 동등화를 실시할 때 최소의 표본 크기를 설정하여 표본 크기가 설정된 값에 미치지 못하면 동등화를 실시하지 않는 것이다(Kolen & Brennan, 2014). 평균동등화는 기준검사와 변환검사의 난이도가 전반적인 점수대에서 일정한 차이로 나타난다고 가정하여, 두 검사의 평균 차이를 이용하고 두 검사를 모두 본 합성 집단에 대해 동등식을 만든 것이다. 명목 가중 평균동등화는 평균동등화 방법의 가정을 그대로 사용하여 기준검사와 변환검사간의 난이도 차이를 모든 점수대에서 동일하게 조정하고, Tucker 선형동등화에

서 회귀계수대신 공통문항 수와 전체문항 수의 비율로 대체하여 사용한 것이다(Babcock et al., 2012). Tucker 선형동등화 방법은 두 모집단 능력의 차이를 교정하기 위해 공통문항 점수의 평균과 표준편차의 차이를 이용하는 것이다(남현우, 2001). 연쇄 선형동등화는 선형동등화 방법의 산출 식에서 Tucker 선형 방법과는 달리 공통문항을 제외한 전체점수와 공통문항간의 관계식을 포함시키는 방법이다(Albano, 2016). 원호동등화는 평균동등화와 선형동등화에서 선형 관계를 가정할 때 나타나는 오차를 없애기 위한 비선형 방법으로 각 검사 형의 최저, 최고, 중간점의 동등화된 점수를 연결하여 원호 형태를 갖는 동등화 곡선을 추정하는 것이다. 합성동등화는 동등화를 실시하지 않은 값과 동등화를 실시한 값을 연결하여 가중 평균을 이용하는 방법으로, 두 개 이상의 동등화 방법 결과의 평균을 구하는 것이다(Kim et al., 2008). 본 연구에서 합성동등화는 임의진과 이규민(2017), Kim 외(2008)의 연구를 참고하여 동일동등화와 연쇄 선형동등화를 합성하였다.

그리고 각 조건 하에서 10,000명의 모의생성자료에 빈도추정 동백분위동등화 방법을 적용하여 얻어진 동등화된 점수를 기준값으로 간주하였다. 빈도추정 동백분위동등화 방법은 모든 점수대에서 동등화의 정확성이 요구될 때 사용하는 방법(Kolen & Brennan, 2014)으로 본 연구에서도 모든 점수대에서 동등화의 정확성을 고려하고자 하여 이 방법을 선택하였다. 연구의 전 과정은 R(3.3.2)을 이용하여 수행되었는데, 분산분석은 aov 함수를 사용하고 동등화를 위해서는 equate package(Albano, 2016)를 이용하였다.

3. 평가준거

본 연구에서 어떤 동등화 방법이 각 조건에서 효과적이고 안정적인지를 비교하기 위해 동등화 오차의 값을 산출하였다. 이를 위해 동등화된 점수 각각의 동등화 오차인, 무선오차와 체계적 오차를 포함하는 평균제곱오차의 제곱근(RMSE)을 신출하였는데, 비교를 더 효과적으로 하기 위해 전체점수를 포함하는 종합적인 오차인 가중된 평균제곱오차의 제곱근(WRMSE)을 이용하였다. WRMSE를 산출하기 위해서는 먼저 RMSE를 계산하여야 하는데 그 식은 다음과 같다.

$$RMSE(x_i) = \sqrt{BIAS^2(x_i) + SE^2(x_i)} \tag{1}$$

$RMSE(x_i)$의 계산을 위한 $Bias(x_i)$와 $SE(x_i)$는 다음과 같이 산출된다.

$$Bias(x_i) = Mean(\widehat{eq_r}(x_i)) - eq(x_i) \tag{2}$$

$$SE(x_i) = \sqrt{\sum_{r=1}^{R} \left[\widehat{eq_r}(x_i) - Mean(\widehat{eq_r}(x_i) \right]^2 / R} \quad (3)$$

여기에서 $\widehat{eq_r}(x_i)$는 r회 대치로 얻어지는 각 원점수 x_i에서의 동등화 값, $eq(x_i)$는 모집단에서의 기준 동등화 값이며, R은 재표집 수로 본 연구에서는 500이다.

다음은 동등화 방법의 수월한 비교를 위한 $WRMSE$의 산출식이다.

$$WRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{I} n_i RMSE(x_i)^2} \quad (4)$$

이때 n은 변환검사를 본 전체자료세트 총 피험자 수이고, I는 총 문항의 수이며, n_i는 변환검사 x_i를 가지는 전체자료세트에서 피험자 수이다.

Ⅳ. 연구 결과

본 연구의 목적은 설정된 다양한 조건에 대해 동등화 방법을 비교하여 더 적절한 동등화 방법을 제시하는 것에 있다. 모의자료의 기술통계치는 〈표 4〉, 〈표 5〉, 〈표 6〉과 같고, 동등화 오차인 WRMSE를 종속변수로 두는 분산분석의 결과는 〈표 7〉과 같다. 지면 제약상 분산분석의 주효과 분석은 표를 제외하고 정리하였으며, 분산분석에서 유의미하게 나타난 조건들에 대한 WRMSE를 표로 제시하였다. 전체 점수를 포함하는 종합적인 동등화 오차인 WRMSE가 낮게 나타날수록 안정적인 동등화 방법이라고 판단하였다.

〈표 4〉 문항 수 20개인 경우 모집단 10,000명의 기술통계치

조건명	공통문항 비율 20%				공통문항 비율 25%			
	기준검사		변환검사		기준검사		변환검사	
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
조건1	11.87	2.69	11.92	2.69	11.59	2.39	11.58	2.41
조건2	11.89	2.68	12.52	2.35	11.59	2.38	11.58	2.31
조건3	11.87	2.71	12.41	1.87	11.57	2.38	12.00	1.81
조건4	11.89	2.66	12.89	1.61	11.56	2.39	12.82	1.37
조건5	11.89	2.69	12.21	2.57	11.56	2.39	11.88	2.15
조건6	11.87	2.72	12.87	1.88	11.56	2.36	11.37	2.02

조건7	11.88	2.68	12.98	1.55	11.57	2.38	12.37	1.46
조건8	11.87	2.69	13.02	2.40	11.59	2.37	10.94	2.40
조건9	11.88	2.69	11.67	2.40	11.56	2.39	11.47	2.35
조건10	11.88	2.69	11.62	2.25	11.62	2.35	12.06	1.87
조건11	11.89	2.69	12.70	1.56	11.57	2.40	11.45	1.66
조건12	11.88	2.70	12.11	2.54	11.58	2.39	11.42	2.34
조건13	11.90	2.68	11.43	2.17	11.57	2.40	11.88	1.88
조건14	11.89	2.68	12.22	1.71	11.58	2.41	12.19	1.54
조건15	11.88	2.70	12.01	2.86	11.58	2.39	9.99	2.50
조건16	11.88	2.68	11.48	2.55	11.61	2.37	10.88	2.19
조건17	11.88	2.69	11.70	2.28	11.61	2.39	10.03	2.20
조건18	11.89	2.68	11.53	1.80	11.59	2.37	11.34	1.69
조건19	11.87	2.70	11.88	2.68	11.59	2.39	12.27	2.03
조건20	11.89	2.69	12.75	2.03	11.68	2.38	11.42	2.08
조건21	11.90	2.70	12.14	1.70	11.58	2.38	11.54	1.68
조건22	8.44	3.17	8.43	3.18	9.80	2.55	9.80	2.55
조건23	8.48	3.17	10.30	2.89	9.81	2.55	9.21	1.91
조건24	8.47	3.18	9.63	2.47	9.79	2.54	8.97	1.89
조건25	8.48	3.18	9.82	1.95	9.81	2.54	9.01	1.88
조건26	8.45	3.19	11.30	2.90	9.78	2.55	9.32	1.92
조건27	8.46	3.16	9.09	2.24	9.75	2.57	8.86	1.93
조건28	9.46	3.20	9.72	1.95	9.81	2.55	8.86	1.88
조건29	8.47	3.18	10.63	2.71	9.79	2.56	9.09	1.91
조건30	8.46	3.18	8.91	3.00	9.78	2.52	8.76	1.94
조건31	8.45	3.17	9.34	2.37	9.78	2.55	8.61	1.92
조건32	8.48	3.17	8.80	2.03	9.82	2.57	9.19	1.88
조건33	8.45	3.16	10.35	2.91	9.80	2.56	9.04	1.93
조건34	8.48	3.18	9.38	2.48	9.79	2.56	9.27	1.91
조건35	8.45	3.16	10.05	2.01	9.80	2.53	8.95	1.88
조건36	8.47	3.20	9.70	2.94	9.80	2.56	9.22	1.91
조건37	8.49	3.17	8.90	2.95	9.81	2.53	8.95	1.94
조건38	8.44	3.21	8.43	2.34	9.82	2.56	8.77	1.90
조건39	8.48	3.18	8.29	1.96	9.80	2.56	8.89	1.90
조건40	8.45	3.18	10.24	2.74	9.78	2.56	8.76	1.94
조건41	8.46	3.18	10.27	2.33	9.78	2.57	8.92	1.88
조건42	8.48	3.16	9.37	1.96	9.77	2.55	8.53	1.89
조건43	7.03	2.62	7.03	2.60	8.38	1.91	8.35	1.91
조건44	7.00	2.63	6.94	2.54	8.34	1.91	8.42	1.97
조건45	6.99	2.60	6.46	2.28	8.36	1.91	8.45	1.95

조건46	7.03	2.64	6.03	1.92	8.37	1.94	8.44	1.92
조건47	7.03	2.61	7.45	2.67	8.34	1.93	8.48	1.93
조건48	7.04	2.60	6.15	2.24	8.36	1.92	8.69	1.91
조건49	7.01	2.63	6.96	1.87	8.33	1.93	8.42	1.93
조건50	7.05	2.63	6.68	2.61	8.31	1.92	8.24	1.95
조건51	7.02	2.60	6.09	2.69	8.39	1.91	8.05	1.95
조건52	7.03	2.64	6.54	2.25	8.34	1.94	8.48	1.93
조건53	7.01	2.61	6.28	1.92	8.36	1.93	8.27	1.91
조건54	6.99	2.60	6.16	2.42	8.36	1.95	8.26	1.95
조건55	7.04	2.64	5.30	2.11	8.34	1.92	8.73	1.91
조건56	7.03	2.62	6.69	1.97	8.33	1.93	8.07	1.91
조건57	7.01	2.62	5.38	2.38	8.34	1.95	8.43	1.92
조건58	7.00	2.61	6.46	2.41	8.34	1.93	8.10	1.98
조건59	7.05	2.63	5.18	2.02	8.37	1.92	8.15	1.94
조건60	7.01	2.62	4.95	1.82	8.35	1.94	8.19	1.91
조건61	7.03	2.61	6.20	2.54	8.37	1.95	8.17	1.96
조건62	7.03	2.63	6.82	2.30	8.34	1.91	8.36	1.93
조건63	7.05	2.64	5.89	1.93	8.36	1.94	8.10	1.91

〈표 5〉 문항 수 40개인 경우 모집단 10,000명의 기술통계치

조건명	공통문항 비율 20%				공통문항 비율 25%			
	기준검사		변환검사		기준검사		변환검사	
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
조건1	23.53	4.82	23.50	4.85	23.06	4.40	23.12	4.40
조건2	23.54	4.83	24.34	4.64	23.08	4.43	23.79	4.40
조건3	23.47	4.82	26.18	3.12	23.09	4.40	24.61	3.17
조건4	23.49	4.82	26.28	2.31	23.08	4.42	24.59	2.21
조건5	23.49	4.84	25.80	4.25	23.07	4.41	24.32	3.57
조건6	23.49	4.86	26.16	2.90	23.10	4.40	24.91	2.88
조건7	23.49	4.85	26.33	2.20	23.06	4.41	25.12	2.10
조건8	23.49	4.85	23.89	4.84	23.10	4.44	23.27	4.30
조건9	23.47	4.86	24.75	4.44	23.10	4.42	22.84	4.40
조건10	23.50	4.87	24.18	3.64	23.11	4.40	23.51	3.01
조건11	23.54	4.85	23.38	2.56	23.10	4.42	23.55	2.35
조건12	23.51	4.85	24.45	4.64	23.08	4.44	23.96	3.84
조건13	23.55	4.82	24.64	3.49	23.08	4.40	22.87	3.49
조건14	23.51	4.85	25.73	2.32	23.07	4.43	23.91	2.25
조건15	23.51	4.86	22.42	5.08	23.06	4.41	21.25	4.35
조건16	23.52	4.86	23.19	4.87	23.07	4.43	22.77	4.44

조건17	23.48	4.85	24.07	3.76	23.09	4.39	22.15	3.17
조건18	23.52	4.84	24.77	2.49	23.05	4.43	21.74	2.49
조건19	23.51	4.82	24.85	4.51	23.06	4.43	22.29	4.65
조건20	23.51	4.85	21.73	3.99	23.08	4.40	23.81	3.08
조건21	23.51	4.88	24.55	2.45	23.08	4.41	22.90	2.41
조건22	19.68	5.11	19.71	5.08	17.63	5.16	17.62	5.15
조건23	19.69	5.09	18.49	5.25	17.65	5.17	19.17	4.60
조건24	19.67	5.11	18.92	3.89	17.61	5.18	19.78	3.71
조건25	19.69	5.09	19.58	2.82	17.62	5.17	18.53	2.67
조건26	19.71	5.07	21.48	4.95	17.65	5.17	17.68	5.33
조건27	19.70	5.10	20.36	4.07	17.63	5.16	17.87	4.18
조건28	19.67	5.06	21.80	2.73	17.66	5.17	19.73	2.68
조건29	19.73	5.05	18.86	5.55	17.65	5.16	17.24	4.93
조건30	19.69	5.09	17.29	5.10	17.62	5.18	15.73	5.28
조건31	19.72	5.10	18.51	4.00	17.60	5.16	17.22	4.15
조건32	19.67	5.06	19.64	2.89	17.63	5.19	18.21	2.72
조건33	19.72	5.09	19.54	5.67	17.65	5.16	18.64	4.79
조건34	19.70	5.08	18.14	4.24	17.61	5.15	19.17	3.66
조건35	19.68	5.06	18.57	2.96	17.65	5.20	18.53	2.80
조건36	19.70	5.06	15.52	5.31	17.64	5.13	15.43	5.20
조건37	19.68	5.07	17.62	4.88	17.64	5.21	15.87	4.90
조건38	19.66	5.10	17.42	3.84	17.67	5.18	15.60	4.06
조건39	19.70	5.05	17.55	2.91	17.69	5.17	15.21	2.83
조건40	19.69	5.07	18.70	5.24	17.63	5.20	17.05	4.86
조건41	19.67	5.06	19.11	3.95	17.65	5.20	16.51	4.05
조건42	19.70	5.07	18.05	2.93	17.59	5.18	15.25	2.81
조건43	12.79	4.45	12.78	4.46	12.93	4.28	12.95	4.28
조건44	12.81	4.45	14.99	5.09	13.00	4.30	12.33	4.28
조건45	12.82	4.48	13.28	3.69	12.92	4.25	12.49	3.63
조건46	12.74	4.49	12.40	2.67	12.94	4.31	11.47	2.69
조건47	12.75	4.47	14.27	4.69	12.98	4.27	13.87	4.77
조건48	12.76	4.43	14.97	3.71	12.99	4.28	12.67	3.72
조건49	12.77	4.48	12.97	2.79	12.94	4.28	14.00	2.74
조건50	12.78	4.46	11.94	4.35	12.96	4.30	12.06	4.28
조건51	12.80	4.46	13.08	4.53	12.97	4.26	11.87	4.51
조건52	12.74	4.45	13.20	3.49	12.92	4.27	11.00	3.29
조건53	12.77	4.48	13.96	2.84	12.92	4.29	11.43	2.68
조건54	12.77	4.45	12.66	4.48	13.00	4.28	13.52	4.74
조건55	12.80	4.48	13.48	3.74	12.91	4.28	13.71	3.51

조건56	12.81	4.42	12.33	2.73	12.95	4.28	13.64	2.73
조건57	12.76	4.44	11.78	4.01	12.95	4.30	10.09	3.71
조건58	12.77	4.45	12.42	4.28	12.96	4.30	10.92	3.87
조건59	12.75	4.45	11.67	3.32	12.87	4.28	11.11	3.43
조건60	12.71	4.46	12.87	2.75	12.97	4.27	11.40	2.60
조건61	12.80	4.44	12.40	4.35	12.97	4.25	11.38	4.27
조건62	12.75	4.46	11.53	3.59	12.93	4.28	11.51	3.56
조건63	12.78	4.45	13.40	2.75	12.93	4.30	11.22	2.61

〈표 6〉 문항 수 60개인 경우 모집단 10,000명의 기술통계치

조건명	공통문항 비율 20%				공통문항 비율 25%			
	기준검사		변환검사		기준검사		변환검사	
	평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
조건1	35.91	6.39	35.93	6.38	36.30	5.70	36.29	5.73
조건2	35.92	6.38	40.03	5.11	36.28	5.77	33.71	6.46
조건3	35.91	6.40	38.27	4.49	36.30	5.67	36.78	3.83
조건4	35.95	6.41	38.68	2.92	36.30	5.69	37.02	2.77
조건5	35.95	6.36	38.70	6.13	36.28	5.76	35.45	5.55
조건6	35.94	6.36	39.37	4.04	36.30	5.68	36.63	4.03
조건7	35.94	6.38	39.29	2.89	36.24	5.73	37.62	2.68
조건8	35.95	6.34	36.80	6.01	36.32	5.71	34.23	6.06
조건9	35.93	6.36	36.83	6.80	36.34	5.68	35.11	6.50
조건10	35.92	6.40	36.11	5.19	36.31	5.73	34.14	4.48
조건11	35.92	6.32	38.18	2.96	36.30	5.70	35.81	2.95
조건12	35.95	6.40	37.40	6.43	36.27	5.70	33.19	6.52
조건13	35.97	6.38	39.05	4.40	36.27	5.72	35.00	4.58
조건14	35.95	6.35	36.39	3.23	36.26	5.72	35.49	2.99
조건15	35.95	6.43	33.97	7.00	36.27	5.71	31.59	6.86
조건16	35.94	6.40	35.45	6.94	36.33	5.73	33.55	6.37
조건17	35.95	6.38	37.42	4.99	36.31	5.70	33.47	4.94
조건18	35.96	6.37	34.87	3.49	36.29	5.69	34.46	3.09
조건19	35.98	6.36	37.95	6.25	36.28	5.69	34.16	6.35
조건20	35.91	6.38	34.32	4.73	36.30	5.71	34.54	4.69
조건21	35.95	6.39	37.70	2.91	36.29	5.71	34.65	2.99
조건22	29.10	8.00	29.09	8.04	27.18	6.91	27.25	6.90
조건23	29.15	7.98	29.60	7.78	27.24	6.96	27.03	7.01
조건24	29.13	8.02	29.18	5.77	27.25	6.94	26.71	5.45
조건25	29.08	7.98	29.30	3.70	27.29	6.91	28.26	3.63
조건26	29.12	8.00	30.87	7.79	27.21	6.92	29.74	6.87

조건27	29.15	8.01	31.43	5.32	27.27	6.86	29.42	5.19
조건28	29.10	8.04	30.81	3.77	27.26	6.96	28.05	3.64
조건29	29.08	7.99	25.85	7.64	27.18	6.89	25.55	7.10
조건30	29.12	7.96	28.63	7.88	27.17	6.94	24.80	7.20
조건31	29.09	8.04	27.14	6.06	27.23	6.91	24.79	5.53
조건32	29.12	8.05	26.26	3.79	27.21	6.89	25.33	3.58
조건33	29.12	7.96	28.91	7.71	27.21	6.95	27.22	7.28
조건34	29.08	8.04	29.28	5.92	27.27	6.89	27.30	5.54
조건35	29.13	8.00	30.21	3.78	27.22	6.90	27.63	3.61
조건36	29.08	7.99	25.58	7.54	27.24	6.94	22.30	6.99
조건37	29.10	8.01	26.55	7.97	27.21	6.93	24.48	7.44
조건38	29.11	8.04	23.94	5.79	27.25	6.95	24.58	5.25
조건39	29.10	8.03	24.48	3.78	27.23	6.93	25.28	3.58
조건40	29.11	8.02	26.79	7.68	27.23	6.89	24.67	6.99
조건41	29.09	8.00	27.00	5.56	27.22	6.95	24.36	5.50
조건42	29.09	8.02	24.72	3.86	27.27	6.91	23.50	3.66
조건43	18.61	6.77	18.61	6.78	18.22	6.13	18.25	6.14
조건44	18.56	6.76	17.59	6.46	18.24	6.08	20.38	6.61
조건45	18.57	6.76	21.35	5.36	18.24	6.07	18.05	4.57
조건46	18.61	6.82	20.41	3.60	18.21	6.08	17.95	3.34
조건47	18.61	6.80	21.40	6.99	18.20	6.11	20.45	6.46
조건48	18.60	6.79	22.04	5.70	18.21	6.12	20.07	5.01
조건49	18.53	6.80	21.59	3.80	18.26	6.09	20.51	3.58
조건50	18.66	6.74	18.32	6.24	18.26	6.13	18.34	5.92
조건51	18.61	6.78	17.37	5.95	18.18	6.11	16.92	5.61
조건52	18.59	6.84	17.79	4.90	18.24	6.10	17.71	4.71
조건53	18.58	6.79	17.97	3.45	18.25	6.07	17.33	3.30
조건54	18.60	6.73	19.30	6.59	18.18	6.11	19.29	6.49
조건55	18.65	6.83	18.05	4.69	18.23	6.13	17.58	4.38
조건56	18.58	6.83	19.81	3.55	18.26	6.07	18.32	3.53
조건57	18.65	6.79	19.37	5.78	18.25	6.08	18.58	6.34
조건58	18.61	6.75	19.15	6.02	18.19	6.15	15.05	5.11
조건59	18.62	6.85	17.40	4.62	18.20	6.09	17.07	4.36
조건60	18.58	6.84	18.22	3.44	18.23	6.05	17.17	3.38
조건61	18.67	6.80	20.25	6.32	18.19	6.11	17.74	5.71
조건62	18.59	6.82	19.24	4.98	18.23	6.05	18.21	4.90
조건63	18.59	6.80	17.61	3.45	18.23	6.06	17.56	3.39

〈표 7〉 동등화 방법과 각 조건에 따른 WRMSE의 분산분석

구분	Df	Sum Sq	Mean Sq	F value	Pr(>F)
동등화 방법(w)	6	2392.91	398.82	36028.00	<.0001
문항 수(i)	2	611.45	305.72	13368.60	<.0001
공통문항 비율(a)	1	4.65	4.65	203.54	<.0001
피험자 수(n)	5	141.05	28.21	1233.54	<.0001
난이도 수준(l)	2	35.14	17.57	768.37	<.0001
난이도 차이(d)	2	78.32	39.16	1712.44	<.0001
피험자 평균 차이(m)	2	8.54	4.27	186.68	<.0001
피험자 표준편차 차이(s)	2	2.13	1.06	46.52	<.0001
w × i	12	393.40	32.78	2961.52	<.0001
w × a	6	3.78	.63	56.92	<.0001
w × n	30	114.56	3.82	344.95	<.0001
w × l	12	39.89	3.32	300.30	<.0001
w × d	12	243.65	20.30	1834.24	<.0001
w × m	12	25.39	2.12	191.13	<.0001
w × s	12	5.69	.47	42.80	<.0001
w × i × a	12	.80	.07	6.03	<.0001
w × i × n	60	28.09	.47	42.29	<.0001
w × i × l	24	12.39	.52	46.64	<.0001
w × i × d	24	44.35	1.85	166.92	<.0001
w × i × m	24	4.29	.18	16.16	<.0001
w × i × s	24	1.04	.04	3.91	<.0001
w × a × n	30	.41	.01	1.23	.1799
w × a × l	12	.07	.01	.51	.9110
w × a × d	12	.27	.02	2.07	.0158
w × a × m	12	.04	.00	.28	.9922
w × a × s	12	.03	.00	.19	.9988
w × n × l	60	2.86	.05	4.30	<.0001
w × n × d	60	.17	.00	.25	1.0000
w × n × m	60	.06	.00	.09	1.0000
w × n × s	60	.03	.00	.05	1.0000
w × l × d	24	53.60	2.23	2017.76	<.0001
w × l × m	24	1.85	.08	6.96	<.0001
w × l × s	24	5.26	.22	19.81	<.0001
w × d × m	24	190.93	7.96	718.65	<.0001
w × d × s	24	.27	.01	1.03	.4230
w × m × s	12	.06	.00	.48	.9252
오차	12888	142.67	.01		

조건의 주효과 모두 유의미한 것으로 나타났으므로 사후분석을 실시하였는데, Tukey 사후분석 결과 동등화 방법 중에서는 원호동등화와 평균동등화 방법이 WRMSE가 작았고, 그 다음으로 Tucker 선형동등화, 합성동등화, 연쇄 선형동등화의 순이었다. 대부분의 조건에서 원호동등화와 평균동등화가 낮은 동등화 오차를 나타내는 것은 Aşiret과 Sünbül(2016)의 결과와 동일하였지만, 임의진과 이규민(2017)의 연구에서는 동일동등화와 합성동등화가 효과적인 것으로 나타나 차이가 있었다. 명목 가중 평균동등화 방법은 좋지 않은 수행을 보였는데, Babcock 외(2012)의 결과와는 차이가 있는 것이었고, 임의진과 이규민(2017)의 연구와는 유사한 결과였다. 이는 각 연구에서 자료의 형태, 문항 수, 문항 종류, 공통문항 비율, 검사 난이도, 피험자 능력 등의 조건이 다르기 때문인 것으로 보이는데, 관련된 연구가 더 필요하다. 문항 수 조건에서는 문항 수가 증가할수록 WRMSE 값이 커짐을 알 수 있었다. 이는 명목 가중 평균동등화의 경우에 국한되기는 하지만 반재천, 김선(2015)의 연구와 유사한 결과였다. 그리고 피험자 수가 증가하면 WRMSE는 감소하는 것으로 나타났는데, 이는 우중호(2016), Parshall 외(1995)의 연구결과와 유사하다. 두 검사의 난이도 수준과 검사간 난이도 차이에 대한 사후분석 결과는 기준검사와 변환검사의 난이도가 쉬운 경우에 WRMSE가 가장 작았고, 그 다음은 어려운 경우와 보통의 경우로 나타났다. 반재천과 김선(2015)의 연구에서도 명목 가중 평균동등화 방법에 국한되기는 하나 난이도가 쉬운 경우에 동등화 오차가 작아졌다. 두 검사간 난이도 차이는 없거나 조금 있을 경우에 WRMSE가 작게, 난이도 차이가 크면 WRMSE 값이 크게 나타났는데, 이는 Kolen과 Brennan(2004)의 결과와 유사하다. 피험자 능력 평균 차이 조건에서는 차이가 많은 경우에 WRMSE가 가장 작았고, 그 다음으로는 조금 있을 때, 차이가 없을 때의 순이었다. 우중호(2016)의 결과에서도 원호동등화 방법에 국한되기는 하나, 피험자 능력 차이가 클수록 동등화 오차가 감소하였다. 피험자 능력 표준편차 차이는 평균 주변으로 어느 정도 떨어져 있는 보통일 때가 가장 오차가 작았고, 그 다음은 차이가 없을 때, 마지막으로 평균 주변으로 밀집되어 있을 때로 나타났다.

동등화 방법과 조건들의 3개의 상호작용을 살펴보면, (1) 동등화 방법, 문항 수, 공통문항 비율의 상호작용, (2) 동등화 방법, 문항 수, 피험자 수의 상호작용, (3) 동등화 방법, 문항 수, 두 검사의 난이도 수준의 상호작용, (4) 동등화 방법, 문항 수, 두 검사간 난이도 차이의 상호작용, (5) 동등화 방법, 문항의 수, 피험자 능력 평균 차이의 상호작용, (6) 동등화 방법, 문항 수, 피험자 능력 표준편차 차이의 상호작용, (7) 동등화 방법, 공통문항 비율, 두 검사간 난이도 차이의 상호작용, (8) 동등화 방법, 두 검사의 난이도 수준, 피험자 수의 상호작용, (9) 동등화 방법, 두 검사의 난이도 수준, 두 검사간 난이도 차이의 상호작용, (10) 동등화 방법, 두 검사의 난이도 수준, 피험자 능력 평균 차이의 상호작용, (11) 동등화 방법, 두 검사의 난이도 수준, 피험자 능력 표준편차 차이의 상호작용, (12) 동등화 방법, 두 검사간 난이도 차이와 피험자 능력 평균 차이의 상호작용 등이 유의미하였다. 이는 해당 조건에서 WRMSE 값이 의미 있게 차이가 있으므로 그 차이를 해석하는 것이 가능함을 나타낸

다. 동등화 방법과 각 조건들 간의 2개 상호작용도 의미가 있는 것으로 나타났으나, 3개의 유의미한 상호작용 12개에 모두 포함되므로 별도 해석은 하지 않았다. 유의미하게 나타난 각 조건에서 7가지 동등화 방법에 대한 WRMSE 결과를 정리하면 다음과 같다.

〈표 8〉 (1) 문항 수, 공통문항 비율의 동등화 방법별 WRMSE

문항 수	공통문항 비율	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	20%	.478	.306	.979	.377	.508	**.301**	.367
	25%	.449	.296	.903	.365	.493	**.292**	.351
40개	20%	.946	.433	1.816	.535	.722	**.429**	.628
	25%	.887	.418	1.690	.515	.697	**.414**	.594
60개	20%	1.422	.530	2.647	.654	.882	**.528**	.880
	25%	1.330	.513	2.466	.633	.855	**.511**	.831

(1) 문항 수, 공통문항 비율 조건에 따른 동등화 방법의 WRMSE 값은 〈표 8〉과 같다. 모든 문항 수와 공통문항 비율 조건에서 원호동등화 방법이 가장 적절한 것으로 나타났으며, 평균동등화 방법도 원호동등화 방법과 유사하게 대체로 좋은 동등화 수행을 보였다. 그 다음으로는 Tucker 선형동등화, 합성동등화 방법이 효과적이었는데, 문항 수가 20개일 때에는 합성동등화 방법이, 문항 수가 40개, 60개일 때에는 Tucker 선형동등화 방법이 비교적 작은 WRMSE를 가졌다.

〈표 9〉 (2) 문항 수, 피험자 수의 동등화 방법별 WRMSE

문항 수	피험자 수	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	25	.464	.416	1.322	.538	.717	**.409**	.441
	50	.460	.326	1.026	.404	.546	**.320**	.372
	75	.464	.290	.904	.353	.478	**.286**	.351
	100	.463	.269	.839	.325	.442	**.265**	.338
	125	.464	.257	.791	.309	.418	**.254**	.329
	150	.463	.248	.764	.297	.403	**.245**	.324
40개	25	.920	.587	2.466	.762	1.015	**.583**	.714
	50	.923	.463	1.913	.573	.776	**.460**	.634
	75	.909	.411	1.682	.500	.677	**.406**	.595
	100	.915	.380	1.553	.461	.626	**.376**	.582
	125	.915	.363	1.474	.437	.593	**.360**	.573
	150	.917	.348	1.429	.418	.570	**.345**	.568
60개	25	1.374	.722	3.593	.935	1.241	**.720**	.966
	50	1.377	.564	2.778	.699	.944	**.560**	.879
	75	1.376	.503	2.458	.612	.831	**.499**	.842
	100	1.372	.468	2.269	.565	.766	**.466**	.823
	125	1.376	.445	2.162	.536	.729	**.444**	.814
	150	1.378	.428	2.078	.514	.700	**.426**	.808

(2) 문항 수, 피험자 수에 따른 각 동등화 방법의 WRMSE 값은 위의 〈표 9〉와 같이 나타났다. 이 조건에서도 역시 원호동등화 방법이 가장 효과적이었고 평균동등화 방법도 대체로 안정적으로 나타났다. Tucker 선형동등화와 합성동등화 방법도 비교적 낮은 WRMSE를 가졌는데, 문항 수가 20개이고 피험자 수가 25명일 경우에는 동등화하지 않는 동일동등화 방법도 좋은 수행을 보였다. 문항 수가 20개이고 피험자 수가 25명일 경우에 합성동등화, 동일동등화 방법이 비교적 좋은 수행을 보이는 것은 Kim 외(2008), Skaggs(2005)의 동일동등화 결과와 유사한 결과이다.

〈표 10〉 (3) 문항 수, 검사 난이도 수준의 동등화 방법별 WRMSE

문항 수	난이도 수준	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	상	.458	.312	1.032	.391	.512	**.309**	.362
	중	.517	.314	.982	.389	.512	**.302**	.384
	하	.415	**.276**	.810	.333	.477	.278	.332
40개	상	.908	**.441**	1.948	.554	.727	**.441**	.616
	중	1.025	.442	1.830	.549	.726	**.432**	.659
	하	.817	.394	1.480	.472	.675	**.392**	.558
60개	상	1.365	**.542**	2.861	.680	.893	.546	.863
	중	1.542	.543	2.657	.674	.888	**.534**	.931
	하	1.220	.481	2.151	.576	.825	**.478**	.773

(3) 문항 수, 두 검사의 난이도 수준에 따른 각 동등화 방법의 WRMSE 값은 위의 〈표 10〉과 같다. 먼저, 문항 수 20개인 경우에는 두 검사의 난이도 수준이 어렵거나 보통일 때 원호동등화, 난이도가 쉬울 때에는 평균동등화 방법이 가장 낮은 WRMSE를 나타냈다. 문항 수 40개인 경우에는 검사 난이도의 모든 조건에서 원호동등화가 효과적인 것으로 나타났는데, 난이도가 어려울 때에는 원호동등화와 함께 평균동등화도 좋은 수행을 보였다. 문항 수가 60개일 때에는 검사가 어려운 경우에는 평균동등화가, 검사가 보통이거나 쉬울 때에는 원호동등화 방법이 적절하였다. Tucker 선형동등화와 합성동등화 방법도 비교적 낮은 WRMSE를 가졌다.

〈표 11〉 (4) 문항 수, 검사간 난이도 차이의 동등화 방법별 WRMSE

문항 수	난이도 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	없음	.328	**.295**	.938	.371	.501	.305	.312
	조금	.340	.299	.942	.370	.500	**.288**	.315
	많이	.722	.309	.943	.371	.501	**.296**	.451
40개	없음	.645	**.420**	1.754	.524	.710	.429	.508
	조금	.672	.423	1.757	.525	.711	**.415**	.512
	많이	1.432	.433	1.747	.525	.708	**.421**	.813
60개	없음	.973	**.516**	2.554	.643	.867	.528	.696
	조금	1.008	.518	2.559	.644	.870	**.510**	.699
	많이	2.146	.532	2.556	.644	.869	**.520**	1.172

(4) 문항 수, 두 검사간 난이도 차이에 따른 각 동등화 방법의 WRMSE 값은 위의 〈표 11〉과 같다. 문항 수 조건에 상관없이 두 검사간의 난이도 차이가 없으면 평균동등화가, 난이도 차이가 조금 있거나 많이 있으면 원호동등화 방법이 효과적인 것으로 나타났다. 두 동등화 방법 외에도 Tucker 선형동등화와 합성동등화 방법이 낮은 WRMSE를 가졌으며, 문항 수 20개이고 검사간 난이도 차이가 없거나 조금 있을 때에는 동일동등화 방법도 효과적이었다.

다음의 〈표 12〉는 (5) 문항 수, 피험자 집단의 평균 차이에 따른 각 동등화 방법의 WRMSE 값을 나타낸다. 문항 수 20개, 40개, 60개에서 피험자 능력 평균 차이가 없거나 조금 차이가 있으면 원호동등화가, 능력 평균 차이가 많이 나면 평균동등화가 효과적이었다. 그 다음으로는 Tucker 선형동등화와 합성동등화 방법이 좋은 수행을 보였다.

〈표 12〉 (5) 문항 수, 피험자 능력 평균 차이의 동등화 방법별 WRMSE

문항 수	피험자 능력 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	없음	.537	.308	.941	.370	.499	**.292**	.398
	조금	.495	.301	.940	.371	.500	**.295**	.369
	많이	.407	**.298**	.942	.371	.501	.299	.336
40개	없음	1.061	.430	1.754	.522	.707	**.415**	.698
	조금	.982	.427	1.750	.526	.710	**.421**	.635
	많이	.802	**.423**	1.755	.525	.709	.424	.558
60개	없음	1.596	.528	2.557	.643	.866	**.514**	.989
	조금	1.470	.522	2.558	.643	.869	**.518**	.890
	많이	1.207	**.519**	2.554	.644	.869	.523	.777

〈표 13〉 (6) 문항 수, 피험자 능력 표준편차 차이의 동등화 방법별 WRMSE

문항 수	피험자 표준편차 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20개	없음	.458	.303	.943	.370	.500	**.292**	.360
	보통	.448	.299	.940	.371	.500	**.297**	.350
	좁음	.487	**.299**	.938	.372	.501	.302	.366
40개	없음	.901	.426	1.754	.524	.708	**.416**	.609
	보통	.891	.425	1.744	.527	.709	**.424**	.598
	좁음	.966	**.424**	1.760	.525	.711	.428	.628
60개	없음	1.357	.524	2.559	.643	.869	**.514**	.854
	보통	1.329	**.520**	2.550	.644	.869	**.520**	.831
	좁음	1.450	**.520**	2.559	.644	.868	.526	.883

위의 〈표 13〉은 (6) 문항 수, 두 피험자 집단 능력의 표준편차 차이에 대한 동등화 방법별 WRMSE 값이다. 모든 문항 수 조건에서 피험자 능력 표준편차의 차이가 없거나 평균 중심으로 어느 정도 퍼져있는 보통의 수준이면 원호동등화가, 능력 표준편차가 평균 중심으로 좁게 분포되어

있으면 평균동등화가 가장 낮은 WRMSE를 가졌다. 다만 문항 수가 60개인 경우 피험자 능력 표준편차의 차이가 보통일 때 원호동등화와 평균동등화가 동일하게 낮은 동등화 오차를 나타냈다. 그리고 Tucker 선형동등화와 합성동등화 방법도 효과적이었다.

(7) 공통문항 비율, 검사간 난이도 차이에 따른 각 동등화 방법의 WRMSE 값이 다음 〈표 14〉에 제시되어 있다. 전반적인 조건에서 적절한 동등화 방법은 평균동등화와 원호동등화 방법이었다. 그 중 공통문항 비율이 20%와 25% 모두에서 두 검사간의 난이도 차이가 없으면 평균동등화가, 난이도 차이가 조금 또는 많이 있으면 원호동등화가 가장 좋은 수행을 보였다. 그리고 Tucker 선형동등화와 합성동등화도 효과적으로 나타났다.

〈표 14〉 (7) 공통문항 비율, 검사간 난이도 차이의 동등화 방법별 WRMSE

공통문항 비율	난이도 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
20%	없음	.670	**.418**	1.814	.522	.703	.428	.518
	조금	.696	.420	1.815	.522	.705	**.411**	.522
	많이	1.480	.432	1.813	.522	.704	**.419**	.835
25%	없음	.627	**.403**	1.684	.504	.682	.413	.493
	조금	.651	.406	1.691	.504	.682	**.397**	.496
	많이	1.387	.417	1.685	.505	.681	**.406**	.788

〈표 15〉 (8) 검사 난이도 수준, 피험자 수의 동등화 방법별 WRMSE

난이도 수준	피험자 수	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
상	25	.909	**.600**	2.738	.788	1.015	**.600**	.715
	50	.916	**.469**	2.121	.589	.775	**.469**	.635
	75	.909	**.415**	1.872	.514	.679	**.415**	.601
	100	.906	**.386**	1.732	.476	.628	.387	.584
	125	.908	**.367**	1.640	.450	.595	.368	.573
	150	.915	**.353**	1.579	.433	.571	.354	.572
중	25	1.028	.601	2.566	.779	1.013	**.586**	.756
	50	1.026	.469	1.983	.584	.771	**.457**	.675
	75	1.024	.419	1.754	.513	.679	**.409**	.645
	100	1.028	.385	1.616	.471	.625	**.377**	.631
	125	1.033	.368	1.535	.448	.594	**.361**	.626
	150	1.027	.354	1.483	.429	.571	**.346**	.614
하	25	.820	**.525**	2.078	.668	.944	.526	.651
	50	.818	.415	1.613	.502	.719	**.414**	.576
	75	.817	.371	1.418	.438	.629	**.368**	.542
	100	.816	.345	1.313	.403	.581	**.343**	.528
	125	.815	.330	1.252	.384	.552	**.329**	.517
	150	.817	.318	1.208	.367	.530	**.316**	.514

(8) 난이도 수준, 피험자 수에 대한 각 동등화 방법의 WRMSE 값은 위의 〈표 15〉와 같이 나타났다. 대체로 평균동등화와 원호동등화 방법이 효과적이었는데, 두 검사가 어려울 때에는 평균동등화, 두 검사가 보통 수준의 난이도일 때에는 원호동등화, 두 검사 모두 쉬운 경우일 때에는 피험자 수가 25명일 때에만 평균동등화가 그 외의 조건에서는 원호동등화가 적절하였다. 이는 원호동등화가 50명 이하 피험자에서 효과적이라는 Kurtz와 Dwyer(2013)의 결과와 유사한 것이다. Tucker 선형동등화와 합성동등화도 비교적 좋은 수행을 보였다.

다음 〈표 16〉은 (9) 두 검사의 난이도 수준, 검사간 난이도 차이에 대한 동등화 방법별 WRMSE 값이다. 두 검사가 어려운 경우에 검사간 난이도 차이가 없으면 동일동등화, 난이도 차이가 조금 있으면 원호동등화, 차이가 많으면 평균동등화가 효과적이었다. 두 검사간 차이가 없을 때 동일동등화가 낮은 동등화 오차를 가지는 것은 Babcock 외(2012), Kolen과 Brennan(2004)의 결과와 유사하였다. 그리고 두 검사 난이도가 보통 수준인 경우에는 두 검사간 난이도 차이와 상관없이 원호동등화가 적절하였는데, 난이도 차이가 없을 때에는 평균동등화도 함께 동등화 오차가 작았다. 두 검사 난이도가 쉬운 경우에는 검사간 난이도 차이가 없으면 평균동등화, 차이가 조금 또는 많이 있으면 원호동등화가 가장 작은 WRMSE를 나타냈다.

〈표 16〉 (9) 검사 난이도 수준, 검사간 난이도 차이의 동등화 방법별 WRMSE

난이도 수준	난이도 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
상	없음	**.385**	.434	1.945	.540	.708	.419	.425
	조금	.827	.430	1.955	.541	.713	**.429**	.563
	많이	1.519	**.431**	1.942	.543	.711	.448	.852
중	없음	.725	**.428**	1.819	.537	.708	**.428**	.533
	조금	.711	.433	1.829	.537	.710	**.417**	.532
	많이	1.648	.437	1.821	.537	.709	**.423**	.908
하	없음	.837	**.370**	1.482	.461	.662	.415	.558
	조금	.482	.376	1.475	.460	.658	**.367**	.431
	많이	1.133	.405	1.484	.460	.658	**.366**	.674

〈표 17〉 (10) 검사 난이도 수준, 피험자 능력 평균 차이의 동등화 방법별 WRMSE

난이도 수준	피험자 평균 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
상	없음	.935	**.431**	1.941	.542	.710	.436	.644
	조금	1.004	**.432**	1.947	.542	.711	.436	.658
	많이	.808	.431	1.949	.541	.710	**.427**	.559
중	없음	1.202	.433	1.827	.535	.706	**.420**	.764
	조금	1.116	.433	1.823	.536	.709	**.421**	.686
	많이	.881	.432	1.821	.539	.710	**.425**	.595
하	없음	1.056	.402	1.485	.458	.656	**.364**	.677
	조금	.827	.384	1.479	.461	.659	**.377**	.551
	많이	.728	**.377**	1.480	.461	.660	.395	.517

위의 〈표 17〉은 (10) 두 검사의 난이도 수준과 두 검사를 치른 피험자 집단 능력의 평균 차이에 대한 각 동등화 방법의 WRMSE 값이다. 두 검사가 어렵고 피험자 능력 차이가 없거나 조금 있는 경우에는 평균동등화가, 피험자 능력 차이가 많을 때에는 원호동등화가 효과적이었다. 두 검사 난이도가 보통 수준에서는 피험자 능력 평균 차이에 상관없이 원호동등화가 적절하였고, 두 검사가 쉽고 피험자 능력 평균 차이가 없거나 조금 있을 때에는 원호동등화 차이가 많을 경우에는 평균동등화가 가장 낮은 WRMSE를 가졌다. 그 다음으로는 Tucker 선형동등화, 합성동등화의 순으로 낮은 동등화 오차를 나타냈다.

다음의 〈표 18〉은 (11) 두 검사의 난이도 수준과 피험자 집단 능력의 표준편차 차이에 대한 각 동등화 방법의 WRMSE 값이다. 두 검사의 난이도가 어렵거나 쉬운 경우에는 피험자 능력 표준편차 차이가 없으면 원호동등화, 표준편차가 평균 중심으로 어느 정도 퍼져 있는 보통이거나 평균 중심으로 밀집되어 있으면 평균동등화가 안정적이었다. 다만, 난이도가 어려운 경우에 표준편차 차이가 보통이면 평균동등화 뿐 아니라 원호동등화도 동일하게 낮은 WRMSE를 가졌다. 그리고 두 검사의 난이도가 보통 수준인 경우에는 모든 피험자 능력 표준편차 차이에서 원호동등화가 가장 효과적인 것으로 나타났다. 평균동등화, 원호동등화 다음으로는 Tucker 선형동등화, 합성동등화가 낮은 WRMSE를 나타냈다.

〈표 18〉 (11) 검사 난이도 수준, 피험자 능력 표준편차 차이의 동등화 방법별 WRMSE

난이도 수준	피험자 표준편차 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
상	없음	.813	.431	1.947	.541	.711	**.429**	.575
	보통	.903	**.431**	1.945	.542	.711	**.431**	.612
	좁음	1.063	**.432**	1.948	.543	.710	.437	.673
중	없음	1.023	.432	1.829	.536	.709	**.421**	.664
	보통	1.004	.433	1.811	.539	.708	**.423**	.641
	좁음	1.059	.433	1.825	.538	.710	**.424**	.666
하	없음	.880	.390	1.480	.460	.657	**.372**	.585
	보통	.760	**.380**	1.478	.461	.660	.387	.526
	좁음	.779	**.378**	1.484	.461	.660	.395	.538

〈표 19〉 (12) 검사간 난이도 차이, 피험자 능력 평균 차이의 동등화 방법별 WRMSE

난이도 차이	피험자 평균 차이	동등화 방법						
		동일	평균	명목가중	Tucker	연쇄	원호	합성
없음	없음	**.043**	.408	1.742	.511	.688	.400	.343
	조금	.461	**.408**	1.747	.513	.693	.413	.428
	많이	1.038	**.414**	1.753	.514	.693	.435	.637
조금	없음	1.187	.422	1.759	.513	.693	**.404**	.695
	조금	.833	.415	1.756	.514	.695	**.405**	.560
	많이	**.342**	.408	1.748	.513	.693	.404	.395

	없음	1.964	.435	1.752	.512	.692	**.416**	1.046
많이	조금	1.653	.427	1.746	.513	.691	**.416**	.907
	많이	1.037	.418	1.750	.514	.694	**.407**	.638

(12) 두 검사의 검사간 난이도 차이와 피험자 집단 능력의 평균 차이에 대한 동등화 방법별 WRMSE 값은 위의 〈표 19〉와 같다. 두 검사의 난이도 차이가 없을 때 피험자 능력 평균 차이가 없으면 동일동등화, 평균 차이가 조금 있거나 많이 있으면 평균동등화가 효과적이었고, 검사간 난이도 차이가 조금 있을 때에는 피험자 능력 평균 차이가 없거나 조금 있으면 원호동등화, 평균 차이가 많이 나면 동일동등화가 가장 낮은 WRMSE를 가졌다. 또한, 검사간 난이도 차이가 많이 나는 경우에는 피험자 능력 평균 차이의 모든 조건에서 원호동등화가 동등화 오차가 낮게 나타났다. Tucker 선형동등화, 합성동등화도 비교적 좋은 수행을 보였다.

V. 논의 및 결론

본 연구는 소규모의 피험자가 검사를 응시하는 실제의 다양한 검사 상황을 반영하기 위해 검사동등화의 실시에서 중요하게 고려되는 조건들을 되도록 많이 포함시키고자 하였다. 연구의 결과는 이전 연구의 결과들과 유사하게 나타나기도 하고 다소 다른 결과를 보이기도 하였다. 연구 결과에서 차이가 나타나는 이유는 연구 조건의 차이로 보이므로 실제에서의 동등화 방법의 선택은 여러 조건과 상황을 고려하여 이루어져야 하고, 차이를 보인 연구들에 대한 후속 연구가 이루어질 필요가 있다. 이전 연구들과 본 연구결과를 모두 고려하여 소규모 집단에 대한 각 조건별 효과적인 검사동등화 방법의 탐색 및 비교를 정리하면 다음과 같다. 첫째, 본 연구에서 고려한 대부분의 조건들에서 평균동등화와 원호동등화 방법의 동등화 오차가 가장 낮게 나타났다. 따라서 소규모 집단을 위한 검사 동등화를 실시하고자 할 때, 본 연구에서 고려한 다양한 검사의 조건과 유사하다면 평균동등화 방법과 원호동등화 방법을 사용하여 동등화를 진행하고 변환된 점수를 사용할 수 있다. 이는 동등화를 하지 않는 동일동등화의 조건도 포함되는 결과이므로 두 가지 동등화 방법은 안정적이라고 할 수 있고 사용하는 것에 무리가 없다.

둘째, 특정 조건에서 동일동등화 방법, 즉, 동등화하지 않는 방법의 선택을 고려할 수 있다. 문항의 수가 20개 내외, 피험자 수가 25명 내외, 기준검사와 변환검사간의 난이도 차이가 없거나 조금 있어야 효과적이다. 검사의 결과에서 대부분의 경우에 평균동등화와 원호동등화 방법의 동등화 오차가 가장 작게 나타났으나, 두 검사가 어렵고 검사간 난이도 차이가 없는 경우, 두 검사간 난이도 차이가 없고 피험자 능력 평균 차이도 없는 경우, 검사간 난이도 차이가 조금 있고 피험자 능력 평균 차이가 많이 나는 경우에는 동일동등화 방법이 가장 낮은 동등화 오차를 가졌다. 따라서 그러한 조

건이라면 동등화하지 않고 원점수를 사용할 수 있다. 하지만 난이도 차이가 커지면 동등화 오차가 매우 커지는 것으로 나타났으므로 사용에 주의를 요하고, 같은 검사 명세표를 사용하는 동형의 문항 제작에 매우 신중해야 한다는 전제가 충족되어야 한다.

셋째, 조건에 따라 더 효과적인 동등화 방법을 선택하여 사용할 수 있다. 평균동등화와 원호동등화 두 가지 동등화 방법 모두 본 연구에서 설정한 다양한 조건에서 좋은 수행을 보였으나 위의 둘째 정리와 같이 동일동등화가 가장 낮은 동등화 오차를 가진 특정한 경우가 있고, 또 다른 특정 조건에 따라 평균동등화와 원호동등화 중 조금 더 효과적인 방법이 있었다. 예를 들어, 피험자 능력 표준편차가 평균 중심으로 좁게 분포되어 차이가 나면 설정된 모든 문항 수에서 평균동등화 방법이 가장 안정적이었다. 그리고 기준검사와 변환검사의 난이도가 보통 수준이면 모든 조건에서 원호동등화 방법의 사용이 권장된다. 또한, 동등화를 실시하는 검사가 준거참조검사이고, 준거 점수가 낮거나 높은 양극단에 위치한다면 역시 원호동등화 방법이 적절하다. 원호동등화 방법이 양극단 점수에서 낮은 동등화 오차를 가졌기 때문이다. 다만, 원호동등화를 사용할 때에는 중간점이나 최저점을 달리하여 최적의 결과를 산출하기 위한 노력을 기울이면 더 정확한 점수 산출이 가능하다. 그리고 Tucker 선형동등화 방법과 합성동등화 방법도 조건에 따라 비교적 좋은 수행을 보였으므로 검사의 다양한 조건을 확인하여 동등화 방법의 선택과 활용에 참고할 수 있다.

본 연구는 소규모 집단에 대한 동등화 방법들을 비교한 연구로 다양한 조건에 대한 적절한 동등화 방법을 탐색하고자 하였으며, 이를 위해 공통문항 비동등집단 설계로 모의자료를 생성하여 연구를 진행하였다. 본 연구의 한계를 고찰하고 후속 연구를 제안하면 다음과 같다. 먼저, 본 연구는 모의실험 연구로 연구자가 설정한 조건의 모의자료를 이용하여 연구를 진행하였으므로 현실의 검사 상황과 부합되지 않을 수 있다. 그리고 연구의 조건에서 문항의 조건은 문항 수만 포함하여 문항의 종류를 고려하지 못하였다는 점, 검사의 난이도와 관련된 조건만 포함시키고 변별도 조건을 포함하지 못하였다는 점 등의 한계를 지닌다. 따라서 본 연구에서 고려하지 못한 선다형, 구성형 문항 모두에 대한 문항의 종류 조건 및 검사의 변별도 조건 등을 포함하는 후속 연구를 제안한다. 이때, 문항의 종류 조건을 포함한 연구에서는 이전 연구에서 거의 다루지 않았던 구성형 문항 비율, 공통문항의 구성형 문항 포함 여부, 채점자 효과 등의 조건이 고려되는 것이 필요하다. 또한, 동등화 오차와 관련된 조건 및 산출 방법을 달리한 연구나, 활용이 수월할 수 있도록 손쉬운 검사 동등화 통계처리 도구에 대한 연구 등도 요구된다.

참고문헌

김연정(2009). An investigation of the effect of the anchor test length and the non-equivalency of equating groups on equating. 석사학위논문, 연세대학교.

남현우(2001). **검사동등화 방법**. 서울; 교육과학사.

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

안수현(2016). 모의실험에 의한 가교피험자설계의 조건에 따른 동등화 방법 비교. 박사학위논문, 성균관대학교.

우중호(2016). 소규모 자료를 이용한 Circle-Arc 동등화 방법의 최저점 변화에 따른 동등화 오차 크기 연구. 석사학위논문, 충남대학교.

이규민(2005). 초등학교 3학년 국가수준 기초학력 진단평가 동등화 방안. **교육평가연구, 18**(1), 125-151.

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

임의진(2011). Comparison of small-sample equating methods for mixed-format tests in a NEAT design. 석사학위논문, 연세대학교.

임의진, 이규민(2017). 공통문항 비동등집단 설계 하에서 혼합 문항유형 검사에 대한 소표본 동등화 방법 비교. **교육평가연구, 30**(2), 199-218.

Albano, A. D. (2016). Equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8).

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.).* Washington D.C.: American Council on Education.

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice, 16*(2), 647-668.

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational and Psychological Measuremnet, 72*(4), 608-628.

Brennan, R. L. (편). (2015). Educational measurement (4nd ed.). 교육측정 1. 한국교육평가학회, 재미한인교육연구자협회 (공역). 서울: 학지사.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and a postsmoothing methods in equipercentile equating (ACT Research Report 94-4).* Iowa

City, IA: American College Testing.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equaing using a synthetic linking function. *Journal of Educational Measurement, 45,* 325-342.

Kim, S., & Livingston, S. A. (2010). Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement, 47,* 286-298.

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.

Kurtz, A, M., & Dwyer A. C. (2013). *Small sample eqauting: Best practice using a SAS Macro.* North Carolina State University, SESUG, Paper BTB-11. https://analytics.ncsu.edu/sesug/2013/BtB-11.pdf

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

Livingston, S. A., & Kim, S. (2008). *Small-sample equating by the circle-arc method (Research report 08-39).* Princeton, NJ: Educational Tesing Service.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46,* 330-343.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47,* 175-185.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32,* 37-54.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309-330.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The kernel method of test equating. New York: Springer.

Wang, T., & Brennan, R. L. (2009). A modified frequency estimation equating method for the common-item nonequivalent groups design, *Applied psychological measurement, 33*(2), 118-132.

* 논문접수 2019년 5월 2일 / 1차 심사 2019년 6월 7일 / 2차 심사 2019년 9월 1일 / 게재승인 2019년 9월 5일

* 김화영: 성균관대학교 사범대학 교육학과를 졸업하고, 동대학원에서 교육학 박사학위를 취득하였다. 현재 한국행동과학연구소에서 연구위원으로 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

A Comparison of Equating Methods in Small Samples : Simulation Study*

Kim, Hwa Young**
Kim, Hyunchul***

The purpose of this study is to compare the equating methods for small samples with various conditions under the common-item nonequivalent groups design. The conditions considered in this simulation study include item sizes, proportions of common-item, sample sizes, levels of test difficulty, levels of difference in difficulty between forms, levels of difference of mean in group ability, and levels of difference of standard deviation in group ability. And seven equating methods were applied to the above mentioned conditions.

The results shows that mean equating and circle-arc equating were mostly stable under all conditions and depending on the conditions, one method may be more effective than the other. However identity equating had the lowest error of equating under the following conditions: if forms were difficult and there was no difference in the difficulty between forms; there was no difference in the difficulty between forms and no difference of mean in group ability; there was a slight difference in the difficulty between forms and a large difference of mean in group ability.

Key words: small samples, small sample equating, test equating, simulation study

* This study was modified and supplemented based on a doctoral dissertation titled 'A Comparison of Equating Methods in Small Samples(2019)' of Sungkyunkwan University Graduate School of General Studies.

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