The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 977-994

Empirical analysis of equating methods for elective subjects of College Scholastic Ability Test[1)]

Hyunchul Kim[2)]

Abstract

Five equating methods for elective subjects of College Scholastic Ability Test were analyzed under a common-items nonequivalent groups design using a real data set of 110 thousand examinees. Five methods were (a)two-stage linear equating, (b)two-stage equi-percentile equating, (c)Tucker equating, (d)Frequency estimation equating, and (e)Braun-Holland equating. The results indicated that Frequency estimation equating fits well, and two-stage linear equating produces most different equating results from the Frequency estimation equating.

1. 서 론

99학년도의 대학입학수학능력시험에는 교육과정의 개정에 의하여 수리·탐구II 영역에 선택과목이 도입되었다. 수리·탐구II 사회탐구의 경우에는 인문계열은 공통사회, 국사, 윤리를 필수로 하고, 정치, 경제, 사회문화, 세계사, 세계지리의 다섯 과목 중 한 과목을 택일하도록 하였으며, 수리·탐구II의 자연탐구 경우에는 자연계열은 공통과학을 필수로 하고, 물리II, 화학II, 생물II, 지구과학II의 네 과목 중 한 과목을 택일하도록 하였다.

대학수학능력시험과 같은 경쟁시험 상황에서는 출제시에 어느 과목을 선택하던지 수험생의 학업능력 이외의 요인에 의하여 점수차이가 생기지 않도록 하여야 한다. 그러나 선택과목의 난이도를 완벽하게 일치시키는 것은 어려운 일이며, 뿐만 아니라 표준점수를 사용하는 경우에는 선택과목의 난이도 이외에 선택과목간 응시집단의 수학능력 차이에 따라서도 점수차이가 생길 수 있다. 즉, 응시집단의 전반적인 수학능력이 높은 과목을 선택한 수험생의 점수는 자신의 능력 이하로 평가될 가능성이 있으며, 응시집단의 수학능력이 낮은 과목을 선택한 수험생의 점수는 자신의 능력 이상으로 평가될 가능성이 있다.

이러한 요인들에 의하여 발생한 수험생의 점수차이는 사후에 보상하여 수험생간의 형평성을 보장하여야만 하기 때문에 이를 위하여 선택과목 점수의 동등화(test score equating) 과정이 요구되게 되었다. 이에 한국교육과정평가원은 99학년도의 입시에 선택과목 점수의 선형 동등화(linear equating) 방식을 적용하여 각 선택과목의 동등화된 점수를 산출하였다. 한국교육과정평가원의 동등화 방식은 Angoff(1971)가 '제5설계(Design V)'라 명명한 방식으로, 이는 한 선택과목(X)의 점

1) This paper was supported by SEOK CHUN Research Fund, Sungkyunkwan University, 1999.
2) Assistant Professor, Sungkyunkwan University, Seoul 110-745, Korea.

수를 전체 수험생이 공동으로 치른 공통과목(V)의 점수척도로 변환하고, 이를 다시 다른 선택과목(Y)의 점수척도로 변환하는 이단계 동등화 방법이다. 이단계 동등화를 실시하는 방식으로 선형 동등화 방식과 등백분위 동등화 방식이 있는데, 한국교육과정평가원은 그 중 선형 동등화 방식을 채택하였다. 한국교육과정평가원은 수험생들이 공동으로 치른 수리·탐구II의 공통과목을 수리·탐구II의 선택과목 점수 동등화를 위한 가교검사(anchor test)로 사용한 것이다.

한편 Angoff(1971)는 선택과목 X를 치른 집단과 선택과목 Y를 치른 집단의 합집단(synthetic population, S)에 대하여 공통과목(V)의 점수를 이용하여 동등화를 실시하는 합집단 동등화 방법을 '제4설계(Design IV)'라 칭하였는데, Braun과 Holland(1982, pp.39-42)는 가교검사(V) 점수를 이용하여 동등화를 실시하는 경우에는 합집단 동등화 방식이 이단계 동등화 방식보다 정확하며, 이단계 동등화 방식은 동등화가 적용되는 집단에 대한 정의가 명확하지 않은 이론적 약점을 가지고 있다고 지적하였다. 반면에 이단계 동등화 방식은 합집단 동등화 방식에 비하여 컴퓨터의 계산시간이 짧고 기억용량이 많이 필요하지 않다는 장점을 가지고 있다. 합집단 동등화의 수행에도 선형동등화와 등백분위 동등화 방법을 사용할 수 있다.

한국교육과정평가원이 99학년도의 대입수능시험에 사용한 선형 동등화 방식은 동등화가 이루어지는 두 자료의 분포가 평균과 표준편차를 제외하고는 동일하다는 것과, 두 시험점수가 선형의 관계를 갖는다는 강한 가정들을 부과하고 있기 때문에 이러한 가정들이 만족되지 않는 경우에는 동등화의 결과가 적절하지 않을 수 있다. 또한 이 방식은 동등화 이후의 점수가 0점보다 아주 작거나 만점보다 아주 크게 나타날 수 있다. 반면에 등백분위 동등화 방식은 자료의 분포에 대하여 약한 가정을 부과하며, 원점수가 0점인 경우와 만점인 경우에, 동등화가 이루어진 후에도 0점과 만점에 근사한 점수로 유지될 수 있다는 장점을 가지고 있다. 그러나 등백분위 동등화 방식은 해당 백분위에 관측 자료수가 많지 않은 경우에는 등화식이 심한 곡률의 변화를 보인다는 문제점이 있다. 대학입학수학능력시험에서는 선택과목에 따라 응시자의 수가 많지 않은 과목이 있을 것으로 예상되며, 특히 응시자의 수가 많지 않은 과목의 최상위 점수와 최하위 점수 부근에서는 관측 자료수가 많지 않아서 이러한 현상이 나타날 가능성이 있을 것으로 보인다.

동등화 방법들은 Angoff(1971), Braun과 Holland(1982), 그리고 Kolen과 Brennan(1995)에서 종합적으로 정리되고 비교되었다. 또한 여러 가지 실험설계하에서 동등화 방법들을 비교한 연구들은 Jarjoura와 Kolen(1985), Kolen과 Jarjoura(1987), Woodruff(1989), 그리고 Harris와 Kolen(1990)에 의하여 수행되었다. 우리나라에서는 남현우(1992a)가 여러 실험설계에서의 동등화 방법들을 정리하였으며, 남현우(1992b)는 문항난이도, 변별도, 추측모수의 변화에 따른 동등화 방법들의 강인성을 모의실험에 의하여 비교하였다. 또한 성태제(1994)는 대학별 고사의 선택과목 점수 동등화를 위한 기존의 방법들을 검토하였으며, 허명회(1995)는 대학별 고사의 선택과목 점수 동등화 방안을 제시하고 이를 실제 자료에 적용한 것을 예시하였다.

이 연구는 한국교육과정평가원이 99학년도 입시에 사용한 이단계 선형 동등화 방식을 비롯한 이단계 동등화 방법들과 합집단 동등화의 여러 가지 대표적인 방식들을 실제 자료에 적용한 결과를 분석하여 수능시험 자료에 대한 각 동등화 방법들의 적절성을 비교하는 것을 목적으로 한다. 비교된 동등화 방법들은 (1)이단계 선형동등화(linear equating), (2)이단계 등백분위 동등화(equipercentile equating), (3)Tucker의 합집단 선형동등화(Tucker equating), (4)합집단 등백분위

동등화 방식인 빈도추정 동등화(Frequency estimation equating), 그리고 (5)빈도추정 동등화의 가정에 의하여 각 선택과목의 평균과 분산을 구하고 이를 선형동등화에 적용하는 Braun-Holland 동등화의 다섯 가지이다.

동등화 방법의 선택은 선택과목 점수 동등화의 결과에 영향을 미쳐서 수험생의 점수에 상당한 차이가 생기게 될 수 있다. 그러므로 대학입시와 같은 경쟁시험 상황에서는 동등화 방식의 선정에 주의를 기울여야 한다. 주요 동등화 방식들을 실제 자료에 적용하고 비교함으로써 대학입학수학능력시험 자료에 대한 각 동등화 방법들의 적절성을 평가할 수 있을 것이며, 그 결과는 이후의 대학입학수능시험의 선택과목 점수 동등화 방법의 선정에 의미있는 시사를 줄 것으로 기대된다.

2. 연구방법

2.1 동등화가 적용된 자료

다섯 가지 동등화 방법들이 실제 자료에 적용되고 그 결과가 비교·분석되었는데, 이를 위하여 사용된 자료는 98년 5월에 서울의 모 대학입학전문 사설학원에서 약 23만 명의 고등학교 3학년 학생들과 재수생을 대상으로 실시한 모의 수학능력시험 자료로, 응시자들 중에서 인문계열은 112,670명, 자연계열은 약 12만 명이었다. 이 모의시험에 응시한 인원은 99학년도 대학입학수학능력시험에 응시한 인원의 약 30%에 해당한다. 이 연구에서는 이중 인문계열 수험생들이 정치, 경제, 사회문화, 세계사, 세계지리의 다섯 과목 중 하나를 선택하여 획득한 선택과목 점수의 자료에 대하여 다섯 가지의 대표적인 동등화 방식들을 적용하여 동등화를 실시하고 그 결과를 비교하였다.

2.2 동등화 방법

가. 이단계 선형 동등화

한국교육과정평가원에서는 다음과 같은 공식을 적용하여 선택과목의 동등화 점수를 산출하였다.

$$x' = \frac{\sigma(V)}{\sigma(X)}[x-\mu(X)] + \mu(V) \quad (1)$$

여기서 x' = 동등화된 선택과목 X의 점수

x = 선택과목 X의 원점수(raw score)

$\sigma(V)$ = X를 선택한 집단의 공통시험과목(V) 점수의 표준편차

$\sigma(X)$ = 선택과목 X 점수의 표준편차

$\mu(V)$ = X를 선택한 집단의 공통시험과목(V) 점수의 평균

$\mu(X)$ = 선택과목 X 점수의 평균

이는 각 선택과목의 점수를 각 선택과목별 응시집단의 공통과목(V) 점수의 척도로 변환함으로써

서로 다른 선택과목의 점수를 비교할 수 있도록 한 방식이다. 각 선택과목 점수의 평균이 각 선택과목별 응시집단의 공통과목 평균과 같아지도록 하여 응시집단의 학업능력 차이가 선택과목 점수에 반영되도록 한 것이다. 공통과목으로는 인문계의 경우에는 수리·탐구II 사회탐구의 공통사회가, 자연계의 경우에는 수리·탐구II 자연탐구의 공통과학이 사용되었다.

이 방법은 Angoff(1971)에 의하여 '제5설계(Design V)'로 명명된 이단계 동등화방법인 한 개의 선택과목(X) 점수를 가교검사인(anchor test)인 공통과목(V) 점수와 동등화하고 이를 다시 다른 선택과목(Y)의 점수와 동등화함으로써 두 선택과목 X와 Y의 점수를 동등화하는 것과 같은 형태이다. Angoff(1971)의 이단계 동등화 방법에는 선형 동등화와 등백분위 동등화가 모두 사용될 수 있는데 한국교육과정평가원은 그 중 선형 동등화의 방식을 사용한 것이다. Braun과 Holland(1982)는 이단계 동등화가 합집단 동등화와는 달리 동등화가 적용되는 집단을 명백히 정의하지 못하는 문제점을 가지고 있다고 지적하였다. Braun과 Holland(1982)는 이단계 동등화의 결과가 특정 집단에 대하여 정확한 결과를 제공하지 못한다는 것을 보였다.

한국교육과정평가원이 인문계 수험생 자료에 적용한 방식은 정치를 포함한 모든 선택과목의 점수를 공통과목 점수의 척도로 변환하여 비교한 것이나, 이 연구에서는 선택과목 중 정치과목의 점수는 변환하지 않고, 공통과목 점수의 척도로 변환된 그 이외의 선택과목 점수를 다시 정치 과목 점수의 척도로 변환하였다. 이는 다른 동등화 방법들과의 비교를 위한 것으로 이 방식은 한국교육과정평가원의 방식과 동일한 성질을 갖는다.

나. 이단계 등백분위 동등화

이단계 등백분위 동등화는 선택과목 X의 점수를 공통과목 V의 점수 척도로 변환하고, 이 변환된 점수를 다시 선택과목 Y의 점수 척도로 변환한다는 점에서 앞의 이단계 선형 동등화와 같은 형태이나, 동등화를 위하여 선형의 관계를 설정하지 않고 등백분위 동등화 방식을 사용한다는 것이 앞의 방식과의 차이점이다.

즉, 첫 번째 단계에서는 선택과목 X를 치른 집단1에 대하여 선택과목 점수(X)와 집단1의 공통과목 점수(V)간에 다음의 등백분위 동등화가 수행된다.

$$e_V(x) = P_1^{-1}[F(x)] \tag{2}$$

여기서 e_v는 V의 척도로 변환된 X 점수

P_1은 X를 선택한 집단의 공통과목 V 점수의 백분위함수(percentile rank function)

F는 X를 선택한 집단의 선택과목 X 점수의 백분위함수

이에 의하여 공통과목 V의 점수척도로 변환된 선택과목 X의 점수는 다음에 의하여 다시 선택과목 Y의 점수척도로 변환된다.

$$e_Y(v) = G^{-1}[P_2(v)] \tag{3}$$

여기서 e_y는 Y의 척도로 변환된 V 점수

P_2는 Y를 선택한 집단의 공통과목 V 점수의 백분위함수
G는 Y를 선택한 집단의 선택과목 Y 점수의 백분위함수

위의 식에서 백분위함수(percentile rank function) P(x)는 x미만의 점수를 획득한 수험생의 백분율에 점수 x를 획득한 수험생의 백분율의 절반을 더한 값이 된다. x 바로 아래의 점수가 x-1일 때 이를 식으로 표현하면 $P(x)=100[F(x-1)+f(x)/2]$가 된다. $P^{-1}(x)$는 이의 역함수이며 보통 백분위수함수(percentile function)라 불린다. 이 연구에서는 정치과목을 제외한 모든 선택과목의 점수를 등백분위 동등화에 의하여 각 집단 내에서 공통과목 V의 척도로 변환한 후에, 이를 다시 정치를 선택한 집단의 백분위함수를 이용하여 정치과목의 척도로 변환하였다.

다. Tucker의 합집단 선형 동등화

한편 Angoff(1971)는 선택과목 X와 Y를 치른 두 집단의 합집단을 고려하여 이 합집단에서의 X와 Y의 점수분포를 구한 후 공통과목 V의 점수에 의하여 두 검사 점수를 동등화하는 합집단 동등화방법을 '제4설계(Design VI)'라 명명하였다. 이 방법에도 역시 선형 동등화 방식과 등백분위 동등화 방식이 모두 사용될 수 있는데, Tucker 동등화(Tucker equating)는 그 중 선형 동등화의 방식을 사용한 것이다.

Tucker의 합집단 선형 동등화는 Gulliksen(1950, pp.299-301)에 의하여 Tucker에 의한 동등화 방식으로 지칭되었으며, Harris와 Kolen(1990), Kolen과 Brennan(1995) 등에 의하여 논의되었다. Tucker 동등화는 선택과목 X를 치른 집단과 선택과목 Y를 치른 집단의 합집단의 평균과 표준편차를 이용하는 다음의 식에 의하여 수행된다.

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

여기서 s는 선택과목 X를 치른 집단과 선택과목 Y를 치른 집단의 합집단을 의미하는데, 각 집단은 그 크기에 따라 가중치 w_1과 w_2를 가지게 된다. 이때 합집단에서의 선택과목 X의 평균과 표준편차, 선택과목 Y의 평균과 표준편차는 다음과 같이 구해진다.

$$\mu_s(X)=\omega_1\mu_1(X)+\omega_2\mu_2(X) \tag{5}$$

$$\mu_s(Y)=\omega_1\mu_1(Y)+\omega_2\mu_2(Y) \tag{6}$$

$$\sigma_s^2(X)=\omega_1\,\sigma_1^2(X)+\omega_2\,\sigma_2^2(X)+\omega_1\omega_2[\mu_1(X)-\mu_2(X)]^2 \tag{7}$$

$$\sigma_s^2(Y)=\omega_1\,\sigma_1^2(Y)+\omega_2\,\sigma_2^2(Y)+\omega_1\omega_2[\mu_1(Y)-\mu_2(Y)]^2 \tag{8}$$

그런데 합집단 동등화에서 집단 1은 선택과목 X만을 치렀고, 집단2는 선택과목 Y만을 치렀으므로 Tucker 동등화 방식은 위의 (식5)-(식8)을 추정하기 위하여 다음 두 가지를 가정한다.

첫 번째 가정은 회귀식의 기울기를 α, 절편을 β라고 할 때

$$\alpha_2(X|V)=\alpha_1(X|V),\ \ \beta_2(X|V)=\beta_1(X|V) \tag{9}$$

$$\alpha_1(Y|V)=\alpha_2(Y|V),\ \ \beta_1(Y|V)=\beta_2(Y|V) \tag{10}$$

여기서 첨자 1과 2는 각각 집단1과 집단2를 지칭
X|V는 종속변수 X에 대한 독립변수 V의 회귀계수임을 의미
Y|V는 종속변수 Y에 대한 독립변수 V의 회귀계수임을 의미
식들의 좌변은 직접 관측되지 않는 값

이라는 것인데, 이는 선택과목 X의 점수와 공통과목 V의 점수간의 관계와 선택과목 Y의 점수와 공통과목 V의 점수간의 관계는 집단 1과 집단 2에서 동일한 선형회귀식을 갖는다는 가정이다.

두 번째 가정은

$$\sigma_2^2(X)[1-\rho_2^2(X,V)]=\sigma_1^2(X)[1-\rho_1^2(X,V)] \tag{11}$$

$$\sigma_1^2(Y)[1-\rho_1^2(Y,V)]=\sigma_2^2(Y)[1-\rho_2^2(Y,V)] \tag{12}$$

여기서 ρ는 상관계수,
첨자1과 2는 각각 집단1과 집단2를 지칭
식들의 좌변은 직접 관측되지 않는 값

이라는 것인데, 이는 주어진 공통시험과목 V의 점수 하에서 X점수와 Y점수의 조건부 분산은 집단 1과 집단 2에서 동일하다는 것이다.

이러한 가정을 부과하면 (식5)-(식8)은 다음 식에 의하여 추정될 수 있다.

$$\mu_s(X)=\mu_1(X)+\omega_2\gamma_1[\mu_1(V)-\mu_2(V)] \tag{13}$$

$$\mu_s(Y)=\mu_2(Y)+\omega_1\gamma_2[\mu_1(V)-\mu_2(V)] \tag{14}$$

$$\sigma_s^2(X)=\sigma_1^2(X)-\omega_2\gamma_1^2[\sigma_1^2(V)-\sigma_2^2(V)]+\omega_1\omega_2\gamma_1^2[\mu_1(V)-\mu_2(V)]^2 \tag{15}$$

$$\sigma_s^2(Y)=\sigma_2^2(Y)+\omega_1\gamma_2^2[\sigma_1^2(V)-\sigma_2^2(V)]+\omega_1\omega_2\gamma_2^2[\mu_1(V)-\mu_2(V)]^2 \tag{16}$$

여기서 γ_1과 γ_2는 회귀선의 기울기로 다음과 같이 추정된다.

$$\gamma_1=\alpha_1(X|V)=\sigma_1(X,V)/\sigma_1^2(V) \tag{17}$$

$$\gamma_2=\alpha_2(Y|V)=\sigma_2(Y,V)/\sigma_2^2(V) \tag{18}$$

여기서 첨자 1과 2는 각각 집단1과 집단2를 지칭
σ(X, V)는 X와 V의 공분산
σ(Y, V)는 Y와 V의 공분산

이 연구에서는 각 선택과목을 선택한 응시자 집단과 정치를 선택한 응시자 집단의 합집단에 대하여 선형 동등화를 수행하여 각 선택과목의 Tucker 동등화 점수를 산출하였다.

라. 빈도추정 동등화(frequency estimation equating)

빈도추정 동등화 방식은 합집단에 대하여 등백분위 동등화 방식을 적용한 것으로 Braun과 Holland(1982), Jarjoura와 Kolen(1985), Kolen과 Jarjoura(1987), Harris와 Kolen(1990)에 의하여 논의되었다. 이 방식은 선택과목 X의 점수와 선택과목 Y의 점수에 대한 합집단의 분포를 다음과 같이 산출한다.

$$f_s(x) = \omega_1 f_1(x) + \omega_2 f_2(x) \tag{19}$$

$$g_s(y) = \omega_1 g_1(y) + \omega_2 g_2(y) \tag{20}$$

여기서 s는 합집단을 의미하며, 첨자 1과 2는 각각 선택과목 X와 Y를 치른 집단을 지칭하며, f와 g는 각각 선택과목 X와 Y의 점수의 확률밀도함수를 의미한다.

이 방식은 Tucker의 합집단 선형 동등화의 경우와 마찬가지로 집단 1은 선택과목 X의 점수에 대한 자료만을 가지고 있고, 집단 2는 선택과목 Y의 점수에 대한 자료만을 가지고 있으므로, 위의 식을 추정하기 위하여 모든 공통과목 점수 v에 대하여 다음을 가정한다.

$$f_1(x|v) = f_2(x|v), \quad g_1(y|v) = g_2(y|v) \tag{21}$$

이는 공통과목(V)의 점수가 v로 주어졌을 때, 선택과목 X의 점수와 선택과목 Y의 점수의 조건부 확률밀도함수가 두 집단에서 동일하다는 것이다.

이러한 가정 하에서 (식19)와 (식20)은 다음과 같이 추정될 수 있다.

$$f_s(x) = \omega_1 f_1(x) + \omega_2 \sum_v f_1(x|v) h_2(v) \tag{22}$$

$$g_s(y) = \omega_1 \sum_v g_2(y|v) h_1(v) + \omega_2 g_2(y) \tag{23}$$

여기서 $h_1(v)$와 $h_2(v)$는 각각 집단 1과 2에서의 공통과목 점수의 주변분포함수(marginal distribution function)이다. 위의 식을 사용하여 구한 선택과목 X와 선택과목 Y의 점수의 백분위 함수(percentile rank function)를 P_s와 Q_s라 할 때 합집단에 대한 등백분위 동등화는 다음과 같이 이루어진다.

$$e_{Ys}(x) = Q_s^{-1}[P_s(x)] \tag{24}$$

빈도추정 동등화는 주변분포만을 요구하는 이단계 동등화나 Tucker의 합집단 선형 동등화와는 달리 선택과목 점수와 공통과목 점수간의 결합확률밀도함수(joint probability density function)를 필요로 하므로 컴퓨터의 계산시간이 길고, 기억용량이 많이 필요하게 된다는 단점이 있다. 이 연구에서는 각 선택과목을 선택한 응시자 집단과 정치를 선택한 응시자 집단의 합집단에 대하여 등백분위 동등화를 수행하여 각 선택과목의 빈도추정 동등화 점수를 산출하였다.

마. Braun-Holland 동등화

Braun과 Holland(1982)는 빈도추정 동등화에 적용된 가정 하에서 합집단의 평균과 분산을 산출한 후 이를 선형 동등화에 이용하는 방식을 제안하였다. 이 방법은 평균과 표준편차를 산출하는 방식을 제외하고는 Tucker의 합집단 선형 동등화와 유사하다. 합집단의 선택과목 X와 선택과목 Y 점수의 평균과 분산은 각각 다음 식에 의하여 산출된다.

$$\mu_s(X) = \sum_x x f_s(x) \tag{25}$$

$$\mu_s(Y) = \sum_y y g_s(y) \tag{26}$$

$$\sigma_s^2(X) = \sum_x [x - \mu_s(X)]^2 f_s(x) \tag{27}$$

$$\sigma_s^2(Y) = \sum_y [y - \mu_s(Y)]^2 g_s(y) \tag{28}$$

이때 $f_s(x)$와 $g_s(y)$는 빈도추정 동등화와 같은 방식으로 앞의 (식22)와 (식23)에 의하여 구해진다.

Braun과 Holland의 동등화는 집단 1에서 X의 V에 대한 관계식과 집단 2에서 Y의 V에 대한 관계식이 선형이며, 동일할 때는 Tucker의 합집단 선형 동등화와 결과가 일치한다. 그러므로 Braun-Holland 동등화는 Tucker 동등화의 일반형이라고 할 수 있으며, X의 V에 대한 관계식이나 Y의 V에 대한 관계식이 선형이 아니거나 동일하지 않아서 Tucker 동등화의 가정이 만족되지 않을 때 유용하게 사용될 수 있다. Braun-Holland 동등화의 단점은 Tucker 동등화에 비하여 계산이 복잡하다는 것이다. 이 연구에서는 각 선택과목의 응시자 집단과 정치과목을 선택한 응시자 집단의 합집단에 대하여 앞의 식에 의한 평균과 표준편차를 산출하고, 이를 사용하여 Braun-Holland의 동등화 점수를 산출하였다.

2.3 동등화 방법들의 비교

우선 동등화되기 이전 점수의 분포와 동등화된 이후 점수의 분포를 비교하여 동등화의 적절성을 검토하였다. 선형 동등화는 동등화되는 두 검사가 평균과 표준편차 이외에는 동일한 적률(moment)을 갖는다고 가정하고, 한 검사의 평균과 표준편차를 기준 검사의 평균과 표준편차에 일치시키는 방식이다. 그러므로 선형 동등화에 의한 동등화 점수는 동등화 이전에는 물론이고, 동등화 이후에도 기준검사와 유사한 왜도와 첨도를 가져야 한다. 등백분위 동등화는 동등화되는 점수를 기준검사의 분포에 의하여 변환시키므로 동등화 이후의 분포가 기준검사와 유사하여야 한다.

다음으로는 선형동등화 방법의 적절성을 검토하기 위하여 동등화 잔차분석이 실시되었다. 동등화에 대한 가정이 잘 만족되는 경우에는 동등화 이전의 점수와 동등화 이후의 점수의 잔차가 0을 중심으로 무작위로 변동하여야 한다. 만약 일정 점수대에서는 양의 잔차가 연속적으로 나타나고 다른 점수대에서는 음의 잔차가 연속적으로 나타난다면, 이는 자료가 동등화를 위한 선형성의 가정을 만족하지 못하며, 비선형 관계식이 적합되어야 할 필요를 나타낸다고 볼 수 있다. 그러므로 잔차의 형태를 통하여 선형 동등화 방법의 적절성이 검토되었다.

빈도추정 동등화 방법과 다른 동등화 방법들에 의한 동등화 점수의 차이가 다음의 세 가지 지수에 의하여 비교되었다. 첫 번째 지수는 차이평균(Mean Signed Difference; MSD)인데 이는 다음과 같이 구해진다.

$$MSD=\frac{\sum_i f_i(A_i-B_i)}{\sum_i f_i} \tag{29}$$

여기서 f_i는 각 점수의 빈도, A_i는 빈도추정 동등화 방법을 제외한 다른 네 가지 방법에 의한 동등화 점수이며, B_i는 빈도추정 동등화 방법에 의한 동등화 점수이다. 이는 빈도추정 동등화와 그 외의 동등화 방법간의 동등화 점수 차이를 각 점수의 관측빈도로 가중평균한 값이다.

두 번째 지수는 절대값 차이평균(Mean Absolute Difference, MAD)인데 이는 다음과 같이 구해진다.

$$MAD=\frac{\sum_i f_i \mid A_i-B_i \mid}{\sum_i f_i} \tag{30}$$

이는 빈도추정 동등화 방법과 다른 동등화 방법간의 동등화 점수 차이의 절대값을 각 점수의 관측빈도로 가중평균한 값이다.

세 번째 지수는 평균제곱오차의 제곱근(Root mean squared error; RMSE)인데 이는 다음과 같이 구해진다.

$$RMSE=\left[\frac{\sum_i f_i(A_i-B_i)^2}{\sum_i f_i}\right]^{1/2} \tag{31}$$

이는 빈도추정 동등화 방법과 다른 동등화 방법간의 동등화 점수 차이의 제곱값을 각 점수의 관측빈도로 가중평균한 값이다. 각 방법에 의한 동등화의 수행과 여러 동등화 방법의 비교를 위한 컴퓨터 프로그램은 SAS와 SAS/IML로 작성되었다.

3. 연구결과

인문계의 수리·탐구II과목은 공통과목 점수 57점과 선택과목 점수 15점을 합하여 72점 만점으로 되어 있다. 선택과목은 정치, 경제, 사회문화, 세계사, 세계지리의 다섯 과목으로 구성되며, 수험생들은 이 중 한 과목을 선택하도록 되어 있다. 이 연구에서 사용한 모의 수능시험에 응시한 총 112,670명의 인문계 수험생들 중 선택과목에서 정치를 선택한 수험생은 31,436명, 경제를 선택한 수험생은 11,965명, 사회문화를 선택한 수험생은 52,372명, 세계사를 선택한 수험생은 9,663명, 그리고 세계지리를 선택한 수험생은 7,234명이었다.

<표 1>은 이들이 각 선택과목에서 획득한 점수의 평균과 표준편차, 그리고 왜도와 첨도이다. 이 자료에 의하면 세계지리를 선택한 학생들의 평균점수가 11.494점으로 가장 높고, 다음으로는 정치, 사회문화, 세계사의 순으로 높은 점수를 받았으며, 경제를 선택한 학생들의 평균점수가 8.148로 가장 낮은 것으로 나타나고 있다.

왜도는 다섯 과목 모두 음수값을 가지고 있는데, 이는 점수분포가 고득점자가 많고 저득점자는 많지 않은 부적편포를 하고 있음을 나타낸다. 그러나 왜도의 값이 가장 작은 세계지리의 왜도가 -0.982로 편포의 정도가 아주 심하지는 않았다. 첨도는 정치와 사회문화는 양수이고 나머지 과목들은 음수인데, 첨도가 양수인 것은 정규분포보다 더 뾰족한 형태이고, 음수인 것은 정규분포보다 납작한 형태인 것을 나타낸다. 경제과목의 첨도가 -0.882로 가장 작았고, 세계지리의 첨도가 0.553으로 가장 컸는데 이것 역시 정규분포에서 아주 심하게 이탈된 정도는 아니었다. 통계패키지 SAS는 첨도의 값에서 3을 제하여 정규분포의 첨도가 0이 되도록 조정한 값을 출력하고 있다.

<표 1> 선택과목 점수의 적률

집단	평균	표준편차	왜도	첨도	N
정치	10.935	3.250	-0.910	0.324	31,436
경제	8.148	3.635	-0.178	-0.882	11,965
사회문화	10.909	3.206	-0.790	0.135	52,372
세계사	10.016	3.843	-0.578	-0.631	9,663
세계지리	11.494	3.129	-0.982	0.553	7,234

각 선택과목을 선택한 응시자 집단이 동일한 학업능력을 가진 무작위 집단(random group)이고, 선택과목별 평균점수의 차이가 오직 각 선택과목의 검사난이도(test difficulty)의 차이에 의하여 발생한 것이라면 이는 표준점수(standardized score)의 사용으로 조정될 수 있다. 즉, 각 선택과목별 평균점수와 표준편차를 사용하여 원점수를 표준화한 표준점수를 수험생들에게 부여하면 검사의 난이도에 상관없이 모든 선택과목은 동일한 평균과 표준편차를 갖게 된다.

<표 2>는 각 선택과목별 공통과목 점수의 평균과 표준편차, 왜도와 첨도를 보여준다. 모든 응시집단이 공동으로 치른 공통과목 점수의 응시집단별 평균차이는 이들 응시집단의 학업능력(trait)이 동일하지 않음을 나타낸다. 이 경우에는 각 집단이 무작위 집단(random group)이 아니라 비동질 집단(nonequivalent group)이 되며, 따라서 선택과목 점수는 검사간 난이도의 조정을 위한 표준점수의 사용

만으로 교정될 수 없고, 응시집단간 학업능력 차이를 고려한 검사점수의 동등화(test equating)가 필요하게 된다.

이때 선택과목별 응시집단간 학업능력의 차이를 측정하는 공통검사(common items)는 집단간 학업능력의 차이를 정확히 반영하기 위하여 선택과목과 검사 내용이 일치하여야 한다. 대학입학수학능력시험의 경우에 인문계와 자연계 모두 선택과목이 공통과목을 구성하는 교과들의 일부이다. 그러므로 선택과목의 교과 내용과 공통과목의 교과 내용이 정확히 일치하지 않으나, 사회탐구나 과학탐구의 선택과목에서 측정하는 학업능력이 수리·탐구II 공통과목에서 측정하는 능력과 크게 다르지 않다고 볼 수 있다.

선택과목별 응시집단간 학업능력 차이를 측정하는 가교검사(anchor test)로 수리·탐구II의 공통과목 이외의 대안을 찾는다면 수능시험 전체의 총점이나 학생생활기록부의 동일과목 성적을 사용할 수도 있을 것이다. 그러나 수능시험 총점은 수리·탐구II의 공통과목에 비하여 선택과목과 교과 내용의 불일치가 심하여 선택과목과 같은 능력을 측정하고 있다고 볼 수 없으며, 학생생활기록부의 동일과목 성적은 자료 수집에 어려움이 있고, 자료를 수집할 수 있다고 하여도 서로 다른 학교간 점수를 비교하는 데 어려움이 있다. 따라서 이 연구에서는 대입수능시험의 수리·탐구II 공통과목 점수를 선택과목 응시집단간 학업능력 차이를 측정하고, 선택과목의 점수를 동등화하는 가교검사로 사용한다.

<표 2> 선택과목별 공통과목 점수의 적률

집단	평균	표준편차	왜도	첨도	N
정치	34.906	8.486	-0.551	0.378	31,436
경제	33.783	9.768	-0.535	-0.036	11,965
사회문화	32.016	9.303	-0.380	-0.159	52,372
세계사	36.460	9.978	-0.827	0.390	9,663
세계지리	36.541	8.858	-0.787	0.778	7,234

<표 3>은 각 동등화 방법에 의하여 산출된 동등화 점수들의 평균, 표준편차, 왜도, 그리고 첨도를 보여준다. 동등화 방법에서 LIN은 한국교육과정평가원에서 사용한 이단계 선형 동등화를, EQUI는 이단계 등백분위 동등화를, Tucker는 Tucker의 합집단 선형 동등화를, FREQ는 빈도추정 동등화를, BH는 Braun-Holland 동등화를 나타낸다. 동등화 점수의 평균을 살펴보면 다섯 가지 동등화 방법들 중에서 LIN, Tucker, FREQ는 평균점수의 순위가 세계지리, 세계사, 정치, 경제, 사회문화의 순으로 나타나 공통과목 점수의 순위와 같도록 조정되었음을 알 수 있다. 그러나 BH는 평균이 높은 순으로 네 번째와 다섯 번째인 경제와 사회문화의 평균점수 순위가 바뀌었으며, EQUI는 두 번째와 세 번째인 세계사와 정치의 평균점수 순위가 바뀐 것으로 나타났다.

공통과목 점수의 순위에서 네 번째와 다섯 번째의 평균점수 차이는 1.767(=33.783-32.016)로 이웃한 순위간의 평균차 중에서 가장 큰 값을 가지며, 두 번째와 세 번째의 평균점수 차이는 1.554(=36.460-34.906)로 이웃한 순위간의 평균차 중에서 그 다음으로 큰 값을 가지고 있다. 그런데도 이들의 순위가 바뀐 것은 이 두 방법이 주어진 자료에 적절하게 작용하지 못하여서 공통과목의 점수에 의한 선택과목별 응시집단의 학업능력 차이를 선택과목 점수에 반영하는 데에 실패하고 있음을 보여주고 있다고 하겠다.

<표 3> 동등화 방법별 동등화된 선택과목점수의 적률

집단	동등화	평균	표준편차	왜도	첨도
경제	LIN	10.504	3.741	-0.178	-0.882
	EQUI	10.452	3.739	-0.799	-0.259
	Tucker	10.645	3.482	-0.178	-0.882
	FREQ	10.602	3.472	-0.824	-0.005
	BH	10.649	3.448	-0.178	-0.882
사회문화	LIN	9.828	3.563	-0.790	0.135
	EQUI	9.740	3.605	-0.599	-0.549
	Tucker	10.191	3.398	-0.790	0.135
	FREQ	10.150	3.596	-0.661	-0.340
	BH	10.569	3.435	-0.790	0.135
세계사	LIN	11.530	3.822	-0.578	-0.631
	EQUI	11.406	3.686	-1.190	0.623
	Tucker	11.353	3.544	-0.578	-0.631
	FREQ	11.322	3.344	-1.052	0.543
	BH	11.212	3.436	-0.578	-0.631
세계지리	LIN	11.561	3.393	-0.982	0.553
	EQUI	11.549	3.281	-1.229	1.088
	Tucker	11.360	3.317	-0.982	0.553
	FREQ	11.327	3.186	-1.059	0.692
	BH	11.169	3.293	-0.982	0.553

표준편차는 동등화 이전과 이후 자료 모두 각 선택과목별로 비슷한 값을 가졌으며, 공통과목 점수와 동등화 이전의 선택과목 점수의 경우에 사회문화의 표준편차가 가장 크고, 세계지리의 표준편차가 가장 작았는데, 다섯 가지 동등화 방법에 의하여 동등화된 점수들이 모두 이러한 경향을 유지하고 있다. 왜도와 첨도는 선형변환을 사용하는 세 가지 동등화 방법들에 있어서는 동등화 이전과 이후의 값이 동일하다. 또한 EQUI와 FREQ의 동등화 이후의 왜도와 첨도는 공통과목의 왜도, 첨도와 유사하여졌고, 특히 동등화 이후의 선택과목의 왜도와 첨도 크기의 순위는 선택과목별 공통과목의 왜도, 첨도 크기의 순위와 거의 일치하였다.

<표 4> 세계사 점수별 공통과목 점수 평균($\overline{V}|X$), 세계사 점수에 의한 공통과목 점수의 선형회귀값($\hat{V}|X$)

점수	N	$\overline{V}\|X$	$\hat{V}\|X$	차이	부호
0.0	78	16.647	16.668	-0.021	-
1.0	18	18.500	18.644	-0.144	-
1.5	214	20.304	19.632	0.672	+
2.0	21	18.262	20.620	-2.358	-
2.5	26	19.269	21.608	-2.339	-
3.0	332	22.548	22.596	-0.048	-
3.5	107	23.238	23.584	-0.346	-
4.0	40	18.288	24.572	-6.285	-
4.5	400	24.465	25.560	-1.095	-
5.0	249	26.669	26.548	0.120	+
5.5	42	21.333	27.536	-6.203	-
6.0	310	28.792	28.524	0.268	+
6.5	462	29.974	29.512	0.462	+
7.0	30	27.167	30.500	-3.334	-
7.5	312	31.793	31.488	0.305	+
8.0	578	33.016	32.476	0.540	+
8.5	41	35.329	33.464	1.865	+
9.0	261	35.322	34.452	0.869	+
9.5	758	35.792	35.440	0.352	+
10.0	34	38.132	36.429	1.704	+
10.5	354	37.682	37.417	0.266	+
11.0	836	38.748	38.405	0.344	+
11.5	52	39.942	39.393	0.550	+
12.0	488	40.49	40.381	0.106	+
12.5	983	41.339	41.369	-0.029	-
13.0	23	41.152	42.357	-1.204	-
13.5	741	43.244	43.345	-0.101	-
14.0	798	44.111	44.333	-0.222	-
15.0	1075	45.899	46.309	-0.410	-

모든 선택과목의 점수는 0점과 만점인 15점을 포함하여 29가지의 가능한 점수가 있으며, 선택과목에서 각 점수를 획득한 수험생들의 공통과목 평균점수를 구할 수 있다. LIN과 Tucker, BH는 공통과목과 선택과목간의 선형관계식을 사용한 동등화를 실시하고 있는데, 만약 선형관계를 가정하는 선형동등화가 적절한 동등화 방식이라면 두 과목 점수간의 선형회귀식과 선택과목의 각 점수를 획득한 수험생들의 공통과목 점수 평균간 차이의 부호가 음과 양이 무작위로 반복되는 경향을 가져야 한다.

<표 4>는 선택과목 중 세계사 과목의 점수별 관측돗수(N), 세계사의 해당점수를 획득한 학생의 공통과목 점수 평균($\overline{V}|X$), 세계사의 점수를 독립변수로 하여 선형회귀식에 의하여 예측한 공통

과목 점수($\hat{V}|X$), 두 값의 차이, 두 값의 차이의 부호를 보여준다. 이 결과에 의하면 차이값의 부호가 음과 양의 값이 무작위로 반복되지 않고 이웃한 점수에서 같은 부호가 연속되는 경향이 있음을 알 수 있는데, 이는 두 검사간에 선형관계가 잘 성립하지 않음을 나타내는 증거로 볼 수 있다. 세계사에서 이러한 현상이 가장 강하게 나타났으나, 다른 선택과목에서도 정도의 차이는 있지만 비슷한 현상이 관측되었다.

<표 5> 일부 원점수에 대한 다섯 가지 동등화 방법들의 동등화 점수

집단	원점수	LIN	EQUI	Tucker	FREQ	BH
경제	0.0	2.118	0.283	2.841	0.442	2.921
	2.5	4.691	3.350	5.235	4.323	5.292
	5.0	7.264	7.872	7.630	8.065	7.663
	7.5	9.837	10.757	10.024	10.794	10.034
	10.0	12.411	12.445	12.419	12.440	12.405
	12.5	14.984	14.625	14.814	14.548	14.777
	15.0	17.557	15.445	17.208	15.416	17.148
사회문화	0.0	-2.297	-0.357	-1.372	-0.306	-1.119
	2.5	0.481	1.455	1.278	1.451	1.559
	5.0	3.260	3.079	3.927	3.318	4.238
	7.5	6.039	5.716	6.577	6.349	6.916
	10.0	8.817	8.179	9.227	9.086	9.595
	12.5	11.596	11.981	11.877	12.224	12.273
	15.0	14.374	14.402	14.527	14.704	14.952
세계사	0.0	1.568	-0.007	2.117	0.313	2.257
	2.5	4.055	2.793	4.422	3.376	4.492
	5.0	6.541	6.679	6.727	7.514	6.728
	7.5	7.231	7.758	7.127	7.517	6.966
	10.0	11.514	12.197	11.338	11.929	11.198
	12.5	14.001	13.728	13.643	13.517	13.433
	15.0	15.363	14.946	15.077	14.784	14.859
세계지리	0.0	-0.902	-0.391	-0.824	-0.208	-0.928
	2.5	1.809	1.545	1.826	1.704	1.703
	5.0	4.520	4.117	4.476	4.570	4.334
	7.5	7.231	7.758	7.127	7.517	6.966
	10.0	9.941	10.302	9.777	9.663	9.597
	12.5	12.652	12.563	12.427	12.401	12.228
	15.0	15.363	14.946	15.077	14.784	14.859

<표 5>는 일부 원점수에서의 다섯 가지 동등화 방식에 의한 동등화 점수이다. 이 결과에 의하면 다섯 가지 방식이 양극단의 값을 제외하고는 원점수에 대하여 거의 같은 동등화 점수를 산출하는 것으로 나타났다. 그러나 LIN과 Tucker, BH와 같이 선형관계식을 사용하는 동등화 방식들은 경제에서 원점수의 만점인 15점에 대하여 만점보다 아주 높은 17점 이상의 점수를 부여하거나,

사회문화나 세계지리에서 원점수에서 0점에 대하여 0보다 아주 낮은 -2점 이하의 점수를 부여하는 것으로 나타났다.

반면에 등백분위 동등화 방식을 사용하는 EQUI나 FREQ에는 이러한 현상이 나타나지 않았다. 등백분위 동등화에 의한 동등화 방식에 나타날 수 있는 최저점은 -0.5점이며, 최고점은 15.5점이다. 최저점 -0.5점은 원점수의 최저점인 0점에서 최저점과 최저점 바로 위 점수간의 간격의 ½을 뺀 값으로 정의되며, 최고점 15.5점은 원점수의 최고점인 15점에 최고점과 최고점 바로 아래 점수간의 간격의 ½을 더한 값으로 정의된다. 이는 이산분포를 하는 점수(discrete integer valued random variable)의 백분위(percentile rank) 산출에 연속화 과정(continuation process)이 적용되었기 때문이다.

<표 6> 빈도추정동등화 방법과 그 외의 동등화 방법들간의 오차

집단	동등화	MSD	MAD	RMSE
경제	LIN	-0.098	0.679	0.791
	EQUI	-0.150	0.233	0.375
	Tucker	0.043	0.586	0.713
	BH	0.047	0.576	0.710
사회문화	LIN	-0.322	0.346	0.403
	EQUI	-0.410	0.413	0.466
	Tucker	0.041	0.246	0.310
	BH	0.419	0.425	0.507
세계사	LIN	0.208	0.642	0.768
	EQUI	0.084	0.356	0.467
	Tucker	0.030	0.491	0.579
	BH	-0.110	0.456	0.553
세계지리	LIN	0.234	0.334	0.380
	EQUI	0.222	0.283	0.339
	Tucker	0.033	0.216	0.253
	BH	-0.158	0.212	0.286

주. MSD는 차이평균(Mean Square Difference),
MAD는 절대값 차이평균(Mean Absolute Difference),
RMSE는 평균제곱오차 제곱근(Root Mean Squared Error)

지금까지의 비교에 의하면 EQUI와 BH는 동등화의 가장 중요한 기능인 응시집단간의 학업능력 차이를 선택과목의 점수에 반영하는 목표에 실패하는 경우가 있는 것으로 나타났다. 또한 선택과목 점수의 동등화에 선형관계를 가정하는 방식들인 LIN, Tucker, 그리고 BH는 만점과 영점을 크게 벗어나는 점수를 부여하는 경우가 있는 것이 관측되었으며, 세계사의 경우에 선택과목 점수와 공통과목 점수간의 관계가 선형이 아닌 것으로 나타났다.

따라서 이 자료에는 합집단에 대하여 등백분위 동등화 방식을 적용하는 FREQ에 의한 동등화가 가장 적절한 것으로 보인다. 이 방식은 계산이 복잡하여 컴퓨터 시간(computing time)이 길지만

이론적으로도 다른 방법들에 비하여 장점을 가지고 있으므로, 이를 기준으로 하여 FREQ와 다른 동등화 방법간의 동등화 점수의 차이를 세 가지 방식으로 계산한 결과가 다음의 <표 6>에 수록되었다.

이 결과에 의하면 경제, 세계사, 세계지리에서는 차이평균(MSD), 절대값 차이평균(MAD), 평균제곱오차 제곱근(RMSE) 모두 LIN이 제일 큰 값을 가졌고, 사회문화에서는 LIN, EQUI, 그리고 BH가 비슷한 크기의 값을 가졌다. 사회문화는 모든 선택과목들 중에서 기준과목인 경제와 가장 유사한 평균, 표준편차, 첨도, 왜도값을 가졌는데, 이러한 경우에는 동등화 방법들의 결과간에 큰 차이가 없어서 나타난 현상인 것으로 생각된다. 전반적으로는 네 가지 동등화 방법들 중에서 한국교육과정평가원이 99학년도의 입시에 사용한 이단계 선형 동등화(LIN) 점수가 FREQ의 동등화 점수와 가장 큰 차이가 있는 것으로 나타났다. 네 가지 방법들 중에서 Tucker가 전반적으로 작은 값을 가져서 그 결과가 FREQ와 가장 근사함을 보여주고 있다.

4. 결 론

이 연구의 결과는 동등화가 이루어지는 응시집단의 합집단을 고려하고 등백분위 동등화를 사용하는 빈도추정 동등화가 이 연구에서 사용한 모의 수능시험 자료에 가장 적합한 방법인 것으로 나타났다. 반면에 99학년도의 입시에서 선택과목 점수의 동등화에 사용된 이단계 선형 동등화는 이 연구에서 비교된 다른 방법들에 비하여 빈도추정 동등화의 결과와 가장 큰 차이를 보이는 것으로 보고되었다.

빈도추정 동등화에서 집단 1은 선택과목 X의 점수에 대한 자료만을 가지고 있고, 집단 2는 선택과목 Y의 점수에 대한 자료만을 가지고 있다. 그러므로 합집단의 X 점수와 Y 점수의 결합밀도함수를 추정하기 위하여 공통과목(V)의 점수가 v로 주어졌을 때, 선택과목 X의 점수와 선택과목 Y의 점수의 조건부 확률밀도함수가 두 집단에서 동일하다는 것을 가정한다. 따라서 두 집단의 점수가 상당히 다를 때는 이 연구의 결과로 추천된 빈도추정 동등화 방법은 적절하지 못하며, 그러한 경우에는 진점수 모형(true score model)에 근거한 Levine 동등화(Levine linear method)나 문항반응이론에 근거한 방법들(Item response methods)을 사용하여야 한다. 그러나 집단간의 차이가 클 때는 어떠한 방법도 검사 점수의 동등화에 성공적이지 못한 것으로 알려져 있다.

검사 점수의 동등화에 가장 많이 사용되는 설계는 무작위집단 설계(random group design)와 공통문항 비동질집단 설계(common-item nonequivalent groups design)이다. 무작위집단 설계는 수험생을 두 개의 무작위 집단으로 분할하거나, 분할되어 있는 집단들이 무작위로 분할된 동질한 집단이라고 간주하고, 각 집단에 대하여 서로 다른 형태의 검사지를 사용하여 검사를 수행하는 방식이다. 이 설계에서는 두 집단이 동질하다고 가정하기 때문에 검사 점수의 차이는 오직 검사지의 난이도 차이에 의하여 발생한 것으로 생각한다. 그러나 대학입학수학능력시험의 경우에는 선택과목별로 응시자를 무작위 할당할 수 없기 때문에 집단간 학업능력에 차이가 있을 수 있으며, 이를 무시하고 선택과목별 집단을 무작위집단으로 간주하게 되면 각 집단의 학업능력차에 따라 수험생의 점수가 불공정하게 평가될 수 있다. 이러한 경우에 선택과목의 점수차이를 집단간 능력차에 의한 부분과 검사지 난이도 차이에 의한 부분으로 분할하는 것이 이 연구에서 사용된 공통문항 비동질집단 설계를 사용하는 가장 큰 목적이다.

공통문항 비동질집단 설계는 내부 공통문항(internal common items)을 사용하는 경우와 외부

공통문항(external common items)을 사용하는 두 가지 종류가 있으며, 내부와 외부는 공통문항의 점수가 총점에 포함되는가, 아니면 오직 동등화의 목적으로 사용되는가에 의하여 구분된다. 따라서 대입수능시험의 공통과목은 내부 공통문항에 해당되며, 이때 내부 공통문항의 조건은 우선 선택과목과 교과 내용이 같아야 하고, 다음으로는 선택과목과 검사지의 통계적 특성이 같아야 한다는 것이다. 즉, 공통문항은 검사를 대표하는 작은 검사(mini version)가 되어야 한다. 이 연구에서는 이러한 이유로 수리·탐구II의 공통과목을 공통문항으로 사용하였다. 앞에서 기술한 바와 같이 공통과목은 검사가 포함하는 교과 내용이 선택과목과 완전히 일치하지는 않지만, 공통문항으로 사용될 수 있는 조건에 공통과목이 가장 적절한 것으로 보이며, <표 1>과 <표 2>에 의하면 공통과목의 통계적 성질은 선택과목과 대체로 유사한 것으로 나타나 이를 공통문항으로 사용하는 데에 큰 문제가 없는 것으로 생각된다.

동등화 방법은 선형동등화와 등백분위 동등화로 대별되는데, 선형동등화 방법은 자료에 대하여 선형관계의 강한 가정을 부과하며, 만점 이상과 영점 이하의 점수를 부여하는 경우가 있다는 단점이 있다. 이 연구의 결과에서도 선형동등화는 선형관계의 가정이 잘 성립하지 않으며, 선택과목에 따라서 영점 이하나 만점 이상의 점수를 부여하는 것으로 나타나고 있다. 한편 등백분위 동등화 방법은 분포함수에 어떤 모수적 가정도 하지 않는 일종의 비모수적 방법(nonparametric method)이나, 계산이 복잡하고 오래 걸려서 높은 컴퓨터 기능이 요구된다는 단점이 있다. 그러나 등백분위 동등화의 단점인 컴퓨터 시간(computing time)의 문제는 컴퓨터 발달로 인하여 점차 큰 문제가 되지 않고 있어서 이의 유용성이 점차 증가하고 있다.

등백분위 동등화 방법의 적용에 가장 큰 문제가 되는 것은 관측돗수가 작은 최저점이나 최고점 근처의 동등화 함수의 곡률이 불규칙할 수 있다는 점이다. 그런데 이 연구에서 분석된 모의 수능시험 자료는 전반적으로 선택과목의 난이도가 낮아서 최고점의 관측돗수가 상당히 많은 것으로 나타났다. 사회문화, 세계사, 세계지리의 경우에는 만점이 가장 관측돗수가 많은 점수였고, 정치도 만점이 두 번째로 관측돗수가 많은 점수였다. 따라서 등백분위 동등화 방식이 가지고 있는 약점으로 지적되고 있는 관측돗수의 부족으로 인한 오차는 최고점 근처에서는 심각한 문제가 되지 않았다. 반면에 최저점 근처에서는 관측돗수가 많지 않았는데, 대학입학수학능력시험의 용도상 최저점 근처의 동등화 점수의 정확도는 최고점 근처의 동등화 점수의 정확도보다 상대적으로 큰 문제가 되지 않을 것으로 생각된다. 또한 한국교육과정평가원은 1999학년도의 대학입학수학능력시험을 상위 50% 집단의 영역별 예상 평균 점수가 100점 만점의 60~70점 정도가 되도록 출제하겠다고 발표하였다. 따라서 수능시험이 앞으로 계속 쉽게 출제될 가능성이 많다고 볼 때, 앞으로 등백분위 방식을 사용한 수능시험 점수의 동등화가 선형 동등화보다 유용한 방법이 될 수 있을 것으로 생각된다.

이 연구는 분포의 모양과 응시자의 수에서 차이가 나는 다섯 개의 선택과목에 대하여 동등화가 실시되었는데, 선택과목에 따라 동등화 방법의 우열이 조금씩 다르게 나타났다. 이 연구와 같은 실증자료의 분석으로는 이러한 현상의 원인을 규명하는 데 한계가 있을 것으로 보이며, 이후에 동등화 방법에 대한 모집단의 분포, 공통과목과 선택과목의 상관계수의 크기, 모집단의 크기, 동등화되는 선택과목들의 분포의 이질성 등의 효과에 대한 연구가 모의실험에 의하여 수행될 필요가 있는 것으로 생각된다.

참고문헌

[1] 남현우(1992a). 검사동등화에 대한 이론적 고찰, 교육학연구, 30(2), 205-221.

[2] 남현우(1992b). 문항모수의 변이에 따른 선형, 동백분위, IRT, 검사동등화 방법의 강인성 비교 연구, 교육평가연구, 5(2), 27-60.

[3] 성태제(1994). 대학별 고사를 위한 문항분석, 표준점수, 검사동등화, 한국통계학회논문집,1(1), 206-214.

[4] 허명회(1995). 선택시험 등화를 위한 통계적 방법론, 성곡논총 26, 949-983.

[5] Angoff, W. H.(1971). Scales, norms, and equivalent scores, In R. L. Thorndike(Ed.) Educational measurement, American Council on Education, Washington, D. C. 508-600.

[6] Braun, H. I. and Holland, P. W.(1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin(Eds.) Test equating, 9-49. New York: Academic Press.

[7] Gulliksen, H.(1950). Theory of mental tests. New York: Wiley.

[8] Harris, D. J., and Kolen, M. J.(1990). A comparison of two equipercentile equating methods for common item equating, Educational and psychological measurement, 50, 61-71.

[9] Jarjoura, D., and Kolen, M. J.(1985). Standard errors of equipercentile equating for the common item nonequivalent populations design, Journal of educational statistics, 10, 143-160.

[10] Kolen, M. J., and Brennan, R. L.(1995). Test equating: Methods and practices, Springer.

[11] Kolen, M. J., and Jarjoura, D.(1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design, Psychometrika, 52, 43-59.

[12] MacCann, R. G.(1989). A comparison of two observed-score equating methods that assume equally reliable, congeneric tests, Applied psychological measurement, 13(3), 263-276.

[13] Wang, T., and Kolen, M. J.(1996). A quadratic curve equating method to equate the first three moments in equipercentile equating, Applied psychological measurement, 20(1), 27-43.

[14] Woodruff, D. J.(1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design, Applied psychological measurement, 13(3), 257-261.