교육과정평가연구
The Journal of Curriculum and Evaluation
2012, Vol. 15, No. 3, pp. 277~294

혼합형태의 공통+행렬표집 가교문항설계의 베이지안 IRT 동등화 방안[1)]

남 현 우(순천향대학교 교수)*

≪ 요 약 ≫

가교 문항의 노출 위험을 적게 하면서 피험자의 수험 부담을 줄일 수 있도록 하는 공통+행렬 표집 가교 문항 설계(Common plus Matrix-sampled Anchor Items Design)에서 사용 가능한 문항반응이론(IRT)의 동등화 방안을 확장하려는 목적으로 본 연구가 수행되었다. 기존의 최대우도추정(MLE) 방식에 의한 동등화 방안들을 베이지안 사전 정보(Bayesian Informative Priors)를 활용하는 새로운 IRT 동등화의 부분 집합으로 보면서, 행렬 표집 설계에서 가장 안정적인 결과를 나타내는 베이지안 IRT 동등화 방안을 찾아보려 했다.

고전적인 가교 문항 설계로 수집된 국가수준 학업성취도 평가의 고1 국어 시험 자료를 최대 6개의 구획으로 행렬 표집되는 자료로 변형하였다. 구획 수가 많아지고 또 구획 당 피험자 수가 적어지면서 모수 추정 및 동등화에 어느 정도의 오차가 발생하는지를 알아보았다. 또한 사전 정보를 제시하는 방식에 따라 모수 추정 및 동등화에 어떤 영향이 있는지를 알아보았다.

MLE 방식의 문항모수고정(FPIP) 방법은 점-사전 정보(Point Prior)를 활용한 베이지안 IRT 동등화로, 그리고 특성곡선전환(CCT) 방법은 사전 정보가 없는(Flat Prior) 베이지안 IRT 동등화와 같은 것으로 보고, 이에 덧붙여 사전 정보의 정도를 달리하는 MCMC 방식의 두 가지 베이지안 IRT 동등화 방법들(즉, Informative Prior, High Informative Prior)을 적용해보았다.

사전 정보의 활용 정도가 지나치게 높거나(Point Prior) 낮은 경우(Flat Prior)보다는 어느 일정 수준의 사전 정보(Informative Priors)를 활용할 때 안정적인 동등화 결과를 나타낼 것이라는 가정을 지지할만한 명확한 증거를 찾을 수 없었다. 가장 안정적인 결과를 보인 점-사전 정보 방법의 한계와, 최적의 사전 정보 수준을 찾는 후속 연구의 필요성 등을 논의했다.

주제어 : 행렬표집설계, 사전정보, 베이지안 IRT 동등화, 최대우도추정, MCMC

1) 이 논문은 2010학년도 순천향대학교 교수 연구년제에 의하여 연구하였음.
* 제1저자 및 교신저자, namhw@sch.ac.kr

Ⅰ. 연구의 필요성 및 목적

국가 교육과정을 평가하거나 국가 간 학력 비교 등을 목적으로 하는 대규모 시험의 경우 각 개인이 치러야 할 문항 수가 많기 때문에 동기 수준이 높지 않은 어린 학생들의 능력 수준을 신뢰롭고 타당하게 평가하는데 어려움이 많다. 이럴 때 학생들의 능력을 측정하는데 지장이 없는 한도 내에서 이들의 수험 부담을 줄일 수 있는 정교한 시험 설계가 필요하다. 학생들의 능력 수준을 측정하기 위한 문항들은 모든 수험생들에게 공통으로 부과하고, 시험 유형 간 비교 가능성을 확보하기 위한 가교 문항들은 일부 수험생들에게만 부과하는 '공통+가교문항 행렬표집설계'(Massachusetts Department of Education, 2002; Driscoll, 2002; Hu, Rogers, & Vukmirovic, 2008)는 대규모 시험에서 수험 부담을 적게 하면서 동등화 오차를 최소화 할 수 있는 시험 설계 중 하나라고 여겨진다. 그런데 문제는 이런 식의 시험 설계로 수집된 자료를 가지고 동등화 할 수 있는 기존의 방안들이 제한돼 있다는 것이다. 행렬 표집된 가교 문항에는 전체 피험자 중 일부만 반응했기 때문에 기존의 고전검사이론에 기초한 동등화 함수를 산출하는 게 쉽지 않다. 그러나 문항반응이론의 경우에는 가능한 방안들이 있다. 모든 피험자들에게 모든 가교 문항들을 부과했으나 특정 구획에 있는 가교 문항들을 제외한 다른 문항들에는 응답하지 않은 것으로(not responded) 가정해서 모수를 추정할 수 있다는 문항반응이론의 특징을 활용하는 것이다.

행렬 표집 설계에서 활용 가능한 문항반응이론의 동등화 방안은 두 가지 정도 있다. 서로 다른 척도에서 구한 두 모수 추정치들은 일정한 선형 관계를 갖는다는 문항반응이론의 특징을 활용하여, 두 문항 반응 함수의 차이가 최소화 되는 점에서 선형 관계식을 찾는 방안이 있다. 이른바 Stocking & Lord(1983)의 특성곡선전환(CCT: Characteristic Curve Transformation) 방법이 그것이다. 다른 하나는 가교 문항의 모수 값은 언제나 같아야 한다는 점을 활용하여, 이미 얻은 가교 문항의 모수 추정치를 고정한 상태에서 동등화 할 검사의 모수를 추정하는 방법이다. 이른바 문항모수고정(FPIP: Fixed Pre-calibrated Item Parameter) 방법이 그것이다.

그러나 최대 우도 추정(MLE) 방법을 근간으로 하는 이들 동등화 방법들만으로는 안정적인 동등화 결과를 기대하기 어려운 상황이 많다. 가교 문항의 노출 가능성을 가능한 한 줄이기 위해서는 구획 수를 늘려야 하는데, 그러다 보면 각 구획 당 응답자 수가 줄어들고, 결과적으로 문항반응이론을 활용하여 문항 및 능력 모수를 추정하기 어려운 지경에 이를 수 있다(남현우, 2002; 남현우 외, 2009).

기준 검사의 가교 문항 모수 특성들을 사전 정보(Priors)로 활용하여 동등화 검사의 모수를

추정하는 베이지안 IRT 동등화(Bayesian IRT Equating)가 좋은 대안이 될 수 있다(Lord, 1986; de la Torre & Patz, 2001). IRT 동등화 방법 중 가장 안정적인 결과를 나타내는 것으로 알려진 특성곡선전환(CCT) 방법은 모수를 추정하는 과정과 동등화 함수를 만드는 두 과정을 거쳐야 하기 때문에 이미 발생한 오차의 원천이 무엇인지를 규명하거나 통제하기가 쉽지 않다는 단점이 있다. 그리고 이미 알고 있는 가교 문항의 모수 추정치를 활용하는 문항모수고정(FPIP) 방법도 검사 맥락 효과 즉, 가교 문항들이 기준 검사와 동등화 검사 내에서 각각 처한 위치가 다를 때 동등화 결과가 다를 수 있다는 단점을 안고 있다. 그런데 베이지안 IRT 동등화는 최대 우도 추정(MLE) 방법을 사용하기 때문에 피할 수 없었던 이들의 단점들도 사전 정보의 활용으로 극복할 수 있다는 특징이 있다.

MCMC(Markov Chain Monte Carlo) 추정 기법의 발달에 따라 포괄적이며 통합적으로 문항반응이론에 의해 모수를 추정하고 동등화 할 수 있게 되었다(Patz & Junker, 1999a, 1999b). 베이지안 IRT 동등화는 기본적으로 베이지안 통계를 바탕으로 하기 때문에 사전 정보에 따라 사후 분포로서의 동등화 결과가 달라진다. 모든 피험자가 모든 가교 문항을 치르는 고전적인 가교 검사 설계가 아니라, 일부 피험자들만이 특정 구획의 가교 문항들을 치른 행렬 표집 가교 문항 설계에서 어느 정도의 사전 정보가 적절한지를 알아야 한다. 행렬 표집 방식으로 가교 문항들을 처리하는 동등화 자료 수집 설계에서 구획 수가 많아지고 구획 당 피험자 수가 줄어들 때 동등화 오차는 얼마나 커지는지 등을 알아야 한다.

대규모 시험에서 피험자의 수험 부담을 줄이면서 가교 문항의 노출 가능성을 낮추기 위해서는 동등화 자료 수집 설계가 정교화 되어야 할 것이다. 고전적인 시험 설계에서 적절하게 기능했던 기존의 방법들만으로는 만족할 만한 동등화 결과를 얻기 어려울 것으로 본다. IRT 동등화의 외연을 확대한다는 목적을 갖고 본 연구를 통해, 시험 설계가 복잡해짐에 따라 동등화 자료를 최적의 상태에서 수집하지 못했을 때에도 안정적인 동등화 결과를 얻을 수 있는 최적의 사전 정보 수준을 알아보려 한다.

Ⅱ. 이론적 배경

Lord(1986, p. 158)에 따르면, 대략적으로 평행인 시험 유형을 새로운 연도에 표집된 유사한 피험자 집단에게 실시하였다면, 전년도 검사 결과로부터 문항 및 피험자 능력 추정을 위한 사전 정보를 연역하는 것이 가능하다고 볼 수 있다(van Rijn & Beguin, 2009, 2010).

Patz & Junker(1999a)는 복잡한 모형에 특별히 잘 적합화 되는 MCMC 기법을 통해 2-

모수 문항반응이론에 따른 문항 및 피험자 모수를 동시에 추정할 수 있다는 것을 예시했다. 이어서 그들(Patz & Junker, 1999b)은 MCMC 기법을 이용한 베이지안 추론을 문항반응이론의 다양한 장면으로 확대했다. 즉, 무 반응이 발생한 경우, 계획적인 탈락이 있는 경우, 다수 평정자가 존재하는 경우, 추측 행동이 두드러진 경우 및 다분 채점 상황 등이 포함된 문항반응이론 모형에까지 베이지안 추론이 확장될 수 있음을 보였다.

de la Torre & Patz(2001)는 동등화 관계의 추정을 모수 추정 단계에 포함시키는 대안적인 IRT 동등화 방안을 제안하고 탐구하였다. 기준 검사로부터 얻은 문항 모수들을 사전 정보로 활용하여 새 검사의 문항 및 피험자 모수를 추정하였다. 그들은 사전 정보의 수준을 다양하게 설정하면서 기존 IRT 동등화 방안들을 포함시켰다. 즉, 무-사전 정보(Flat Prior)에서부터 점-사전 정보(Point Prior)에 이르는 다양한 수준의 사전 정보를 설정했다. 특성곡선전환(CCT) 방법은 동등화 검사의 모수를 추정할 때 기준 검사로부터 얻은 사전 정보를 전혀 사용하지 않기 때문에 무-사전 정보(Flat Prior)의 한 예에 해당하고, 문항모수고정(FPIP) 방법은 점 추정치를 사전 정보로 활용한다는 점에서 사전 정보 활용의 극단적인 경우(Point Prior)에 해당한다. 그들은 이러한 극단적인 사전 정보 수준 사이에, 기준 검사에서 얻은 가교 문항 모수의 분포와 동일한 수준의 사전 정보를 가정하는 사전 정보(Informative Prior), 기준 검사의 분포보다 변산도가 작은 사전 정보를 가정하는 고-사전 정보(High Informative Prior) 등을 설정하고, 가장 안정적인 동등화 결과를 나타내는 사전 정보 유형을 밝히려 했다.

van Rijn & Beguin(2009, 2010)은 장기적인 학력 추이를 조사하는 평가 연구에서 베이지안 IRT 동등화가 적용될 수 있을지를 탐구했다. 가교 문항을 노출 없이 장기간 활용하는 것이 쉽지 않고 그렇다고 해서 가교 피험자를 확보할 수도 없는 상황에서 동등화 함수를 도출하는 적절한 방안을 찾기 어렵다. 그러나 장기적으로 볼 때 매년 시험을 치르는 피험자들은 다르지만 이들의 능력 수준이 해마다 크게 달라지지 않을 것으로 가정할 수 있기 때문에 학력 수준 설정에 베이지안 IRT 동등화가 적용될 수 있다고 보았다.

Kim & Kang(2012)은 문항 모수를 고정하는 방식으로 동등화 할 때, 기존의 최대 우도 방식(MMLE/EM)과 MCMC 기법에 의한 베이지안 방식을 사용한 동등화 결과 간에 차이가 없음을 보였다. 이들은 MMLE/EM 알고리즘을 이용한 모수 추정 및 동등화가 가능한 검사 상황이 한정적인데 반해, MCMC 기법을 활용한 베이지안 방식은 복잡한 검사 상황으로 적용 가능성이 확장될 수 있다는 점을 감안할 때, 매우 의미 있고 가치 있는 결과를 얻었다고 평가했다.

Ⅲ. 연구 내용 및 방법

1. 연구 문제

본고를 통해 규명하고자 하는 연구 문제는 두 가지다. 하나는 전혀 사전 정보가 없는 것에서부터 완전한 사전 정보에 이르는 네 가지 사전 정보 유형(Flat Prior, Point Prior, Informative Prior, High Informative Prior)에 따라 동등화 결과에 차이가 있는지를 알아보는 것이다. 다른 하나는 행렬 표집 설계에서 구획 수를 늘리고 구획 당 피험자 수가 줄어들 때 가장 안정적으로 동등화 결과를 보이는 사전 정보 유형이 무엇인지를 밝히는 것이다.

- 가장 안정적인 동등화 결과를 나타내는 사전 정보 유형은 무엇인가?
- 구획 수의 변화에도 동등화 결과에 차이가 없는 사전 정보는 무엇인가?

2. 검사 자료

행렬 표집 가교 문항 설계로 시행된 검사 자료를 구하기 어렵기 때문에 고전적인 가교 문항 설계로 시행된 국가수준 학업성취도 평가 2006년-2007년(박정, 2006) 국어 시험 자료를 최하 2개에서 최고 6개 구획에 이르는 행렬 표집 설계로 변형하여 사용했다.

기준이 되는 2006년도 국어 시험 자료는 동등화를 목적으로 별도 표집된 4,680명에게만 실시된 것이다. 이 시험은 30개의 선택형과 10개의 수행형 문항으로 이루어져 있는데, 그 중 9개의 선택형과 3개의 수행형 문항들은 가교 문항들로서 2007년 시험에도 포함되었다.

동등화될 2007년도 국어 시험 자료는 전국의 고1 학생들을 대표하는 표본 24,932명에게 실시된 것이다. 이 시험도 30개의 선택형과 10개의 수행형 문항들로 구성돼 있다.

3. 검사 설계 변형

국어 시험을 2개부터 6개 구획으로 이루어진 행렬 표집 설계로 변형했는데, 그 중 6개의 구획으로 이루어진 행렬 표집 설계에서 가교 문항을 아래 〈표 Ⅲ-1〉과 같이 배정했다. 구획 1에는 선택형 문항 5번, 10번, 12번을 배정하고, 구획 2에는 선택형 문항 13번, 14번, 19번을 배정했으며, 수행형 문항 1번은 구획 4에, 2번과 3번 문항은 각각 구획 5, 6에 배정하였다. 국어 시험의 가교 문항에서 추정되는 난이도 모수가 17개인데, 이들을 6개의 구획에 골고루 배정하기 위해 한 구획에 3개씩(단, 구획 4에는 2개만) 나누어지도록 했다(남현우 외, 2009).

〈표 III-1〉 국어 6구획 행렬표집설계

전체 피험자	2007년 고유문항	행렬표집가교문항	구획
24,932명	선택형 21문항, 수행형 7문항	선택 5, 10, 12	4,156명
		선택 13, 14, 19	4,156명
		선택 22, 23, 24	4,155명
		수행 1	4,155명
		수행 2	4,155명
		수행 3	4,155명
		선택형 9문항, 수행형 3문항	

4. IRT 모형

선택형 문항은 2-모수 로지스틱 모형을, 수행형 문항은 F. Samejima의 등급반응모형(Graded Response Model)을 가정했다. 선택형 문항에서 추측에 의해 정답을 할 가능성이 있음에도 불구하고 3-모수 모형을 가정하지 않은 이유는, 선택형과 수행형 문항이 혼합된 시험에서 두 모형을 동시에 적용할 때 추측 모수 추정의 오차가 매우 컸기 때문이다. 2-모수 모형은 채점 범주가 두 개인 등급 반응 모형의 한 예에 불과하다고 보고 선택형과 서답형 문항들을 동시에 묶어서 분석하였다.

등급 반응 모형에서 능력이 θ_i인 피험자가 문항 j의 범주 k 또는 그 이상의 범주 점수를 얻을 누적 확률을 아래 (1)식 P_{ijk}^{*}로 표현할 수 있다. K_j는 범주 수, a_j는 문항 변별도, b_{jk}는 범주 2로부터 K_j에 이르는 범주의 난이도 (또는 위치) 모수이며, D는 척도화 상수로서 1.7이 보통이다.

$$P_{ijk}^{*} = P_{jk}^{*}(\theta_i) = P^{*}(\theta_i \diagdown a_j, b_{jk}) = \begin{cases} 1 & k = 1 \\ \dfrac{\exp[D a_j(\theta_i - b_{jk})]}{1+\exp[D a_j(\theta_i - b_{jk})]} & 2 \le k \le K_j \\ 0 & k > K_j \end{cases} \quad (1)$$

$$P_{ijk} = P_{jk}(\theta_i) = P_{ijk}^{*} - P_{ij(k+1)}^{*} \quad (2)$$

범주 반응 함수 P_{ijk}는 식 (2)와 같이 인접한 두 누적 함수의 차이로 계산한다. 첫 번째 범주 즉, k가 1일 때의 범주 반응 함수(P_{ij1})는 $P_{ij1} = 1 - P^{*}_{ij2}$이고, 마지막 범주 즉, k가 K_j일 때의 반응 범주 함수(P_{ijK_j})는 $P_{ijK_j} = P^{*}_{ijK_j}$이다.

5. 모수 추정 및 동등화

모수 추정을 위해서는 WinBUGS 1.4 프로그램(Spiegelhalter et al., 2003)에 내재된 MCMC 알고리즘을 이용하였다. 2-모수 로지스틱 모형과 등급 반응 모형의 모수 추정을 위한 WinBUGS 코드는 Curtis(2010)를 바탕으로 연구자가 혼합 문항 형태에 맞게 수정하여 작성했다.

기준 검사인 2006년도 시험의 모수를 추정하기 위한 사전 분포로서 변별도 모수는 $N(1,1)$, 난이도 모수와 능력 모수는 $N(0,1)$을 가정했다. 15,000번 씩 반복하여 사후 분포를 생성했는데, 그 중 첫 5,000번의 값은 'burn-in'으로 버리고 나머지 10,000번의 값을 사용했다. 대체로 5,000번 정도의 반복이 이루어진 후에는 'autocorrelation'이 .0으로 낮아졌고, 'BGR'이 1.0으로 안정되었고, 3개의 체인으로 확인한 'trace'도 특이한 돌출 흔적 없이 평범한 양상을 보였다(박찬호, 강태훈, 2011; Fox, 2010; Kim & Bolt, 2007; Cowles, 2004).

베이지안 IRT에서 모수 추정과 모수 동등화는 하나의 과정에서 이루어진다. 다만, 사전 정보가 전혀 활용되지 않는 경우(Flat Prior)에는 기존의 IRT 동등화 방식을 따를 수밖에 없었다. 즉, 2007년도 시험 자료를 가지고 MCMC 알고리즘을 이용한 WinBUGS 프로그램으로 모수를 추정한 후, 특성곡선전환(CCT) 방식(Kim & Kolen, 2004; Kim & Lee, 2004)으로 2006년도 기준에 모수 동등화를 수행하였다. 이 경우의 사전 정보는 기준 검사인 2006년도 검사와 마찬가지로 변별도 모수는 $N(1,1)$, 난이도 모수와 능력 모수는 $N(0,1)$을 가정했다.

완전한 사전 정보를 활용하는 경우(Point Prior)에는 2006년도 검사에서 추정한 가교 문항 모수 추정치를 2007년도에 그대로 고정하였다. 가교 문항을 제외한 채점 문항들의 사전 정보 역시 변별도 모수는 $N(1,1)$, 난이도 모수는 $N(0,1)$을 가정했다. 이 방식은 최대 우도 추정 방식의 문항모수고정(FPIP) 방식의 동등화와 다르지 않다.

사전 정보를 활용하는 경우(Informative Prior)에는 2006년도 검사에서 추정한 가교 문항 모수 추정치들의 분포와 유사한 값을 사전 정보로 활용했다. 12개 가교 문항 변별도 모수 추정치는 $N(\widehat{\mu_a}, \widehat{\sigma_a^2})$, 난이도 모수 추정치는 $N(\widehat{\mu_b}, \widehat{\sigma_b^2})$를 가정했다. $\widehat{\mu_a}$는 12개 가교 문항의 변별도 사후 분포의 평균값들의 평균과 분산을, $\widehat{\mu_b}$는 12개의 가교 문항의 17개 난이도 사후 분포의 평균값들의 평균과 분산을 상위 사전 정보(Hyper Prior)로 지정했다. $\widehat{\sigma_a^2}$ 와 $\widehat{\sigma_b^2}$ 는 12개

가교 문항의 변별도와 난이도 사후 분포의 변산도에 맞는 감마(Gamma) 분포를 상위 사전 정보로 지정했다. 다만, 감마 분포를 결정하는 α, β 값 중에서 α가 2.0인 분산 분포를 가정했다. 이들을 정리하면 〈표 Ⅲ-2〉와 같다.

〈표 Ⅲ-2〉 2007년도 시험의 베이지안 IRT 동등화를 위한 사전 정보들

동등화 방식	모수	사전정보	상위사전정보
무-사전정보	θ	$N(0,1)$	
	a	$N(1,1)$	
	b	$N(0,1)$	
점-사전정보	θ	$N(0,1)$	
	a	2006년추정치	
	b	2006년추정치	
사전정보	θ	$N(0,1)$	
	a	$N(\widehat{\mu_a}, \widehat{\sigma_a^2})$	$\widehat{\mu_a} \sim N(.9311333, .084)$, $\widehat{\sigma_a^2} \sim G(2.0, .0013604)$
	b	$N(\widehat{\mu_b}, \widehat{\sigma_b^2})$	$\widehat{\mu_b} \sim N(-.0589018, .800)$, $\widehat{\sigma_b^2} \sim G(2.0, .0011622)$
고-사전정보	θ	$N(0,1)$	
	a	$N(\widehat{\mu_a}, \widehat{\sigma_a^2})$	$\widehat{\mu_a} \sim N(.9311333, .0084)$, $\widehat{\sigma_a^2} \sim G(2.0, .00013604)$
	b	$N(\widehat{\mu_b}, \widehat{\sigma_b^2})$	$\widehat{\mu_b} \sim N(-.0589018, .0800)$, $\widehat{\sigma_b^2} \sim G(2.0, .00011622)$

높은 수준의 사전 정보를 활용하는 경우(High Informative Prior)에는 사전 정보(Informative Prior)인 경우와 비교 하여 상위 사전 정보의 변산도를 0.1배 작게 했다. 즉, 변별도 평균 분포의 분산을 .084에서 .0084로, 난이도 평균 분포의 분산을 .80에서 .08로 작게 했고, 변별도 분산에 관한 감마 분포의 β 값을 .0013604에서 .00013604로, 난이도 분산에 관한 감마 분포의 β 값을 .0011622에서 .00011622로 작게 했다. 변별도와 난이도 사후 분포의 분산에 관한 감마 분포의 α 값은 여전히 2.0으로 동일하게 하였다.

베이지안 IRT 동등화로 모수 추정치들이 동일 척도에 놓인 후에는 진 점수(True Scores) 방법으로 검사 점수 동등화(Kolen, 2004)를 수행했다.

6. 동등화 결과 평가

사전 정보의 유형에 따라 동등화 결과가 달라지는지를 알아보고, 행렬 표집 설계의 구획 수가 늘어나고 구획 당 응시자 수가 줄어들 때에도 여전히 안정적인 동등화 결과를 보이는지를 확인

하기 위해 두 가지 통계 방식을 사용하여 평가했다. 하나는 식 (3)과 같은 평균자승편차(RMSD: Root Mean Squared Difference)이고, 다른 하나는 식 (4)와 같은 여러 종속 표본 집중 경향의 차이에 관한 프리드먼(M. Friedman) 이원 분산분석이다.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=0}^{55}(T_i - \widehat{T}_i)^2}{56}} \tag{3}$$

사전 정보의 유형을 달리하는 새로운 베이지안 IRT 동등화 방법들이 안정적인 결과를 나타내는지를 알기보기 위한 평균자승편차의 준거 점수(T_i)는 동일한 구획 수의 특성곡선전환(CCT) 동등화 결과로 하고, 동등화 점수들($\widehat{T}_i$)은 다양한 사전 정보들을 사용해서 얻은 동등화 점수들로 했다. 그리고, 수획 수와 구획 당 응시자 수가 달라져도 여전히 안정적인 동등화 결과를 나타내는지를 알아보기 위한 평균자승편차의 준거 점수는 고전적인 가교 문항 설계(무-구획 설계)에서 각 사전 정보 유형에 따라 시행한 동등화 결과로 하고, 각각의 동등화 점수들($\widehat{T}_i$)은 특정 사전 정보 하에서 구획 수를 달리하면서 얻은 동등화 점수들로 했다.

프리드먼 검정은 n개의 실험 대상을 J번 반복 측정한 임의 구획 설계나 반복 측정 설계에서 F 검정의 대안 검정 방법이다(이종성 외, 2009, p. 921). 프리드먼 검정을 하려면 n개의 각 사례에 대해 J번의 측정값은 1부터 J까지 작은 크기부터 등위를 정한다. J개 측정값에 대하여 등위 합 R_j를 산출하여 아래 식 (4)로 주어지는 χ^2 검정을 한다.

$$\chi^2 = \frac{12}{nJ(J+1)}\sum_{j=1}^{J} R_j^2 - 3n(J+1) \simeq \chi^2(J-1) \tag{4}$$

이 때, n은 0점에서 55점에 이르는 정수 점수의 개수로서 56이고, J는 진점수를 산출한 방식의 수로서 사전 정보 유형의 개수 4개 또는 구획 수가 다른 설계의 개수 6개이다. R_j는 j 측정의 등위 합이다.

프리드먼 검정에서 "J번의 측정값 사이에는 차이가 없다"는 영 가설이 기각된다면, 사전 정보 유형이 다른 동등화 결과 간에 또는 구획 수가 다른 동등화 결과 간에 차이가 있음을 의미한다.

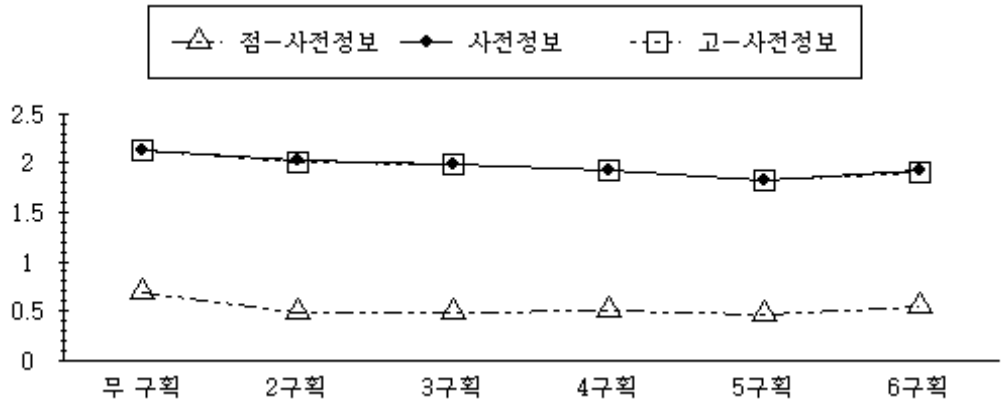
Ⅳ. 연구 결과

1. 사전정보유형의 영향

첫 번째 연구 문제(가장 안정적인 동등화 결과를 나타내는 사전 정보 유형은 무엇인가?)를 규명하기 위해 고전적인 IRT 동등화 방법 중 가장 양호한 것으로 알려진 특성곡선전환(CCT)을 준거로 한 평균자승편차(RMSD)를 계산하여 아래 〈표 Ⅳ-1〉과 〔그림 Ⅳ-1〕에 제시했다. 구획 수와 구획 당 응시자 수가 달라서 생기는 오차를 없애기 위해 각각의 평균자승편차를 계산할 때 구획 수는 동일하게 했다.

〈표 Ⅳ-1〉 특성곡선전환 동등화를 준거로 한 평균자승편차(RMSD)

	무 구획	2 구획	3 구획	4 구획	5 구획	6 구획
점-사전정보	.69	.48	.48	.50	.47	.54
사전정보	2.13	2.02	1.99	1.92	1.83	1.92
고-사전정보	2.13	2.01	1.98	1.93	1.83	1.91



〔그림 Ⅳ-1〕 특성곡선전환 동등화를 준거로 한 평균자승편차(RMSD)

점-사전 정보(Point Prior) 동등화가 모든 구획 상황에서 가장 낮은 편차를 보였고, 사전 정보(Informative Prior)와 고-사전 정보(High Informative Prior) 동등화는 거의 유사하게 높은 편차 값을 보였다.

평균자승편차의 준거가 특성곡선전환(CCT) 동등화라는 점을 고려할 때, 점-사전 정보 동등화는 특성곡선전환 동등화(즉, 무-사전 정보 동등화)와 유사한 동등화 결과를 보이는 반면, 사

전 정보 동등화와 고-사전 정보 동등화는 이들과 다른 결과를 보임을 알 수 있다.

무-구획 설계의 동등화 편차가 구획이 여럿 있는 설계의 동등화 편차보다 높게 계산된 것은 이례적이다. 본 연구에서 계산한 평균자승편차 값이 관찰 점수를 통해 계산된 것이기 때문에 이 결과를 두고 각 동등화 방법의 양호도를 직접적으로 판단할 수는 없다. 다만, 평균자승편차를 계산할 때 사용한 준거와 같거나 다른지 또는 얼마나 많이 다른지 등을 알 수 있을 뿐이다.

〈표 Ⅳ-1〉의 평균자승편차는 첫 번째 연구 문제를 규명하는데 충분하지 않은 것 같다. 그래서 종속 표본 집중 경향의 차이에 관한 프리드먼(M. Friedman) 이원 분산분석을 실시해서 〈표 Ⅳ-2〉에 정리했다. 각 칸의 값들은 각각의 사전 정보를 가지고 동등화한 결과 값의 평균 등위이다. 예를 들어, 무-구획 설계로 수집한 자료를 가지고 네 가지 방식의 동등화를 수행한 결과 무-사전 정보 동등화 값은 평균 1.48 등위를, 점-사전 동등화 값은 평균 1.66 등위를, 사전 정보 동등화 값과 고-사전 정보 동등화 값은 평균 3.43 등위를 나타낸다는 것이다. 즉, 무-사전 정보 동등화 점수가 가장 낮고 사전 정보 또는 고-사전 정보 동등화 점수가 상대적으로 높다는 뜻이다.

모든 구획 상황에서 네 가지 동등화 점수들이 비슷한 양상을 보이고 있다. "네 가지 동등화 점수들 사이에는 차이가 없다"는 영 가설을 프리드먼 방식으로 검정한 결과 모두 기각되었다 ($p<.001$). 사전 정보 유형이 다른 동등화 결과 간에 차이가 있음을 의미한다. 대체로 무-사전 정보 동등화와 점-사전 정보 동등화 결과끼리 비슷하였고, 사전 정보 동등화와 고-사전 정보 동등화 결과끼리 비슷하였다.

〈표 Ⅳ-2〉 사전정보유형에 따른 동등화의 차이에 관한 프리드먼 검정

	무-구획	2 구획	3 구획	4 구획	5 구획	6 구획
무-사전정보	1.48	1.38	1.38	1.39	1.39	1.41
점-사전정보	1.66	1.70	1.70	1.68	1.68	1.66
사전정보	3.43	3.95	3.75	2.98	3.70	3.73
고-사전정보	3.43	2.98	3.18	3.95	3.23	3.20
χ^2(자유도)	134.15(3)	147.76(3)	137.09(3)	147.22(3)	134.78(3)	135.69(3)
확률	.000	.000	.000	.000	.000	.000

평균자승편차와 프리드먼 검정 결과를 종합해 볼 때, 네 가지 동등화 방법 중 어느 것이 가장 안정적인 결과를 나타내는지는 알 수 없으나 무-사전 정보 동등화와 점-사전 정보 동등화는 서로 비슷하고, 사전 정보 동등화와 고-사전 정보 동등화 역시 서로 비슷함을 알 수 있었다. 무-사전 정보 동등화는 기존의 특성곡선전환(CCT) 동등화의 다른 이름이고 점-사전 정보 동등화는 문항모수고정(FPIP) 동등화와 같은 방식이라는 점을 감안 할 때, 기존의 IRT 동등화 방법들 간에 큰 차이가 없음을 다시 확인할 수 있었다. 사전 정보 동등화와 고-사전 정보 동등화는

기존의 IRT 동등화 개념을 확장한 것인데, 이들의 결과가 기존의 IRT 동등화 결과들과 다르다는 것을 확인 할 수 있었다.

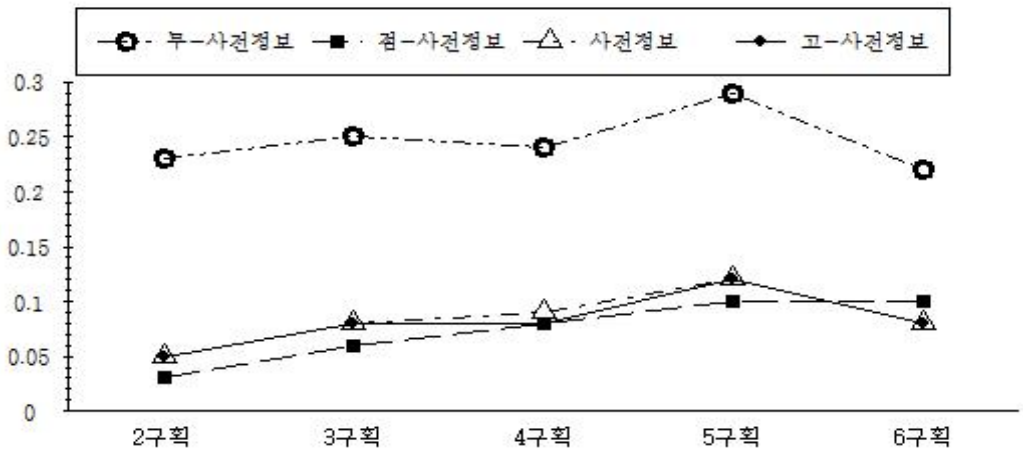
2. 검사설계의 영향

두 번째 연구 문제(구획 수의 변화에도 동등화 결과에 큰 변화가 없는 사전 정보 유형은 무엇인가?)를 규명하기 위해 무-구획 설계의 동등화 값을 준거로 하는 평균자승편차(RMSD)를 계산하여 〈표 Ⅳ-3〉과 〔그림 Ⅳ-2〕에 정리했다.

모든 구획 상황에서 무-사전 정보 동등화 결과가 가장 불안정했다. 나머지 동등화 방법들은 대체로 유사한 안정성을 보였다. 소수점 둘째 자리에서만 차이가 있을 뿐이었다. 구획 수가 많아지면서 평균자승편차가 높아지는 양상은 모두 같았는데, 구획 수가 가장 많은 6 구획 상황이 5 구획 또는 4 구획 상황보다 안정적인 결과를 보인 점은 이례적이다.

〈표 Ⅳ-3〉 무-구획 설계를 준거로 한 평균자승편차(RMSD)

	2 구획	3 구획	4 구획	5 구획	6 구획
무-사전정보	.23	.25	.24	.29	.22
점-사전정보	.03	.06	.08	.10	.10
사전정보	.05	.08	.09	.12	.08
고-사전정보	.05	.08	.08	.12	.08



〔그림 Ⅳ-2〕 무 구획 설계를 준거로 한 평균자승편차(RMSD)

구획 수가 많아지고 구획 당 응시자 수가 줄어드는 상황에서도 안정적인 동등화 결과를 나타내는 사전 정보 유형이 무엇인지를 보다 세밀히 살펴보기 위해, 종속 표본 집중 경향의 차이에 관한 프리드먼(M. Friedman) 이원 분산분석을 실시해서 그 결과를 〈표 Ⅳ-4〉에 정리했다. 각 칸의 숫자는 특정 동등화 값의 평균 등위를 말한다. 예를 들어 무-사전 정보 동등화를 무-구획 자료로 실시했을 때에는 평균 등위가 2.77이었으나 2 구획 자료에서는 1.96으로, 3 구획 자료에서는 3.00 등으로 바뀌었다는 것이다.

프리드먼 검정의 유의 수준을 5%로 했을 때, 네 가지 사전 정보 유형의 동등화들 모두 "여섯 가지 구획 설계로 수집한 자료의 동등화 결과는 같다"는 영 가설을 기각하였다. 한편 카이 자승값으로 볼 때, 무-사전 정보 동등화와 점-사전 정보 동등화는 구획 수가 달라짐에 따라 평균 순위의 변화가 많았음을 알 수 있고, 사전 정보 동등화와 고-사전 정보 동등화 결과는 상대적으로 평균 순위의 변화가 적었음을 알 수 있다. 특히 고-사전 정보 동등화는 변화가 가장 적었다(χ^2 =14.20, p=.014).

〈표 Ⅳ-4〉 검사설계에 따른 동등화의 차이에 관한 프리드먼 검정

	무-구획	2 구획	3 구획	4 구획	5 구획	6 구획	χ^2(자유도)	확률
무-사전정보	2.77	1.96	3.00	3.98	5.09	4.20	106(5)	.000
점-사전정보	1.16	2.13	3.11	3.88	5.61	5.13	245(5)	.000
사전정보	3.27	4.04	4.04	3.34	2.88	3.45	17.38(5)	.004
고-사전정보	3.33	3.93	3.96	3.46	2.85	3.46	14.20(5)	.014

Ⅴ. 논의 및 결론

1. 논의

가장 안정적인 동등화 결과를 나타내는 사전 정보 유형이 무엇인지를 묻는 첫 번째 연구 문제를 규명하기 위해, 특성곡선전환(CCT) 동등화를 준거로 한 평균자승편차(RMSD)를 계산하고, "네 가지 정보 유형에 의한 동등화 결과가 같다"는 영 가설을 프리드먼 카이 자승으로 검정했다. 어떤 방식이 가장 안정적인 동등화 결과를 나타내는지를 규명하는 데는 성공하지 못했지만 무-사전 정보(Flat Prior) 동등화와 점-사전 정보(Point Prior) 동등화가 유사한 결과를 나타내고, 이들은 새로운 베이지안 IRT 동등화 방법인 사전 정보(Informative Prior) 동등화나 고-사전 정보(High Informative Prior) 동등화와는 다른 결과를 나타낸다는 것을 알 수 있었다.

무-사전 정보 동등화는 기존의 특성곡선전환(CCT)에 의한 동등화와 같고, 점-사전 정보 동

등화는 기존의 문항모수고정(FPIP)에 의한 동등화와 같다는 점을 감안하면, 본고에서 적용해 본 베이지안 IRT 동등화는 기존의 IRT 동등화 방법들과 다른 결과를 낳을 수 있다는 가능성을 확인한 셈이다.

구획 수의 증가에도 불구하고 동등화 결과에 큰 변화가 없는 사전 정보 유형은 무엇인지를 묻는 두 번째 연구 문제를 규명하기 위해, 무-구획 설계에서의 동등화를 준거로 한 평균자승편차(RMSD)를 계산하고, "다양한 구획 수를 가진 설계에서의 동등화 결과는 같다"는 영 가설을 프리드먼 카이 자승 방식으로 검정했다.

네 가지 정보 유형 모두 구획 수가 증가함에도 불구하고 평균자승편차 값이 크게 증가하지 않았으나, 무-사전 정보 동등화가 다른 정보 유형들에 의한 동등화에 비해 평균자승편차 값이 가장 컸다. 프리드먼 카이 자승 검정 결과 1% 유의 수준에서 모두 영 가설을 기각함으로써 구획 수를 달리한 설계에서의 동등화 결과가 같다고 볼 수 없었다. 그런데 점-사전 정보와 무-사전 정보 동등화의 프리드먼 카이 자승 값이 가장 컸고 사전 정보와 고-사전 정보 동등화의 카이 자승 값이 가장 작았다. 무-사전 정보 동등화나 점-사전 정보 동등화에 비해 구획 수가 증가함에도 불구하고 사전 정보 또는 고-사전 정보 동등화 결과가 크게 달라지지 않았다는 뜻이다.

무-사전 정보 동등화의 평균자승편차가 유난히 큰 가장 주된 이유는 모수 추정 과정에서 모수 동등화가 한꺼번에 이루어지는 다른 방식들에 비해 모수 추정과 동등화의 두 단계를 별도로 거쳐야 하고 그 과정에 오차가 발생할 가능성이 크기 때문이다(de la Torre & Patz, 2001). 새롭게 제안된 베이지안 IRT 동등화는 사전 정보를 다양하게 달리할 수 있을 뿐만 아니라 기존의 IRT 동등화 방법들과 달리 모수 추정과 동등화를 한꺼번에 처리한다는 장점이 있다. 실제로 본고에서도 사전 정보를 사용한 동등화(점-사전 정보, 사전 정보, 고-사전 정보)가 사용하지 않은 동등화(무-사전 정보)보다 안정적인 결과를 보였다.

비록 본고에서는 점-사전 정보 동등화가 구획 수의 증가에도 불구하고 안정적인 동등화 결과를 나타냈지만, 실제 검사 상황에서는 다른 결과를 보일 수 있음을 지적하고자 한다. 본고에서는 고전적인 가교 문항 설계에서 얻은 검사 자료를 연구 목적에 맞게 여러 구획으로 이루어진 행렬 표집 설계로 변형했을 뿐이기 때문에 가교 문항의 위치 변화에 의한 영향은 없었다. 그러나 실제 행렬 표집 상황에서는 가교 문항들의 위치가 달라질 것이기 때문에 이른바 맥락 효과(Li, Lissitz, & Yang, 1999)가 작용할 가능성이 있다. 따라서 점-사전 정보 동등화가 구획 수의 변화에도 불구하고 안정적인 결과를 보일 것이라는 결론을 섣불리 내려서는 안 될 것이다.

특성곡선전환(CCT) 동등화를 준거로 한 평균자승편차(RMSD) 값들(〈표 Ⅳ-1〉 참조) 중 무-구획의 것이 다른 것들에 비해 크게 계산된 것과, 무-구획 설계를 준거로 한 평균자승편차 값들(〈표 Ⅳ-3〉 참조) 중 6 구획의 값이 5 구획 또는 4 구획의 값보다 작게 계산된 것들은 좀 이례적이다. 본 연구를 위한 자료가 경험적 검사 자료이기 때문에 동등화 방법과 구획 수를 제외한 다양한 변인들을 통제하지 못했고, 그래서 나타난 우연 오차 때문으로 판단된다. 또한 고전

적인 가교 문항 설계로 수집된 검사 자료를 행렬 표집 설계로 임의 변형하는 과정에서 오차가 생겼을 가능성도 있다. 이런 문제는 모의 자료를 이용한 연구가 아니고서는 통제할 수 없기 때문에 후속 연구를 기대할 수밖에 없을 것이다.

사전 정보 동등화와 고-사전 정보 동등화의 평균자승편차 값들이 대부분 비슷하게 계산되었는데 그 주된 이유는 이들의 사전 정보가 크게 다르지 않았기 때문으로 판단된다. 사전 정보 동등화와 고-사전 정보 동등화의 상위 사전 정보로서 평균은 서로 같았고 분산만 다를 뿐이었다. 사전 정보에 비해 고-사전 정보의 분산은 0.1배 작았을 뿐이다(〈표 Ⅲ-2〉 참조). 그런데 사전 정보의 분산이 이미 작은 상태였기 때문에 이것의 0.1배에 해당하는 분산과 큰 차이가 없었을 것으로 보인다. 예를 들어, 사전 정보 동등화의 변별도 모수 추정치의 분산은 .084에 불과했고 고-사전 정보 동등화의 변별도 분산은 .0084였다. MCMC 방식으로 15,000번 회전 하여 모수의 사후 분포를 추정했기 때문에 분산의 작은 차이는 큰 영향력을 발휘하지 못한 것으로 보인다(Fox, 2010). 오히려 저-사전 정보 동등화 상황을 가정하여 분산이 더 큰 상위 사전 정보를 고려했어야 하지 않았나 하는 아쉬움이 있다.

2. 결론 및 제언

경험적인 검사 자료를 이용한 연구였기 때문에 가장 안정적인 동등화 결과를 나타내는 사전 정보 유형을 밝힐 수는 없었다. 그러나 기존의 IRT 동등화 방법들과 새로운 베이지안 IRT 동등화 방법들은 분명 다른 결과를 나타낸다는 것을 알 수 있었다.

구획 수가 늘어나고 구획 당 응시자 수가 줄어드는 행렬 표집 설계 상황에서 무-사전 동등화 방법보다는 사전 정보를 사용하는 베이지안 IRT 동등화 방법들이 더 안정적인 결과를 나타낼 수 있다는 부분적인 증거만을 찾을 수 있었다. 최적의 사전 정보 수준을 찾고자 한 당초의 연구 목적을 충분히 달성하지는 못했다.

본 연구의 결론을 보완할 후속 연구가 필요함을 제언하고 싶다. 우선, 고-사전 정보 동등화에 더해 모수 추정치 분산을 크게 하는 저-사전 정보(Low Informative Prior) 상황을 설정한 연구가 필요하다. 더 나아가 분산에 관한 상위 사전 정보(Hyper Priors)로서 감마 분포의 α 값의 변화에 의한 결과를 확인 할 필요도 있다.

국어 시험의 측정학적 특징은 다른 교과와 다를 수 있을 것이다. 수학이나 영어와 같은 다른 교과의 시험 자료를 활용한 후속 연구도 필요하다고 생각한다.

경험적인 검사 자료를 이용한 연구에서 나타나는 이례적인 현상들을 최소화 할 수 있도록 모의 자료를 이용한 후속 연구가 이루어져야 할 것이다. 그리고 점-사전 정보 동등화의 안정성에 대한 과대 평가 여부를 확인하기 위해서는 행렬 표집 설계로 실제 검사를 시행한 결과를 이용한 연구가 이루어져야 할 것이다.

참 고 문 헌

남현우(2002). 행렬표집설계에 적합한 IRT 동등화 방법 탐색. **교육평가연구, 15**(2), 85-103

남현우, 남명호, 양길석, 김진하(2009). 고부담시험에서 문항노출위험과 동등화편차를 최소화 할 수 있는 행렬표집설계 방안. **교육평가연구, 22**(3), 633-657.

박정(2006). **국가수준 학업성취도평가-기술보고서-**. 한국교육과정평가원 연구보고 RRO 2006-4.

박찬호, 강태훈(2011). 전문가 판정에 의한 차등 배점을 활용한 제한적 일반화부분점수 모형의 적용. **교육평가연구, 24**(3), 781-798.

이종성, 강계남, 김양분, 강상진(2007). **통계방법(4판)**. 서울: 박영사.

Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician, 58*(4), 330-336.

Curtis, S. M. (2010). BUGS code for Item Response Theory. *Journal of Statistical Software, 36*(1), 1-34.

de la Torre, J., & Patz, R. J. (2001). Item Response Theory Equating using Bayesian Informative Priors. Paper presented at the annual meeting of the National Council on Measurement in Education in Seatle, WA.

Driscoll, D. P. (2002). MCAS 2001 Technical Report. Massachusetts Department of Education.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Application*. Springer.

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311-333.

Kim, J., & Bolt, D. M. (2007). Estimating Item Response Theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practices,* An NCME Instructional Module.

Kim, S., & Kolen, M. J. (2004). *STUIRT(A computer program for scale transformation under unidimensional Item Response Theory models)*. Iowa Testing Programs, University of Iowa.

Kim, S., & Lee, W. (2004). IRT scale linking methods for mixed-format tests. ACT research report series 2004-5. Iowa City: ACT.

Kim, S., & Kang, T. (2012). A comparison of MCMC and MMLE/EM algorithms for fixed item parameter calibration. *Journal of Educational Evaluation, 25*(2), 337-350.

Kolen, M. J. (2004). *POLYEQUATE(Windows console version)*. University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking - methods and practice -*, 2nd edition. Springer.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). Estimating IRT equating coefficients for test with polytomously and dichotomously scored items. Paper presented at the annual meeting of the NCME, Montreal, Quebec.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in Item Response Theory. *Journal of Educational Measurement, 23*(2), 157-162.

Massachusetts Department of Education (2001). Overview of the MCAS 2001 tests. http://www.doe.mass.edu/mcas/

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for Item Response Theory. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Spigelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS version 1.4 user manual.* Cambridge, England: MRC Biostatistics Unit.

van Rijn, P., & Beguin, A. (2009). Test equating using prior information on population. A paper presented at the International Meeting of the Psychometric Society at Cambridge, UK.

van Rijn, P., & Beguin, A. (2010). Exposing trends in populations in the context of test equating. Paper presented at the annual meeting of National Council on Measurement in Education at Denver, Colorado.

· *논문접수 : 2012-09-01/ 수정본접수 : 2012-10-11/ 게재승인 : 2012-10-17*

ABSTRACT

Bayesian IRT Equating in the Common plus Matrixed-sampled Anchor Items Design with Mixed Item Formats

Hyun-Woo Nam
(Professor, Soon Chun Hyang University.)

This study was intended to find out the appropriate priors for the Bayesian IRT equating in the common plus matrix-sampled anchor items design. For the study, the 10th grade students' Korean Test in the National Assessment of Educational Achievement was used. The item responses sampled by non- equivalent group anchor test design were revised to the common plus matrix-sampled anchor items design with 6 blocks at most. Traditional IRT equating methods, CCT and FPIP in the context of MLE, were included to the Bayesian IRT equating as those of flat priors and point priors methods. In addition to those of completely uninformative priors and completely informative priors, intermediate informative priors in the context of MCMC were introduced to this study. The traditional IRT equating method(CCT) showed its equating result a little bit different from those of Bayesian IRT equating with intermediate priors (point prior, just prior, high prior) in the context of MCMC. This study also revealed high possibilities of Bayesian IRT equating with appropriate priors in the non-optimal testing designs such as matrix-sampled anchor items design.

Key Words : Matrixed-sampled Design, Prior Information, Bayesian IRT Equating, MLE(Maximum Likelihood Estimation), MCMC(Markov Chain Monte Carlo)