강 태 훈 교수지도

석사학위 청구논문

다분 문항반응모형에서 사후기대추정법의 능력모수 추정 정확성에 관한 연구

- EAPrp와 EAPss를 중심으로 -

2015

성신여자대학교 대학원

교육학과

심 혜 진

다분 문항반응모형에서 사후기대추정법의 능력모수 추정 정확성에 관한 연구

- EAPrp와 EAPss를 중심으로 -

강 태 훈 교수지도

이 논문을 석사학위논문으로 제출함.

2014년 11월

성신여자대학교 대학원

교육학과

심 혜 진

인 준 서

심혜진의 석사학위 논문으로 인준함.

2014년 11월

심사위원장 김 명 랑 인

심 사 위 원 조 영 일 인

심 사 위 원 강 태 훈 인

성신여자대학교 대학원

논 문 개 요

본 연구는 다분 문항으로 이루어진 검사에서 피험자의 능력을 추정하기 위한 방법 중에서, 검사 총점을 고려한 사후기대추정법(EAPss)이 문항별 반응양식을 고려한 사후기대추정법(EAPss)에 비하여 상대적으로 어떠한 기능을 보이는가를 모의실험을 통하여 살펴보고자 한다. 수리적인 특성으로 인하여 EAPrp가 EAPss에 비하여 조금 더 정확한 능력을 추정하지만, EAPss의 경우 원점수와 일대일로 대응하는 속성으로 인하여 일반 대중의 이해가 용이하다는 장점을 갖는다. 이 연구에서는 EAPss의 진모수 복원력이 EAPrp와 실제로 어느 정도 차이가 나는지를 살펴보고자 하였다. 만약 실질적 차이가 존재하지 않는다면, 대중의 이해를 돕고 거부감이 덜할 수 있는 IRT 베이지안 능력추정 방식으로서 EAPss의 사용 선택을 정당화할 수 있을 것으로 기대한다.

연구에서 적용되는 다분 문항반응모형은 등급반응모형(GRM)과 일반화부분점수모형(GPCM)이며 두 EAP 방법이 다양한 연구조건에서 수행하는 기능을 파악하기 위해 '검사 길이', '응답 문항 범주 수', '피험자 능력분포' 등 다양한 조건을 추가하여 연구를 수행한다. 연구는 크게 두 가지 측면으로 구성되어 수행되며 두 측면은 측정학적으로 서로 상호보완적인 기능을 한다. 첫째, 주변(marginal)분포에 관한 연구를 수행한다. 이를 통해 두 EAP 방법의 유사성과 진능력 모수에 대한 복원력을 능력 수준에 관계없이 피험자 전체에 대하여 비교할 수 있다. 보다 다양한 형태의 상황속에서 모형의 적용가능성을 탐색하기 위하여 '모형의 종류', '검사 길이', '응답 문항 범주 수', 및 '피험자 능력분포'를 다양하게 조건화하여 두 EAP 방법이 어떠한 양상을 보이는지 탐구한다. 둘째, 조건부(conditional) 분포에 관한 연구를 수

행한다. 능력모수를 17개의 구간으로 나누어 각 구간 즉 다양한 능력 수준 대별로 EAPss와 EAPrp의 수행을 비교한다. 이를 위해 '모형', '검사 길이', 및 '응답 문항 범주 수'를 다양하게 조건화하여 두 EAP 방법에 따른 능력 수준별 분포는 어떠한 양상을 보이는지 탐구한다.

먼저 모의실험을 통해 산출한 두 EAP 추정치들의 주변 분포에 대한 결과는 다음과 같다. 첫째, 전반적으로 EAPrp가 EAPss에 비하여 조금 더 적은 MSE, SB, VAR 값을 산출하였지만 그 차이는 소수점의 자리에서 발생하는 미미한 차이였다. 둘째, 베이지안 능력모수 추정을 위한 사전정보의 다양한 사용은 능력모수 추정의 정확성에 영향을 미쳤으며 균일분포를 사전정보로 활용하는 것보다 표준정규분포나 피험자분포를 사전분포로 재사용하는 것이 상대적으로 더 안전한 선택임을 확인하였다. 셋째, 베이지안 능력모수 추정을 위한 피험자 분포 중에서 정규분포 형태일 때 능력모수 추정의 정확성이 상대적으로 높은 것으로 확인되었다. 넷째, 문항 범주와 검사 길이가 증가하면 EAPrp와 EAPss를 사용한 능력모수 추정치의 MSE값은 줄어든다. 마지막으로 GRM과 GPCM 적용에 따라 두 EAP 능력추정치의 정확도에 차이가 있었으며 GPCM을 사용한 경우가 GRM을 사용한 경우보다 조금 더 정확하게 진능력모수 복원력을 보여주었다. 능력 수준별 연구에 대한 두 EAP 능력모수 추정치의 결과는 주변 분포에 대한 결과와 매우 유사하다.

능력 수준별 분포 연구결과를 통해 보다 자세히 확인할 수 있는 것은 다음과 같다. 표준정규 사전분포를 사용한 경우, 두 EAP 능력추정 방법 모두 [-2, 2]의 능력범위에서 0에 가까운 SB와 VAR 값을 보였다. 다만, 균일분포 사전분포를 사용한 경우, 양극단에 위치한 능력수준의 피험자를 더 정확히 추정하는 것으로 나타났다.

본 연구는 EAP 능력추정 방법을 상관계수 및 MSE 등을 통하여 비교하였으며, 연구의 결과 그 차이는 매우 미미하였다. 따라서 실제 검사 프로그

램 운영 하에서 EAPss가 가능한 선택임을 확인 및 시사해줌과 더불어 여러 검사 상황에서 선택 가능한 적절한 '모형', '검사 길이', '응답 문항 범주 수', '사전분포', 및 '피험자 분포'를 경험적으로 확인해 주고 있다.

목 차

표 목 차

그 림 목 차

Ⅰ. 서 론

1. 연구의 필요성 및 목적

교육 현장에서 교사들은 학생들의 능력과 특성을 파악하고 그에 따른 적절한 교육적인 조취를 제공하기 위하여 학생들의 능력과 특성을 표상화할 수 있는 도구가 필요하다. 뿐만 아니라 사회 과학의 다양한 분야에서도 인간 행동의 본질을 파악하고 이를 일반화하기 위하여 인간의 특성을 표상할 수 있는 도구와 방법이 필요하다. 교육 및 사회과학 분야에서 인간의 지적인 특성이나 정의적인 특성들을 파악하기 위하여 일반적으로 사용하는 방식은 수학능력 검사지나 각종 심리검사지 그리고 설문지를 사용하는 것이다. 그러나 학생의 수학적 능력이나 지능과 같은 인간의 지적인 특성이나 정의적인 특성들은 물리적 특성처럼 직접 관찰할 수 있는 특성이 아니다. 따라서 사회과학의 측정에서는 가설적인 인간의 특성을 가장 타당하고 신뢰롭게 파악할 수 있는 방법을 강구하기 위하여 검사의 측정에 관한 다양한 연구들이 진행되었다.

사회과학 분야에서 현재 이론적으로 정교화되어 가고 있으며, 현실적으로 널리 사용되고 있는 대표적인 측정이론은 문항반응이론(Item Response Theory, 이하 IRT)이다. IRT는 각 문항에 근거하여 피험자의 능력과 특성을 설명한다. 따라서 IRT 모형에 근거하여 피험자의 능력을 추정할 경우, 피험자 능력모수치의 문항불변성이라는 논리에 의하여 피험자들이 각기 다른 문항들로 이루어진 검사를 치루어도 능력을 추정할 수 있다는 장점이 있다. 이 장점으로 인하여 IRT는 폭 넓게 사용되고 있다.

대표적인 예로 미국의 SAT(Scholastic assessment test), TOEFL(Test of English for Foreign language), GRE와 NAEP(National Assessment Educational Progress), 그리고 TIMSS(the Third International Mathematics and Science Study)등의 연구와 OECD 주관의 PISA(Program for International Student Assessment) 등의 연구들에서 문항의 분석이나 피험자 능력의 추정을 위하여 문항반응이론이 활발히 사용되고 있다.

다분 문항반응이론(Polytomous Item Response Theory)은 이분 문항반응이론을 사용하여 검사지를 '맞고/틀림' 또는 'Pass/Non-Pass'와 같이 이분으로만 분류할 수 있었던 채점방식을, '매우 아니다/아니다/보통이다/그렇다/매우 그러다'와 같이 다양한 반응지에 대한 문항의 특성에 대한 피험자의 특성을 산출할 수 있도록 만든 문항반응이론의 일종이다(한국성인교육학회, 1998). 다분 문항반응모형을 이용하여 검사를 분석할 때 피험자의 능력모수를 추정하기 위하여 일반적으로 많이 사용되고 있는 방법 중 하나는 사후기대추정법(Expected A Posterior Estimation, 이하 EAP)이다. 사후기대추정법은 다른 추정 방법에 비해 이론적 및 실제적 장점을 가진다. 계산이 용이하고 피험자의 반응이 극단적인 경우에도 능력 모수치를 측정할 수 있다는 특징이 있다(박정, 1999a, Bock & Mislevy, 1982; Lord, 1986). EAP는 피험자의 문항별 반응형태에 근거한 방법(EAP by response pattern, 이하 EAPrp)과 검사총점에 근거한 방법(EAP by summed score, 이하 EAPss)으로 분류할 수 있다(Kim, 2007; Thissen & Orland, 2001; Tong & Kolen, 2010).

IRT 모형을 사용할 때, 흔히 그 장점으로 언급되는 것은 검사의 총점에 기반하여 피험자의 능력이 추정되지 않고 반응 형태 즉 개별 문항에서 획득한 점수들의 양상에 따라 피험자의 능력추정치가 산출된다는 점

이다. 하지만, 실제 검사 결과에 대한 분석 문맥에서 문항별 반응 형태에 근거하여 검사 점수를 계산하는 것보다 총점에 근거하여 검사 점수를 계산하는 것이 필요한 경우가 종종 발생한다. 예를 들어, 대규모 검사 결과를 분석하는 경우 IRT 모형을 사용하여 척도화된 점수를 산출하기 전에 피험자들의 원점수를 계산하여 기록하여 두는 것이 일반적이다. 또한 점수를 보고할 때에도 해석적 편의를 위하여 백분율을 계산하여야 할 때, 원점수의 총점 분포를 사용할 수 있다. 총점 분포의 모형을 기반으로 한 추정치는 모수 분포 추정의 타당도를 포함하여, IRT 모형의 적합도를 판단하는 통계적 지표로서 기능할 수 있다(Stucky, 2009; Thissen, 1995).

일반적인 IRT 모형의 조건하에서 EAPrp를 사용하여 능력모수를 추정하면 검사 총점이 같은 피험자라 하더라도 문항별 반응형태에 따라 서로 다른 능력모수가 추정될 수 있다. 이 결과는 일반 학부모나 학생들처럼 능력모수 추정법에 대한 전문지식이 없는 일반인에게 직관적인 이해를 주기 어려울 수 있다. 한국교육과정평가원의 대학 입시 전형 간소화 정책과 연계한 NEAT(2,3급) 개선 방안 연구 보고서에서 박태준 외(2013, p.140, 재인용)는, 패턴 스코링 즉 문항반응 형태에 기반한 방식으로 능력을 산출하는 경우 추정된 능력과 원점수 순위의 불일치로 인하여 피험자들이 채점방식에 거부감을 가질 수 있으므로 이론적인 정확성에도 불구하고 피험자들의 능력추정방식 이해의 어려움으로 인하여 논란이 될 수 있음을 지적한다.

점수 총점에 근거한 척도화 방식의 문제점으로 지적되는 것은 문항 반응 형태에서 사용하던 정보를 사용하지 않음으로 인하여 갖게 되는 정보의 손실이다. 그러나 많은 학자들이 이러한 단점을 감수하고도 점수총점 방식을 사용하는데 이점이 있다고 주장한다. Yen(1984)은 3모수 로지스틱 IRT 모형을 사용하였을 때, 문항 반응 형태에 의한 점수와 총점에 의

한 점수의 표준 오차는 소수점의 자리에서 발생하는 작은 차이임을 강조하였다.

이러한 상황 속에서 일반인들에게 보다 설득력 있는 능력척도 점수를 제공하기 위해서 피험자의 문항 반응형태가 아닌 검사총점을 사용하여 사후기대추정법으로 능력모수를 추정하는 EAPss방식에 대한 연구가 활발히 진행되었으며 특히 EAPrp와 EAPss를 함께 고려하여 능력 추정의 결과를 비교한 경우를 자주 발견할 수 있다(강태훈, 백순근, 2007; 강태훈, 2014; 김성훈, 2012b; Kolen & Tong, 2010; Kolen, 2012; Thissen & Orland, 2001; Tong & Kolen, 2010).

본 연구의 주된 목적은, 이제껏 주로 이분문항반응모형 맥락에서 연구되어 온 EAPrp와 EAPss 간의 비교를 다분 문항반응모형으로 확장하는 데에 있다. 본 연구에서는 다분 문항반응모형하에서 EAPss의 수행력을 EAPrp의 수행력과 비교하여 EAPss를 선택하였을 때의 능력모수의 추정 정확도가 구체적으로 어느 정도인지를 체계적으로 정리함으로써, 검사 자료 분석을 실시하는 연구자가 이를 감수하고 EAPss를 사용할지 또는 EAPrp를 사용할 지에 대한 선택을 도울 수 있을 것이다.

2. 연구문제

본 연구는 다분 문항반응모형에 기초한 모형들에 대하여 피험자의 능력모수를 추정하는 다양한 모의실험 조건 하에서 모의실험 연구를 수행함으로써 두 EAP 방법(EAPrp와 EAPss)의 진능력모수 복원의 정확성을 비교 및 평가하는데 목적이 있다. 이 연구에서 적용되는 다분 문항반응모

형은 Samejima(1969)의 등급반응모형과 Muraki(1992)의 일반화부분점수 모형이며 두 EAP 방법이 다양한 연구조건에서 수행하는 기능을 파악하기 위해 '검사 길이', '응답 문항 범주 수', '피험자 능력분포' 등의 조건을 추가하여 연구를 수행한다.

연구는 크게 두 가지 측면으로 구성되어 수행되며 두 측면은 측정학적으로 서로 상호보완적인 기능을 한다(김성훈, 2010). 첫째, 주변(marginal) 분포에 관한 연구를 수행한다. 이를 통해 두 EAP 방법의 유사성과 진능력 모수에 대한 복원력을 비교할 수 있다. 보다 다양한 형태의 상황속에서 모형의 적용가능성을 탐색하기 위하여 '모형의 종류', '검사 길이', '응답 문항 범주 수', 및 '피험자 능력분포'를 다양하게 조건화하여 두 EAP 방법의 능력 추정 정확도의 차이를 탐구한다. 둘째, 조건부(conditional) 분포에 관한 연구를 수행한다. 능력모수를 17개의 구간으로 나누어 각 구간에서 두 사후기대추정법의 수행을 비교 및 탐구한다. 이를 위해 '모형', '검사 길이', 및 '응답 문항 범주 수'를 다양하게 조건화하여 두 EAP 방법에 따른 능력 수준별 분포의 능력 추정 정확도의 차이를 탐구한다. 이를 위한 연구문제는 다음과 같다.

첫째, 주어진 모의실험 조건하에서 두 사후기대추정법으로 피험자 능력모수를 추정하였을 때, 피험자의 능력에 대한 두 사후기대추정법의 진모수 복원력은 검사에 응시한 피험자 전체를 함께 고려하는 맥락에서 두 추정치 상호 간 상관계수, 진값과 추정치 간의 상관계수 및 평균제곱오차 등의 차이는 어떠한가?

둘째, 주어진 모의실험 조건하에서 두 사후기대추정법으로 피험자 능력모수를 추정하였을 때, 여러 수준의 피험자 능력모수 구간 각각에서 두 추정치 상호 간 상관계수, 진값과 추정치 간의 상관계수 및 조건부 편향, 조건부 표준오차, 평균제곱오차 등의 차이는 어떠한가?

Ⅱ. 이론적 배경

본 연구는 다분 문항반응모형하에서 EAPss의 수행력을 EAPrp의 수행력과 비교하여 EAPss를 선택하였을 때의 능력모수의 추정 정확도가 구체적으로 어느 정도인지를 체계적으로 정리하기 위하여 다분 문항반응모형에 기초한 모형들에 대하여 피험자의 능력모수를 추정하고 두 EAP 방법의 진능력모수 복원의 정확성을 비교 및 평가하기 위하여 모의실험 연구를 수행한다. 이를 위하여 본 연구에서는 베이지안을 적용한 능력모수 추정방법 및 다분 문항반응모형 중 등급반응모형과 일반화부분점수모형에 관한 이해를 요한다.

1. 능력모수 추정 방법

IRT 모형들은 피험자의 검사 점수를 계산하기 위해 다양한 추정방법을 사용하여 채점 가중치(optimal scoring weight)를 부여한다. 각 IRT 모형에서 유도된 채점 가중치를 사용할 때 문항정보함수는 최대화한다. 따라서 각 추정밥법을 이용하여 검사 결과를 점수화할 때, 어떠한 채점 가중치가 사용되느냐에 따라서 검사 점수의 신뢰도는 영향을 받는다 (Stucky, 2009).

피험자의 검사 점수를 추정하기 위한 모든 채점 가중치는 변별도 모수와 추측도 모수와 같은 모형 내의 문항 모수들의 함수를 포함한다 (Baker, 1992). 이와 같은 방법은 문항별 반응형태에 근거한 채점 방식 (item pattern scoring; IP scoring)이다. 문항별 반응형태에 근거한 채점

방식의 조건하에서, 피험자들의 검사 원점수의 총점이 동일할지라도 개별 피험자들의 각 문항 반응 형태에 따라서 다른 능력 추정치가 할당될 수도 있다.

문항에 차별적 기여를 허락하지 않는 모형하에서는 모든 문항들에 동일한 가중치가 부여된다. 동일한 검사를 치른 두 명의 피험자가 같은 원점수 총점을 받았다면, 피험자들이 응답한 문항별 반응의 형태와는 관계없이, 두 피험자는 같은 능력 추정치를 얻게 된다. 이와 같은 방법은 검사 총점에 근거한 채점 방식(summed score)이다. 검사 총점에 근거한 채점 방식은 문항반응모형에도 확장하여 적용될 수 있다.

여러 IRT 능력 추정 방법들 가운데 본 연구는 베이지안 방법의 하나인 EAP 방법에 주된 관심이 있다. EAP는 다시 피험자의 문항별 반응형태에 근거한 방법(EAP by response pattern, 이하 EAPrp)와 검사총점에 근거한 방법(EAP by summed score, 이하 EAPss)로 나누어진다.

1) 베이지안 추론방법

베이지안 추론의 기본 개념은 통계적으로 추정해야 할 모수나 결측치 등은 불확실하며 이 불확실의 정도는 확률로써 표현된다는 가정에서 출발한다. 즉 주어진 자료 또는 정보를 통하여 추정하고자 하는 모수의 불확실성을 확률로 나타낸다. 이는 전통적인 통계학의 기본 가정인 '모수는 알려지지 않았지만 고정된 것'이라는 것과 그 근본을 달리한다(강기훈 외, 2012). 미지의 모수를 확률변수로 가정하고 주어진 자료를 조건으로 하는 사후분포를 추론함으로서 확률에 관한 분포를 구할 수 있다. 이를

이용하여 사전적 확률(prior probability)을 사후적 확률(posterior probability)로 활용할 수 있다(Hamada, 2008). 어떠한 사전분포(prior distrivution)를 사용하는가는 사후분포(posterior distribution)의 형성에 큰 영향을 미치므로 적절한 사전분포의 선택은 무엇보다 중요하다(Cass & Raftery, 1995).

베이지안 추론은 특정 결과가 관찰되었을 때, 그 결과가 어느 특정한 원인 때문에 발생할 확률을 다루는 조건부 확률인 베이즈 정리(Bayes' theorem)를 기본이론으로 하며 식 (1)과 같이 나타낼 수 있다.

$$p(\theta|x) \propto p(x|\theta)p(\theta) \tag{1}$$

θ : 추정하고자 하는 확률 모수

x : 유한개의 측정된 자료

$p(x|\theta)$: θ하에서 측정된 자료의 우도

$p(\theta)$: θ의 사전확률밀도함수

$p(\theta|x)$: 측정된 자료 x하에서 업데이트되는 사후확률밀도함수

새로운 자료가 추가되면 사후분포 $p(\theta|x)$는 다시 사전분포 $p(\theta)$로 업데이트되어 활용되며 이러한 반복적인 과정은 결국 추정하고자 하는 모수 θ의 신뢰도를 향상 시킨다. 따라서 베이지안 추론방법에서는 "모수의 사전분포 결정", "자료의 확률모형과 사전분포를 이용한 사후분포의 계산", "사후분포를 이용한 모수에 대한 추론"의 과정을 이해하는 것이 중요하다.

사후분포를 결정하기 위해서는 신뢰할 수 있는 추정을 유도할 수 있는 사전분포를 우선 선택해야 한다. 사전분포는 모수가 취할 수 있는 가능한

값에 대한 사전정보(prior information)를 확률분포를 이용해 나타낸다. 그 후 우도함수와 사전분포가 공액(conjugate)을 이루게 하는 방법을 사용하여 사전분포와 사후분포가 동일한 모수분포의 형태를 갖도록 한다(선은주, 2011). 베이지안 추정 시 주로 사용되는 사전분포와 우도함수, 사후분포는 <표 1>에 제시하였다.

<표 1> 사전 표본 분포와 이에 대한 사후분포

Prior	Likelihood	Posterior
Poisson	Poisson(λ)	Gamma(a,b)
Binomial	Binomial(p,N)	Beta(a,b)
Normal(Known σ^2)	N(μ,σ^2)	N(μ_0,σ_0^2)
Normal	N(μ,σ^2)	$[\mu,\sigma^2] \ni G(\mu_0,c,a,b)$ $[N(\mu_0,c\sigma^2) \times IG(a,b)]$
Gamma(v known)	Gamma(v,θ)	Gamma(a,b)
Exponential	Exponential(θ) =Gamma($1,\theta$)	Gamma(a,b)
Multinomial	Multinomial(a,N)	dirichlet(a_0)

모형에 대한 사전정보를 가정할 수 있는 경우, 자료를 통해 사전분포를 만들어낼 수 있으며 이러한 사전분포를 주관적 사전분포(subjective prior) 또는 정보 사전분포(informative prior)라 한다. 사전정보가 있다고 하여도 대부분의 경우 모수의 평균이나 분산 등의 단편적인 정보만을 제공하므로 주어진 사전정보를 이용하여 모수에 대한 확률분포를 완벽하게 추정하는 것은 쉬운 일이 아니다. 모형에 대한 충분한 사전정보 없이 사전분포를 형성할 경우 논리적인 모순을 최소화할 수 있는 정보를 사용하는 것이 필요하며 이러한 사전분포를 객관적 사전분포(objective prior)

또는 무정보 사전분포(non-informatiove prior)라 한다. 균일분포(uniform distribution)는 대표적인 비 정보적 사전분포로 사용된다.

모수에 대한 주관적 사전정보와 자료로부터 얻은 객관적 정보를 종합하여 모수의 확률 통계적 성격을 반영하는 사후분포를 얻을 수 있다. 따라서 베이지안 추론의 근간은 전적으로 사후분포에 의존하여 이루어진다. 즉, 우도함수에 내포된 자료의 정보와 사전밀도함수에 내포된 사전정보를 베이즈 정리에 의해 합성함으로써 사후분포 또는 사후밀도함수를 구할 수 있으며, 그 결과는 우도함수와 사전밀도함수의 곱에 비례한다. 베이지안의 실제 적용 시 근간이 되는 사후밀도함수는 함수 형태의 정보만 제공하는 경우가 많기 때문에 사후 추정을 위한 베이지안 계산의 절차는 다음과 같다.

(1) $x_{1,}\cdots,x_n$ 은 분포 $f(x|\theta)$ 에서 나온 관측치이다.

(2) θ 에 대한 사전정보를 알고 있다.

(3) 사전정보와 관측값을 함께 이용하여 θ 를 추정한다.

즉, 현재 가지고 있는 자료인 $x_1,\cdots,x_n$ 이 관측되기 이전의 θ 에 대한 과거 경험에서 얻은 정보인 사전분포 $f(\theta)$ 와 표본정보인 $x_1,\cdots,x_n$ 을 함께 고려하여 미지의 모수인 θ 를 추정하는 것이 베이지안 방법이며, 따라서 $f(\theta)$의 선택에 따라 추정 결과가 달라진다(송정무, 2013).

사후분포 $f(x|\theta)$ 는 θ 에 대한 사전 정보와 관측된 표본을 종합한 분포로서 신뢰할 수 있는 추정이 가능하다.

$$f(\theta,x) = f(x|\theta) \times f(\theta) \qquad (2)$$
$$= f(\theta|x) \times f(x)$$

식 (2)의 함수 $f(\theta,x)$는 확률 변수 θ와 x의 결합함수(joint density function)이며, $f(x)$와 $f(\theta|x)$는 개별변수의 주변 확률분포(marginal distribution)와 조건부 확률분포(conditional distribution)을 나타낸다. 위의 식은 다음과 같이 재표현 될 수 있다.

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{f(x)} \qquad (3)$$

이 때, $f(x)$를 상수취급하고 $f(x|\theta)$를 우도함수 형태로 바꾼다면 위의 식은 다음의 식 (4)의 형태가 된다.

$$f(\theta|x) \propto l(\theta|x) \times f(\theta) \qquad (4)$$

위의 식을 통해 사후확률분포는 사전확률분포에 대한 정보와 표본으로부터 얻은 정보를 결합하여 얻을 수 있음을 확인할 수 있다.

2) 문항별 반응에 근거한 사후기대추정법

검사에 응답한 피험자의 모집단에서 얻을 수 있는 피험자들의 능력 모수는 정보로서 사용될 수 있다. 이러한 피험자들의 능력 분포에 대한

정보는 문항반응 분포와 결합할 수 있다. 이론적으로 모집단 분포는 피험자들이 문항에 응답하기 전에 사용되는 정보로서 사전 정보라고 불린다. 그리고 피험자들이 응답한 문항에 대한 정보를 통해 우도를 도출할 수 있다. 이렇게 생성된 모집단 분포와 우도의 곱은 공액 또는 결합우도(joint likelihood)로서 사후분포를 이룬다.

EAP 방법은 어떤 능력모수의 사후분포의 기댓값(즉 평균)을 그 능력 추정치로 삼는다. 어떤 능력모수의 사전분포에서 비롯되는 사후분포의 밀도함수는 피험자의 문항반응 u와 모집단 분포에 기초하여 정의된다.

각 능력수준에서 피험자들이 i문항에 응답할 확률을 나타내는 우도는 식 (5)에 제시된 것처럼 능력모수 θ의 모든 수준에서의 발생 가능을 합하여 얻어진다. 따라서 이는 문항의 형태와 피험자의 반응 양식에 따라 다양한 함수를 가질 수 있다.

$$L = \prod_{i=1}^{n} T_i(u_i \mid \theta)\Phi(\theta) \tag{5}$$

$i = 0, 1, \cdots, I$

$u = \{u_1, u_2, u_3, \cdots\}$

$\Phi(\theta) =$ 모집단분포

$T_i(u_i \mid \theta) =$ 능력수준에 따라피험자가응답할수있는문항i의 반응형태

우도는 이론적으로 함수의 기댓값을 계산하여 얻은 결과이며 따라서 이 값은 관찰된 자료에서 기댓값을 계산하여 얻을 수 있는 실제 결과와 동일한 값을 갖기 힘들며 대게 약간의 차이를 보이기 마련이다. 관찰된 자료의 기댓값은 관찰빈도에 대한 각 관찰된 값의 합인 반면에 이론적

함수에서는 모든 능력 수준에서 어떤 우도를 구할 수 있으며, 어떤 두 적분값의 비율로서 기댓값을 구할 수 있다. 식 (6)에 제시된 후자의 방법을 이용하여 사후 밀도의 기댓값을 추정하는 것이 사후기대추정법이다(Bock & Mislevy, 1982).

$$EAP[\theta] = \frac{\int_{-\infty}^{\infty} \prod_{i=1}^{n} T_i(u_i)\Phi(\theta)\theta d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} T_i(u_i)\Phi(\theta) d\theta} \tag{6}$$

결과적으로 EAP 추정치는 사후분포의 기댓값으로 정의된다. 그러나 이 해는 열린 형태의 적분을 포함하고 있기 때문에 구분구적법을 이용하여 식 (7)과 같이 근사적으로 구한다.

$$EAP[\theta] \approx \frac{\sum_{1}^{q} \prod_{i=1}^{n} T_{iq}(u_i)\Phi(\theta_q)\theta_q d\theta_q}{\sum_{1}^{q} \prod_{i=1}^{n} T_{iq}(u_i)\Phi(\theta_q) d\theta_q} \tag{7}$$

근사적분을 위해 q개의 구분점을 지정하면 각 적분은 주어진 구분점의 합으로서 추정되며 각 가분에 대한 가중치를 설정한다. 이렇게 구해진 EAP 분포의 표준편차는 식 (8)을 이용하여 구할 수 있다.

$$SD[\theta] \approx \left(\frac{\sum_{1}^{q} \prod_{i=1}^{n} T_i(u_i)\Phi(\theta_q)(\theta_q - EAP[\theta])^2 d\theta_q}{\sum_{1}^{q} \prod_{i=1}^{n} T_i(u_i)\Phi(\theta_q) d\theta_q} \right)^{1/2} \tag{8}$$

사전분포가 표준정규분포를 따를 경우, 적분을 정확하게 수행하기 위해서 가우스-헤르마이트 구분구적법(Gauss-Hermite quadrature)을 사용할 수 있다. EAP 방법은 적절한 사전분포가 부여되는 한, 모든 경우의 문항 벡터에 대한 능력모수 추정치의 산출이 가능하다. 이렇게 추정된 값은 사전분포의 평균 쪽으로 축소되는 경향을 보이는데 이는 베이지안 추정치의 특성이다. 이로 인해서 피험자가 보일 수 있는 모든 가능한 문항 반응 패턴에 대한 능력모수 추정치를 제공하는 EAP 능력추정 방법은 적절한 사전분포가 사용되지 않았을 때, 왜곡된 능력모수를 추정할 수 있다는 단점이 있다(김성훈, 2010).

3) 검사총점에 근거한 사후기대추정법

다양한 문항의 종류에 따라 문항들은 상대적으로 다른 가중치와 다양한 범위의 난이도를 가지기 때문에 검사총점에 근거한 결과를 추정하는 것은 간단한 일이 아니다. 이를 위해서는 검사 문항들이 일차원성 가정을 충분히 만족시켜야만 한다. 따라서 검사총점에 근거한 사후기대추정법의 수치적 계산은 복잡하다.

총점 j가 문항별 반응형태에 대한 총합일 때, 이에 대한 우도는 아래의 식 (9)로 표현되며

$$L_j(\theta) = \sum_{j=\sum k_i}^{patterns} L(k|\theta) \tag{9}$$

i = 문항번호

$k = 0, \ldots, K_i$, 서열화된 문항 점수

$j = \sum k_i$, k의 총점

각 문항별 반응형태는 다음의 식과 같이 나타낼 수 있다.

$$L(k|\theta) = \prod_i T_{k_i}(\theta)\phi(\theta) \tag{10}$$

$T_{k_i}(\theta) = k$개의 범주를 가진 문항 i의 범주형 반응 함수

$\phi(\theta) =$ 모집단 밀도함수

그러므로 각 점수에 대한 우도는

$$L(\theta) = \sum_{j=\sum k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta) \tag{11}$$

이며, 각 점수 j에 대한 확률은 아래의 식 (12)와 같다.

$$P_j = \int L_j(\theta)\, d\theta \tag{12}$$

$$= \int \sum_{j=\sum k_i}^{patterns} L(k|\theta)\, d\theta$$

$$= \int \sum_{j=\sum k_i}^{patterns} T_{k_i}(\theta)\phi(\theta)\, d\theta$$

제시된 알고리듬의 식 (9)를 피적분함수로 활용함으로써 각 점수를 고

려한 EAP와 이에 대한 표준편차를 계산할 수 있다(Bock & Mislevy, 1982).

$$EAP(\theta|j=\sum k_i)=\frac{\int \theta L_j(\theta)d\theta}{P_j} \tag{13}$$

$$SD(\theta|j=\sum k_i)=\left\{\frac{\int \left[\theta-EAP(\theta|j=\sum k_i\right]^2 L_j(\theta)d\theta}{P_j}\right\}^{1/2} \tag{14}$$

IRT 모형이 주어진 자료에 적합하면, 식 (12)는 히스토그램상에서 백분율표와 같은 기능을 한다. 따라서 실제 검사를 진행하기 전에 시험용 문항을 이용하여 검사 총점에 근거한 백분율표를 구조화할 수 있으며, 이를 관측된 자료의 점수 분포와 비교함으로써 모형의 적합도를 통계적으로 진단할 수 있다(Thissen, 1995).

Lord와 Novick(1968)은 θ 의 함수로써 검사 총점에 의한 우도를 계산하기 위해 Walsh(1963)가 제안한 근사법(Approximation Method)를 사용할 것을 주장하며 간단한 재귀적 알고리듬을 제시하였다. 알고리듬은 분배의 법칙을 따르며, 어떠한 반응 범주 개수를 가진 문항에도 적용 가능하다.

$$L_j^n(\theta) = \sum_{j=\sum k_i}^{patterns} \prod_i T_{k_i}(\theta) \tag{15}$$

$n =$ 문항의 개수

$i = 0, 1, ..., n$

$k = 0, 1, ..., K_i$, 문항 i의 응답범주

$T_{k_i}(\theta) =$ 문항 i의 범주 k에 대한 함수

일련의 문항 $[0...n^*]$ 의 검사 총점은 $j = 0, 1, ..., \sum_{n^*}(K_i)$ 이고 일련의 문항 $[0...n^*]$ 의 검사 총점 j 에 대한 우도는 $L_j^{n^*}(\theta)$이며, 모집단 분포는 $\phi(\theta)$ 이다.

일반화된 재귀적 알고리듬은 다음의 과정을 따른다.

1) $n^* = 0$ 을 가정한다.

2) $L_j^{n^*}(\theta) = T_{jn^*}(\theta)$

$j = 0, 1, ..., K_{n^*}$

3) $L_{j+k}^{n^*+1}(\theta) = \sum_{k_{n^*+1}} L_j^{n^*}(\theta) T_{k_{n^*+1}(\theta)}$

$j = 0, 1, ..., \sum_{n^*}(K_i)$

4) $n^* = n^* + 1$ 을 가정한다.

5) $n^* = n$ 을 충족할 때까지 3) ~ 4) 의 과정을 반복한다.

제시된 재귀적 알고리듬을 통해 모집단 분포인 $\phi(\theta)$ 로부터 총점 j 에 대한 우도인 $L_j(\theta) = L_j^n(\theta)\phi(\theta)$ 를 구할 수 있으며, 그 결과 $L_j(\theta)$ 를 이용하여 $EAP(\theta|j = \sum k_i)$ 와 $SD(\theta|j = \sum k_i)$, $P_j(\theta)$ 를 계산 할 수 있다.

원칙적으로 위의 알고리듬은 범주화 반응을 가진 어떠한 모형에도 사용가능하며 정확한 결과를 도출한다. 하지만 반응을 서열화 시키지 않은 문항에 재귀적 알고리듬을 사용한다면, 우도함수를 형성할 때, 응답 반응이 동등한 가치의 θ 값 근처에 집중되지 않게 되므로 이러한 우도함수는 결과적으로 매우 큰 표준편차를 가지게 된다.

하지만, 제시된 재귀적 알고리듬은 보편적으로 사용 가능하다. 간단한 프로그래밍을 통해, 적분값을 도출할 수 있으며, IRT에서는 조금 더 복잡한 방법인 가우스-에르미트 구분구적법을 사용하여 값을 구할 수 있다 (Stroud, 1974). 사전분포와 우도함수의 공액은 어떤 사전분포를 사용하는지에 크게 영향을 받으며, Stroud는 사전분포를 결정하기 위해 가능한 다양한 구분구적법을 사용할 것을 제언한다.

2. 다분 문항반응이론 모형

다분 문항반응이론(polytomous Item Response Theory: polytomous-IRT)은 초기에는 모수추정, 모형적합도 관련연구가 주를 이루었으며

1969년 Samejima의 등급반응모형(Graded Response Model: GRM)의 발표를 기점으로 새로운 모형의 개발과 적용에 관한 연구가 시작되었다. 특히 1990년에 이르러 수행평가와 논술형 문항에 대한 관심이 상승하며 상, 중, 하 등의 등급점수를 주어야 할 때 사용 가능한 모형들이 필요로 하게 되면서 다분 문항반응모형에 관한 연구들이 본격화되기 시작하였다.

자료의 성격에 따라 다분 문항반응모형은 명명반응모형과 서열반응모형으로 분류할 수 있다(Dodd, De Ayala, & Koch, 1995). 명명반응모형은 반응지의 수치가 척도상에서 의미가 없는 특성만을 가진 답지로 구성된 문항에 사용할 수 있는 모형이다. 대표적인 명명반응모형은 명명반응모형(Norninal Response Model: NRM; Bock, 1972)과 선다형 모형(model for multiple-choic items; Tissen & Steinberg, 1984) 등이 있다. 서열반응모형은 반응지의 수치가 척도상에서 서열척도(ordinal scale)를 가진 답지로 구성된 문항에 사용할 수 있는 모형으로 대표적인 서열반응이론 모형은 등급반응모형(Graded Response Model: GRM; Samejima, 1969), 부분점수모형(Partial Credit Model; PCM; Masters, 1982), 평점척도모형(Rating scale model: RSM; Anrich, 1978), 일반화부분점수모형(Generalized Partial Credit Model: GPCM; Muraki, 1992), 연속모형(Sequential model; Tutz, 1990) 그리고 다국면 라쉬모형(Many-faced Rasch model; Linacre, 1989) 등이 있다. 서열반응모형들은 수행평가에서 부분점수를 부여할 때나 리커르트 척도상의 태도검사에서 반응지에 따른 점수를 부여할 때 유용하게 사용 가능하다.

본 연구는 여러 다분 문항반응모형 가운데 GRM과 GPCM 모형을 사용하여 두 EAP 능력모수 추정방법의 복원력 정확성을 비교하고자 한다.

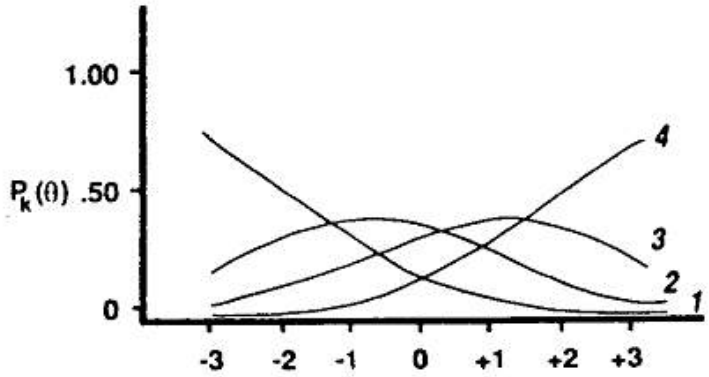
1) 등급반응모형(Gradid Response Model)

등급반응모형은 기존의 2모수 로짓 모형을 다분화된 문항반응에 적용할 수 있도록 Samejima(1968)가 고안한 다분반응모형이다. 이전의 문항반응모형은 4지선다나 5지선다 문항처럼 한 문항내에 응답할 수 있는 범주의 수가 다분화되어 있음에도 정답을 하였으면 1, 오답을 하였으면 0으로 이분화하여 문항을 분석하였다. 이에반해 등급반응모형은 반응의 범주를 m 개로 분류하여 정답과 오답 뿐 아니라 그 사이의 피험자들의 중간 능력을 함께 고려하였다. 따라서 피험자의 반응을 m 개의 범주로 분류하였을 때, 첫 번째 범주는 가장 낮은 등급이 되고 m 번째 범주는 가장 높은 등급이 되며 중간의 범주들은 그 안에서 서열성을 갖는다. 예를 들어, 어떤 수행 과제에 대해서 0, 1, 2, 3점으로 부분점수를 인정하는 채점방식을 택한다든지, 심리검사에서 리커르트 척도에 대해 '매우불만족', '불만족', '보통', '만족', '매우만족'으로 응답의 범주가 서열성을 갖도록 척도화하면 피험자의 반응은 등급반응이 된다(성태제, 1998).

이분반응모형은 범주의 수가 2개인 등급반응모형의 특별한 경우로서, 등급반응모형은 이분반응모형의 확장된 형태로 설명될 수 있다. 문항변수 Γ_i 가 한 문항에 대한 피험자의 반응경향을 나타내는 연속변수이고, r_k 는 문항에 대한 피험자의 각 반응을 구별하는 기준치라고 할때, 각 능력수준에서의 피험자들의 반응에 대한 조건분포들은 기준치 r_{k-1} 에 의해 분화된다. 한 피험자의 문항변수가 기준치 r_{k-1} 과 r_k 사이에 존재할때 피험자의 반응은 k 번째 범주에 속할 것으로 기대한다. 즉 능력 θ_j 를 가진 피험자의 반응이 k 번째 범주에 속할 확률은 기준치 r_{k-1} 과 r_k 사이에 있는 Γ_i 에 대한 조건분포의 면적인 $P_k(\theta_j)$ 된다. 첫 번째 범주에서 m 번째 범주까지 각 능력수준 범주에 속할 확률을 구할 수 있으며 그 합은 1이 된

다.

각 반응범주별로 전 능력수준 범위에 대하여 $P_k(\theta_i)$를 연결하면 문항반응범주특성곡선(item response category characteristic curve)를 얻을 수 있다. [그림 3]과 같은 곡선은 문항반응범주의 기능특성곡선(operating characteristic curve)이라고도 불리며(Samejima, 1969), 문항범주특성곡선(item category characteristic curve) 또는 문항범주흔적곡선(trace curve)이라고도 불린다(박정, 2001).



[그림 1] 문항반응특선 곡선

능력수준에 따라 피험자가 각 범주를 선택할 확률은 달라진다. 한 문항 내에서 선택 가능한 범주들은 서열화되어있고 한 능력수준에서 모든 범주에 대하여 피험자가 각 범주에 속할 확률을 모두 합하면 1이 될 때, 능력이 낮은 피험자 집단일수록 낮은 등급의 범주에 반응할 확률이 크며, 능력이 높은 집단에서는 높은 등급의 범주에 반응할 확률이 크게 나타난다. 문항반응범주특성곡선을 각 능력수준별로 누적하여 표현할 수 있다.

2) 일반화 부분점수 모형(Generalized Partial creditl model)

일반화 부분점수 모형은 Muraki(1992)가 부분점수모형에 문항 변별도 지수를 사용할 수 있도록 제안한 모형이다. 부분점수모형은 이분 반응모형의 하나인 Rasch 모형의 확률 원리를 Masters(1982)가 다분 문항반응으로 그대로 확장한 것이다.

부분점수모형의 응답지 선택 확률은 이분화 원리에 따른다. 즉, 여러 개의 선택 가능한 응답지가 있을지라도 응답지에 반응할 확률은 인접하는 두 개의 응답지 중에서 하나를 선택하는 것과 같은 확률이라고 정의한다. 따라서 부분점수 모형은 선택 가능한 모든 응답지의 범주를 고려하여 모형을 형성하기보다는, 인접한 두 범주만을 고려하여 모형을 구성하게 된다. j 문항이 m 개의 범주를 가지고 있을 때, 능력이 Θ 인 피험자가 k 번째 응답지에 반응할 확률은 k-1 과 k 중에서 k 를 선택할 조건부 확률과 같으며 전체 응답지에 대한 확률은 모두 더하여 1이 되어야 한다.

문항의 변별도 a_j 는 부분점수모형에 포함되지 않으며 모든 문항에 대하여 상수 1로 고정한 것처럼 취급한다. “문항범주 난이도” 또는 “문항단계 난이도”라고 불리는 b_{jk}는 문항 j 에서 k 라는 점수를 받을 때의 어려운 정도를 의미한다. 부분점수모형의 b_{jk}는 두 개의 이웃한 응답지 중에서 하나를 선택하는 것을 의미하므로, 범주 난이도라고 부르기 보다는 단계 난이도라고 부르기도 한다(박정, 2001). 범주 난이도 b_{jk}는 범주 k 를 선택할 확률값과 범주 k-1 을 선택할 확률값이 같은 지점으로 두 범주특성곡선이 만나는 지점의 값이 된다. Masters는 “범주 난이도(category difficulty)”라는 개념 대신 “단계 난이도(step difficulty)”라는 개념을 도입하였으며 서열적인 범주점수(ordered category scores)란 한 검사 또는

한 문항 내에서 성공적으로 이수해야할 하부단계라고 정의하였다. 그러나 각 단계의 난이도는 단계의 순서와 비례하는 것은 아니다. 예를 들어 1단계를 성공적으로 수행하는 것이 2단계를 성공적으로 수행하는 것보다 어려울 수 있다. 그러나 점수는 단계를 성공적으로 수행한 횟수와 비례해서 받는 것이며, 난이도의 서열로 받는 것은 아니다. 이와 같은 가정을 하게 될 때 사용할 수 있는 모형이 부분점수모형으로, 수행의 순서를 서열화한 모형이다(박정, 2001). 부분점수모형하에서 3단계의 범주점수를 받는다는 것은 1단계와 2단계를 성공적으로 수행하였음을 의미한다. 이러한 응답 범주는 태도검사에서 주로 사용되는 리커르트 척도처럼 “매우 불만족”에서 “매우 만족”의 범위를 갖을 수 있다(Dodd & Koch, 1987).

일반화부분점수모형은 부분점수모형을 확장한 모형으로 승산비의 가정하에서 전개된 모형이다. 즉, 일반화부분점수모형은 등급반응모형처럼 문항 변별도 지수를 사용할 수 있으나, 등급반응모형처럼 범주 난이도 지수가 서열화 되어 있다는 가정을 할 필요는 없다. 동시에 부분점수모형의 특성인 단계난이도의 서열성을 가정하지 않기 때문에 현실에서 많이 사용되고 있다. 즉, 일반화부분점수모형은 등급반응모형의 장점과 부분점수모형의 특징을 모두 가지고 있다고 할 수 있다(박정, 2001).

3. 선행연구

문항반응모형을 사용하여 피험자의 능력을 추정하면 문항에 따른 특정 가중치를 더해진 척도화된 능력 점수가 보고된다. 이는 여러 검사 문항들의 반응양식 정보를 모두 사용하기 때문에 정보의 손실이 적다는 장점이

있지만, 일반 대중에게 설득력이 떨어진다는 단점이 있다. 이러한 상황속에서 원점수와 일대일로 대응하는 능력 추정치를 산출하는 방법에 대한 연구가 관심을 받고 있으며 특히 문항별 반응에 근거한 사후기대추정법과 총점에 근거한 사후기대추정법의 비교 연구를 자주 볼 수 있다.

강태훈과 백순근(2007)은 3모수 문항반응이론 하에서 MLE(maximum likelihood estimate), MAP(maximum a posteriori), EAPrp, 및 EAPss의 네 가지 능력모수 추정 방식을 활용하여 검사의 문항곤란도 구성이 서로 다른 3가지 실제 검사 자료와 4가지 모의실험 자료를 분석할 때 능력 추정의 표준오차가 어떻게 달라지는지를 분석하였다. 그 결과 EAPrp의 경우가 능력 추정의 표준오차가 상대적으로 작고, 또 적게 변하는 것으로 드러났지만 EAPss 사용으로 인한 정보의 손실 혹은 능력 추정의 표준오차가 그리 크지 않다고 보고하고 있다. 하지만 이 연구는 구체적인 차이를 수치로 제시하지 않고 있으며 제한적인 모의실험 조건 하에서 연구를 진행하였다는 제한점이 있다.

강태훈(2014)은 다양한 모의실험 조건 속에서 각 방법이 나타내는 진모수 복원력을 비교함으써 '검사총점에 근거한 사후기대추정법'이 정보 손실에 따라서 갖게 되는 문제의 정도를 경험적으로 확인하고자 하였다. 그 결과, 1모수 로지스틱 모형하에서는 두 방법간의 차이를 발견할 수 없었지만 모형이 복잡해짐에 따라서 여전히 작긴 하지만 상대적으로 좀 더 분명하게 문항별 반응양식에 다른 사후기대추정법이 보다 정확한 능력모수 추정을 가능하게 하는 것으로 나타났다. 그러나 두 방법을 상관계수 및 MSE 등을 통하여 비교한 결과 그 차이는 매우 미미하였다.

김성훈(2012b)의 연구는 Kolen과 Tong(2010)의 추수연구(follow-up study)로서, ML, TCF, EAPrp와 EAPss의 능력 추정량의 측정학적 특성을 보다 심도 있게 탐구하였다. 그 결과, EAPrp와 EAPss 추정량의 편향

은 능력모수와 부적 상관을 보였다. 능력 수준별및 주변적 측정의 정확성이 있어서 문항반응패턴에 근거한 EAP 추정량은 검사 총점에 근거한 EAP 추정량보다 근소하게 우수한 기능을 보였다.

Kolen와 Tong(2010)은 문항반응이론을 사용하여 피험자의 숙달 정도를 축정할 때 측정학적 적절성에 대하여 검사 총점에 근거한 방식과 문항별 반응 양식에 근거한 베이지안 방식 및 비(非)베이지안 방식을 연구하였다. 연구에 의하면, 측정도구의 선택은 점수와 피험자의 숙달 정도의 분포에 영향을 미치며 베이지안 방식와 비(非)베이지안 방식의 선택이 검사 총점에 근거한 방식이나 문항별 반응 양식에 근거한 방식의 선택보다 이에 더 심각한 영향을 미친다.

Thissen 외(1995)는 다분 문항을 포함한 검사에서 점수의 척도화에 관한 연구를 수행하였다. 그 결과, EAPrp와 EAPss의 표준오차의 가장 작은 증가율 차이는 대략 10%정도였다. 이 10%의 정확성의 손실은 문항반응에 근거한 EAP 방식에 비하여 총점에 근거한 EAP 방식이 감수하여야 할 부분이다.

Thissen & Orland(2001, pp.126-127)는 2모수 로지스틱 모형하에서 두 EAP 능력모수 추정방식에 따른 표준오차를 제시하였다. IRT의 능력모수 척도 상에서 대게 EAPss를 사용하여 능력을 추정하였을 경우 더 큰 표준오차를 나타내어, 문항별 반응양식에 대한 정보를 무시함에 따른 정보의 손실이 존재함을 보여주었다. 그러나 간혹 EAPrp가 더 큰 추정의 표준오차를 가지는 경우도 발견할 수 있었는데 이는 피험자가 쉬운 문항은 맞추고 보다 어려운 문항은 틀리는 방식으로 응답하지 않을 때 나타날 수 있는 현상이라고 보고 있다.

Tong과 Kolen,(2010) 다양한 조건에서 다양한 추정방법의 영향을 연구하였다. 그 결과, 다양한 추정방법들 사이에서 눈에 띄는 차이가 발견되

었으나 검사 길이가 길어지면서, 여러 추정방법들은 비슷한 추정치의 결과를 산출하였다. EAPss는 다양한 검사 조건에서 가장 다양한 추정치를 보이는 경향이 있었다.

EAPss와 관련된 기존의 심리측정학적 연구들은 EAPss를 사용할 때의 정보 손실의 정보를 보여주는데 그치고 있으며 다분 문항반응이론 하에서 EAPrp와 EAPss의 능력모수 추정 수행에 대한 진능력모수 복원의 정확성을 주변 분포와 능력 수준별 연구를 통하여 확인할 수 있는 연구는 쉽게 찾기 어려운 것으로 보인다. 따라서 본 연구에서는 다분 문항반응모형하에서 EAPss의 수행력과 EAPrp의 수행력을 두 추정치 간 상관계수, 진값과 추정치 간의 상관계수 그리고 평균제곱오차 등을 통하여 체계적으로 비교하고자 한다. 이는 EAPss를 선택하였을 때의 능력모수의 추정 정확도가 구체적으로 어느 정도인지를 체계적으로 비교함으로서 검사 자료 분석을 실시하는 연구자가 이를 감수하고 EAPss를 사용할지 또는 EAPrp를 사용할 지에 대한 선택을 도울 수 있을 것이다.

Ⅲ. 연구 방법

1. 연구 자료

본 연구는 모의실험 연구를 통하여, 다분 문항반응모형에서 EAPrp와 EAPss가 피험자들의 능력모수를 추정할 때의 상대적 정확성을 비교하는 데 목적이 있다. '적용 IRT 모형', '자료의 크기', '피험자 능력분포', '검사의 길이' 등 다양한 모의실험 조건이 두 EAP 능력모수 추정 방법에 미치는 영향을 알아보기 위하여 실제 자료를 바탕으로 자료를 생성하였다. 본 연구에서는 자료 생성을 위하여 Kang, T. H., Cohen, A. S., & Sung, H. J.(2005)이 사용한 40개의 문항에 대한 모수를 참조하였으며 이는 2000년도에 시행된 미국의 NAEP(National Assessment of Educational Progress) 평가의 8학년 수학영역의 자료로 부록 1과 부록 2에 제시된 바와 같다. 두 EAP 능력모수 추정은 MATLAB을 이용하여 작성한 프로그램을 통하여 이루어졌다.

2. 모의 실험 조건

두 EAP 방법의 진능력모수 복원의 정확성을 보다 명확히 이해하고 비교하기 위하여, 본 연구에서는 측정학적으로 상호보완적인 기능을 하는

능력 추정치들의 주변 분포와 능력 수준별 분포의 특성을 다양한 모의실험 조건을 통해 분석하였다.

1) 능력 추정치들의 주변 분포 사용 연구

두 EAP 방법의 수행 결과로 나온 능력 추정치들의 주변 분포를 비교하기 위하여, 다음과 같은 검사와 관련된 세 가지 모의실험 요인과 피험자의 진능력분포를 모의실험 요인으로 고려하였다.

가. 검사와 관련된 모의실험 조건

(1) 검사 길이 = 20 문항, 40 문항

(2) Polytomous IRT 모형 = GRM, GPCM

(3) 문항 범주 = 3, 5

나. 피험자와 관련된 모의실험 조건

(1) 표준정규분포 $N(0,1)$

(2) 균일분포 $U(-4,4)$

(3) 편포 $B4(3,12,-2.1,5.9)$

능력 추정치들의 주변 분포는 GRM와 GPCM 모형에서 3개의 문항 범주와 5개의 문항 범주를 갖는 20문항과 40문항 검사에 대해서 생성하였으며 40문항 검사는 20문항 검사의 모수를 반복하여 사용함으로써 생성하였다. 이 주변 분포의 생성을 위해, 다양한 능력 수준을 가진 충분한 수의 피험자들이 다루어질 수 있도록 N=10,000 명에 해당하는 능력모수들을 표준정규분포, 균일분포, 및 부적편포로부터 무선적으로 선발하였다. 이상의 모의실험 요인 및 편포를 만들기 위한 4모수-베타분포 모수 등은 김성훈(2012a)와 강태훈(2014)의 연구를 참조하였다. 피험자 집단에 대한 검사 자료는 선발된 능력모수들과 해당 검사의 문항모수를 이용하여 생성하였으며, 이 때 문항모수는 추정하지 않고 실제 자료의 문항모수를 사용하였다. 이를 통해, 각 능력 추정 방법을 적용하여 능력 추정치들의 주변 분포를 산출하였다. 결과적으로 총 24개의 모의실험 조건(= 3 능력분포 X 2 검사 길이 X 2 모형 X 2 문항 범주)을 고려하였다.

2) 능력 추정치들의 능력 수준별 연구

두 EAP 능력 추정 방법 각각에 대한 능력 추정치들의 능력 수준별 분포를 생성하기 위하여는 검사와 관련된 모의실험 요인과 피험자와 관련된 모의실험 요건을 고려하여 다음과 같은 문항반응 자료를 생성하였다.

가. 검사와 관련된 모의실험 조건

(1) 검사 길이 = 20 문항, 40 문항

(2) Polytomous IRT 모형 = GRM, GPCM

(3) 문항 범주 = 3, 5

나. 피험자와 관련된 모의실험 조건

(1) 17개의 능력 수준별 분포

먼저 검사와 관련된 모의실험 요건들은 능력 추정치들의 주변 분포를 생성하기 위한 능력 조건과 동일하며 다음과 같다. GRM와 GPCM 모형에서 3개의 문항 범주와 5개의 문항 범주를 갖는 20문항과 40문항 검사에 대해서 생성하였으며 40문항 검사는 20문항 검사의 모수를 반복하여 사용함으로써 생성하였다. 다음으로, 능력 수준별 분포의 생성을 위해 피험자의 능력모수 θ를 -4~+4 범위에서 0.5 간격을 두어 17개의 능력모수들을 선택하였다(Warm, 1989; 김성훈, 201a). 그 다음, 해당 검사의 문항모수들과 각 17개의 능력모수 구간에서 주어진 능력모수를 이용하여 각 구간에 1,000 개의 문항반응 관찰 벡터를 모의 생성하여 전체 17개 능력모수 구간에서 해당하는 능력 모수를 무선적으로 선발하였다. 이 때 문항모수는 추정하지 않고 실제 자료의 문항모수를 사용하였다. 결과적으로 총 8개의 모의실험 조건(= 1 능력 수준별능력 분포 X 2 검사 길이 X 2 모형 X 2 문항 범주)을 고려하였다.

3. 자료 분석

각 자료를 분석할 때 문항모수는 추정하지 않고 실제 자료의 문항모수를 사용하였는데, 이는 본 연구의 주된 관심이 두 EAP 능력모수 추정 방법의 비교에 있기 때문이다.

본 연구에서는 두 EAP 방법이 능력모수의 사전분포의 정보적 특성에 의해서 얼마나 영향을 받는지를 살펴보기 위해서 다음의 세 가지 사전분포를 고려하였다.

(1) 표준정규 사전분포 $N(0,1)$

(2) 균일 사전분포 $U(-3,3)$

(3) 표준정규 사전분포 $N(0,1)$을 사용하여 구한 피험자의 사후분포

표준정규분포는 EAP 방법으로 피험자의 능력모수를 추정할 때, PARSCALE 등의 각종 상업용 소프트웨어 등에서 가장 흔하게 사용하는 사전분포이다. 균일 사전분포는 정보가 가장 적은 사전분포(less information prior)로 다른 사전분포와의 비교를 위한 준거가 될 수 있다. 마지막으로 사후분포를 사전분포로 재사용함으로써 피험자의 능력분포가 정규분포와 큰 차이를 보일 경우, 정규분포를 사전분포로 사용한 것에 비하여 원래 능력분포의 모습에 조금 더 가까워진 형태의 사전분포로 기능할 수 있기 때문이다. 이러한 사후분포는 PARSCALE 등의 상업용 소프트웨어에서 산출되는 '*ph.2' 확장자를 가진 결과 파일에서 문항모수 추정 결과 발견할 수 있는 분포와 같다(강태훈, 2014).

두 EAP 추정치들의 주변 분포와 능력 수준별 분포의 특성을 모의실험을 통해 분석함으로써 세 가지 사전분포의 상대적 기능을 구체적 및 포괄적으로 진단할 수 있다. 능력 추정치들의 능력 수준별 분포와 주변 분포는 그 능력 추정치들에 대한 상호보완적인 측정학적 정보를 제공한다(김성훈, 2010). 주변부 분포와 관련된 두 EAP 능력모수 추정방법을 위한 모의실험에서, 사전분포는 표준정규분포 $N(0,1)$, 균일분포 $U(-3,3)$, 그리고 추정된 능력분포가 사전분포로서 재사용 되는 반면, 능력 수준별 분포와 관련된 두 EAP 능력모수 추정방법을 위한 모의실험에서는 사전분포로서 표준정규분포 $N(0,1)$와 균일분포 $U(-3,3)$만이 사용된다.

4. 연구 결과에 대한 평가 준거

능력 추정치들의 주변분포와 능력별 분포는 그 능력 추정치들에 대한 측정학적 정보를 상호보완적으로 제공한다. 먼저, 주어진 능력모수에서의 추정치들의 주변 분포를 분석하면 특정 검사를 다수의 피험자에게 실시하는 일반적인 검사 상황에서 그 추정치들의 능력모수에 대한 편향(bias; 편의, 편파성), 표준오차(standard error, 이하 SE), 그리고 능력모수와의 평균적인 오차(root of mean squared error. 이하 RMSE) 등을 파악할 수 있다. 다음으로, 여러 수준의 능력모수들에 대한 정보를 얻을 수 있는 능력 수준별 분포에서는, 각 능력 조건별로 능력 추정치의 주변 편향, 표준오차 그리고 능력모수와의 평균적인 차이 등의 특성을 파악할 수 있다(김성훈, 2010).

1) 주변 분포 사용 연구 평가 준거

두 EAP 능력모수 추정방법의 진능력모수 복원 정확성을 비교하고 두 추정 방법의 상호간에 얼마나 비슷한 추정치를 산출하는 지에 대한 경향을 파악하기 위하여 두 능력추정 방법 적용 결과 간의 상관계수를 산출하였다.

다음으로 두 가지 EAP 능력추정 방법의 기능을 비교하기 위해서 능력추정치들의 주변 분포에 대해서 다음과 같은 분석을 시행하였다. 먼저 주어진 주변 분포를 구성하는 추정치들($\hat{\theta}_g; g = 1, \ldots, 10000$)에 대해서 평균 $E(\hat{\theta})$, 표준오차 SE($\hat{\theta}$), 편향 Bias($\hat{\theta}$), 그리고 평균제곱오차의 제곱근 RMSE($\hat{\theta}$)를 다음 식들을 사용하여 계산하였다.

$$Bias(\hat{\theta}) = E(\hat{\theta}) - E(\theta) = \sum_{g=1}^{10000} (\hat{\theta} - \theta)/10000 \tag{16}$$

$$g = 1, \ldots, 10000$$

$$SE(\hat{\theta}) = \sqrt{\sum_{g=1}^{10000} [\hat{\theta} - E(\theta)]^2/10000} \tag{17}$$

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\sum_{g=1}^{10000} (\hat{\theta} - \theta)^2/10000} \tag{18}$$

이 값들은 다시 평균제곱오차(mean square error, 이하 MSE), 제곱편향(squared bias, 이하 SB), 그리고 분산(variance, 이하 VAR)로 표현할

수 있다. 이들은 MSE=SB+VAR의 관계를 가지며 MSE 값이 작을수록 더 정확한 추정을 한다고 볼 수 있다(Lee & Ban, 2010).

2) 능력 수준별 연구 평가 준거

주어진 능력모수 θ를 17구간으로 나누어 능력 수준별로 추정된 능력모수 추정치 $\hat{\theta}|\theta$에 대하여, 조건부 편향 Bias($\hat{\theta}|\theta$), 표준편차 SE($\hat{\theta}|\theta$), 및 능력 모수들의 평균적인 차이 RMSE($\hat{\theta}|\theta$)를 다음 식을 이용하여 구하였다.

$$Bias(\hat{\theta}|\theta) = E(\hat{\theta}|\theta) - \theta = \sum_{r=1}^{1000} \hat{\theta}_r / 1000 - \theta \quad (19)$$

$$r = 1, ..., 1000$$

$$SE(\hat{\theta}|\theta) = \sqrt{\sum_{r=1}^{1000} [\hat{\theta}_r - E(\hat{\theta}|\theta)]^2 / 1000} \quad (20)$$

$$\mathrm{RMSE}(\hat{\theta}|\theta) = \sqrt{\sum_{r=1}^{1000} (\hat{\theta}_r - \theta)^2 / 1000} \quad (21)$$

이들은 MSE값으로 표현할 수 있으며 작은 값을 가질수록 더 유사한 추정을 한다고 해석할 수 있다.

Ⅳ. 연구 결과

모의실험을 통해 산출한 두 EAP 추정치들의 결과를 주변 분포와 능력 수준별 분포로 나누어 제시한다. 본 연구에서는 그림과 표 등의 표기의 간략화를 위하여 사전분포의 이름을 다음과 같이 표기한다. 표준정규분포 사전분포는 “N-Prior”로, 균일 사전분포는 “U-Prior”로, 그리고 사후분포를 재사용한 사전분포는 “P-Prior”로 나타낸다.

1. 능력 추정치들의 주변 분포에 대한 결과

1) 두 EAP 능력모수 추정방법의 상관계수

EAPss의 능력모수 추정방법의 진능력모수 복원 정확성을 EAPrp와 비교하고 두 추정 방법의 상호간에 얼마나 비슷한 추정치를 산출하는 지에 대한 경향을 파악하기 위하여 두 능력추정 방법 적용 결과 간의 상관계수를 산출하였다. 제시된 <표 2>는 EAPrp와 EAPss 간의 상관계수를 나타내며, EAPss 능력모수 추정치가 EAPrp 능력모수 추정치와 얼마나 유사한 능력모수 추정치를 산출하는가를 확인할 수 있다. 또한 <표 3>과 <표 4>에서는 EAPrp와 EAPss의 각 방법에 의한 능력모수 추정치가 진능력모수와 갖는 상관계수를 확인할 수 있다.

<표 2>에서 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력

추정 방법이 GRM과 GPCM 모형 하의 모든 조건에서 1 또는 .999와 같은 1과 매우 가까운 상관계수를 보임을 확인할 수 있다. 다만 검사 길이가 40일 때가 검사 길이가 20일 때보다 1의 상관계수를 더 많이 나타내는 경향이 있었다. 문항 범주의 측면에서도 3개의 문항범주 조건보다 5개의 문항범주 조건에서 더 많은 1의 상관계수를 확인할 수 있었다.

<표 2> EAPrp와 EAPss간의 상관계수

모형	문항 범주	피험자분포	검사길이					
			20			40		
			사전분포					
			N-Prior	U-Prior	P-Prior	N-Prior	U-Prior	P-Prior
GRM	3	$N(0,1)$	.999	.999	.999	1.000	1.000	1.000
		$U(3,-3)$	1.000	1.000	.999	1.000	1.000	1.000
		$B4(3,12,-2.1,5.9)$	.999	.999	.999	1.000	1.000	.999
	5	$N(0,1)$	1.000	1.000	1.000	1.000	1.000	1.000
		$U(3,-3)$	1.000	1.000	1.000	1.000	1.000	1.000
		$B4(3,12,-2.1,5.9)$	1.000	.999	.999	1.000	1.000	1.000
GPCM	3	$N(0,1)$	.999	.999	.999	1.000	1.000	1.000
		$U(3,-3)$	1.000	.999	.999	1.000	1.000	.999
		$B4(3,12,-2.1,5.9)$	.999	.999	.999	1.000	1.000	1.000
	5	$N(0,1)$	1.000	1.000	1.000	1.000	1.000	1.000
		$U(3,-3)$	.999	.998	.999	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	1.000	1.000	1.000	1.000	1.000	1.000
사전분포별 평균(표준편차)			.999 (0.000)	.999 (0.000)	.999 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
전체평균(표준편차)								1.000 (0.000)

아래의 <표 3>을 통해 GRM 모형에서 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력모수 추정치와 진능력모수와의 상관계수는 모든 조건에서 1 또는 1과 매우 가까운 상관계수를 확인할 수 있었다. 다만 EAPrp의 적용 하보다 1에 일치하는 상관계수를 더 많이 산출하는 경향이 보였으나 EAPrp와 그 차이는 미미하였다. 특히 생성 피험자 분포 측면에서 볼 때, 정적편포일 때 상대적으로 작은 상관계수가 나타나는 경향이 있었다.

<표 3> GRM에서 EAPrp와 EAPss의 진능력모수와의 상관계수

능력추정조건	문항범주	피험자분포	검사길이					
			20			40		
			사전분포					
			N-Prior	U-Prior	P-Prior	N-Prior	U-Prior	P-Prior
EAPrp	3	$N(0,1)$	1.000	1.000	1.000	1.000	1.000	1.000
		$U(3,-3)$	1.000	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.995	.994	.996	.997	.997	.997
	5	$N(0,1)$	1.000	1.000	1.000	1.000	1.000	1.000
		$U(3,-3)$	1.000	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.996	.996	.997	.997	.997	.998
EAPss	3	$N(0,1)$	.999	.999	.998	1.000	1.000	1.000
		$U(3,-3)$	.999	.998	.999	.999	.999	1.000
		$B4(3,12,-2.1,5.9)$	.993	.992	.995	.996	.996	.997
	5	$N(0,1)$	1.000	1.000	1.000	1.000	1.000	1.000
		$U(3,-3)$	.999	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.995	.994	.996	.997	.997	.998
사전분포별 평균(표준편차)			.998 (0.002)	.997 (0.003)	.998 (0.002)	.999 (0.001)	.999 (0.001)	.999 (0.001)
전체평균(표준편차)								.998 (0.002)

GPCM 모형에서 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력모수 추정치와 진능력모수와의 상관계수는 <표 4>에 제시된 것처럼 모든 조건에서 1 또는 .992에서 .999의 범위를 갖는 1과 매우 가까운 상관계수를 보였으며 GRM 모형에서 두 EAP 능력모수 추정치와 진능력모수 추정치와의 상관계수와 매우 유사한 결과를 산출하였다. 즉, EAPrp의 적용하보다 1에 일치하는 상관계수를 더 많이 산출하는 경향이 보였으나 EAPrp와 그 차이는 미미하였으며, 특히 생성 피험자 분포 측면에서 볼 때, 정적편포일 때 상대적으로 작은 상관계수가 나타나는 경향이 있었다.

<표 4> GPCM에서 EAPrp와 EAPss의 진능력모수와의 상관계수

능력 추정 조건	문항 범주	피험자분포	검사길이					
			20			40		
			사전분포					
			N-Prior	U-Prior	P-Prior	N-Prior	U-Prior	P-Prior
EAPrp	3	$N(0,1)$	.999	.999	1.000	1.000	1.000	1.000
		$U(3,-3)$	1.000	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.994	.992	.995	.996	.995	.997
	5	$N(0,1)$	1.000	.999	1.000	1.000	1.000	1.000
		$U(3,-3)$	.999	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.996	.994	.996	.997	.997	.998
EAPss	3	$N(0,1)$	.998	.998	.999	.999	.999	1.000
		$U(3,-3)$	.999	.999	.999	.999	.999	1.000
		$B4(3,12,-2.1,5.9)$	.993	.989	.994	.995	.994	.996
	5	$N(0,1)$	1.000	.999	.999	1.000	1.000	1.000
		$U(3,-3)$	.999	.999	1.000	1.000	.999	1.000
		$B4(3,12,-2.1,5.9)$	.995	.993	.996	.997	.997	.997
사전분포별 평균(표준편차)			.998 (0.002)	.997 (0.003)	.998 (0.002)	.999 (0.002)	.998 (0.002)	.999 (0.001)
전체평균(표준편차)								.998 (0.002)

2) 두 EAP 능력모수 추정방법의 MSE

세 가지 조건의 사전 분포를 사용한 두 EAP 방법의 능력모수의 추정과 모의실험 자료 생성을 위해 사용된 다분 문항반응모형이 GRM과 GPCM일 때의 두 능력모수 추정 방법 각각에 대하여 MSE, SB, VAR 값을 구하였다. 표준정규 피험자분포 $N(0,1)$를 상정한 두 EAP 능력모수 추정치에 대한 MSE 값은 [그림 2], [그림 5]에 도표로 제시하였으며, 균일 피험자분포 $U(-3,3)$를 상정한 능력모수 추정 결과의 MSE 값은 [그림 3]와 [그림 6]에, 그리고 편포 $B4(3,12,-2.1,5.9)$로 피험자분포를 상정하여 EAPrp와 EAPss 방법으로 능력모수 추정치를 구한 결과에 대한 MSE 값은 [그림 4]과 [그림 7]에 도표로 제시하였다.

[그림 2], [그림 3] 및 [그림 4]은 GRM이 피험자 분포로 표준정규분포 $N(1,0)$, 균일분포 $U(-3,3)$, 그리고 편포 $B4(3,12,-2.1,5.9)$를 모의실험 자료 생성과 능력모수 추정을 위해 상정한 경우 EAPrp와 EAPss가 보여준 진능력모수 복원 정도를 MSE, SB, VAR 값을 통하여 제시하고 있다. [그림 5], [그림 6], 그리고 [그림 7]은 GPCM이 세 가지 피험자 분포를 사용하여 여러 조건을 가지고 두 EAP 능력모수 추정방법에 구해진 능력모수 추정치의 진능력모수값에 대한 복원력을 확인하기 위해 MSE, SB, VAR 값을 구한 도표이다. 여기에서 발견할 수 있는 특징을 정리하면 다음과 같다.

첫째, 여러 조건들 속에서 EAPrp가 EAPss에 비하여 조금 더 적은 MSE, SB, VAR 값을 산출하였다. 이는 다분 문항반응모형 조건하에서 능력모수를 추정할 때, EAPss를 사용하면 피험자의 원점수와 일대일로 대응시켜서 능력모수를 추정하는 반면, EAPrp는 피험자의 반응유형이라는 더 많은 정보를 사용하여 능력모수를 추정하기 때문에 생기는 당연한

결과이다. 그러나 두 EAP 능력모수 추정 방식에 의한 MSE, SB, VAR의 차이는 소수점의 자리에서 발생하는 작은 차이였다.

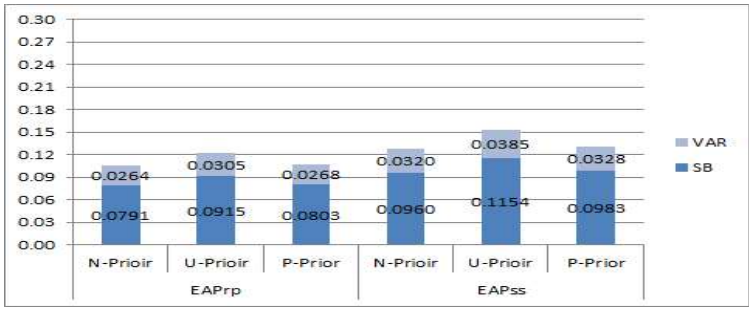
둘째, 베이지안 능력모수 추정을 위하여 사전정보를 표준정규분포 $N(0,1)$과 균일분포 $U(-3,3)$, 그리고 피험자분포를 사전분포로 재사용하였을 때, 각 사전분포에 따라 EAPrp와 EAPss의 능력모수 추정의 정확도에 차이가 있었다. [그림 2]과 [그림 3], 및 [그림 4]에서 볼 수 있듯이 균일분포를 사전정보로 사용하였을 때, 두 EAP 능력모수 추정치의 MSE가 증가한 것을 확인 할 수 있다. [그림 2]과 [그림 4]에서 보듯이 균일분포와 편포가 피험자 분포로 기능하면, "N-Prior">"P-Prior">"U-Prior" 순으로 더 정확한 능력모수 추정이 가능하였으나 [그림 3]에 의하면 균일분포로 피험자 분포를 상정한 경우, "P-Prior">"N-Prior">"U-Prior" 순으로 더 정확한 능력모수 추정이 가능하였다. 이는 적어도 GRM과 GPCM을 적용하여 베이지안 능력모수를 추정할 때는 균일 사전분포를 사용하는 것 보다는 표준정규분포나 피험자분포를 사전분포로 재사용하는 것이 상대적으로 더 안전한 선택일 수 있음을 함의한다.

셋째, 베이지안 능력모수 추정을 위하여 피험자 분포를 표준정규분포 $N(0,1)$과 균일분포 $U(-3,3)$, 그리고 편포 $B4(3,12,-2.1,5.9)$로 상정하였을 때, 각 피험자분포의 사용에 따른 능력모수 추정의 정확도에 차이가 있었다. 두 EAP 방법과 세 가지 사전정보 조건을 사용하여 능력모수를 추정하였을 때, 전반적으로 표준정규분포를 피험자 분포로 가지는 경우가 가장 작은 MSE 값을 가졌으며, 두 번째로 큰 MSE 값은 편포를 피험자 분포로 가지는 경우였다. 즉, 가장 큰 MSE 값을 갖게 하는 피험자 분포는 균일분포였다. 따라서 GRM과 GPCM을 사용하여 베이지안 능력모수를 추정할 때는 편포나 균일분포로 피험자 분포를 상정하기 보다는 표준정규분포를 상정하여 피험자 분포를 사용하는 것이 더 안전한 선택이 될

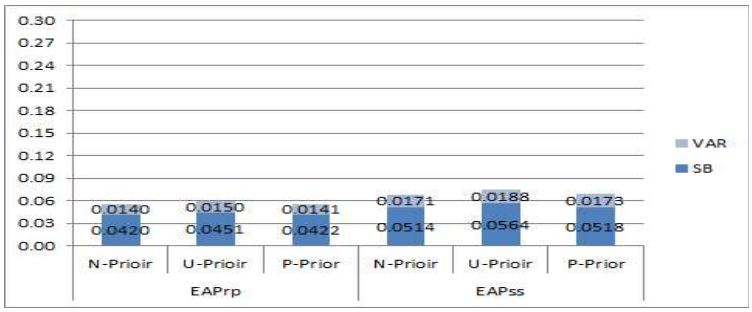
수 있다.

넷째, 문항 범주와 검사 길이가 증가하면 EAPrp와 EAPss를 사용하여 능력모수를 추정하였을 때, MSE값은 줄어든다. 이는 반응범주의 수가 증가할수록 척도의 신뢰도가 증가하기 때문이며(Guilford, 1954; Nunnally, 1978; Finn, 1972; Garner, 1960), 5점 척도나 6점 척도가 가장 신뢰롭다고 보고한 Mckelvie(1978) 연구 결과와도 일치한다. 또한 문항수의 증가는 검사정보함수가 선형적으로 증가하게 함으로써 측정의 정확성에 영향을 주기 때문에 신뢰도가 증가한다(김경희; 1993). 즉, 신뢰도의 증가는 MSE의 감소를 의미한다.

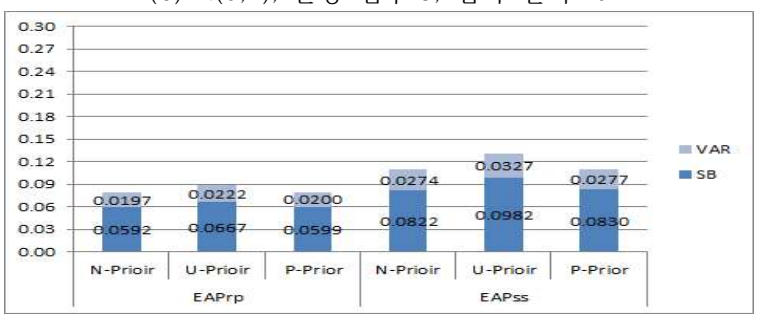
마지막으로, GRM과 GPCM 적용에 따른 두 EAP 능력추정치의 MSE 값에 차이가 존재한다. GPCM 하에서 표준정규분포를 피험자 분포로 사용한 [그림 5]와 부적편포를 피험자 분포로 사용한 [그림 7]을 살펴보면, 동일한 피험자 분포를 사용하였지만 GRM을 사용하여 능력모수를 추정한 [그림 2]와 [그림 4]와 MSE 값에 차이가 있음을 확인할 수 있다. 이들은 GPCM 하에서 더 적은 MSE값을 가지며 특히 문항 범주 수가 5인 (c)와 (d)의 경우 상대적으로 눈에 띄게 MSE가 줄었음을 확인할 수 있다.



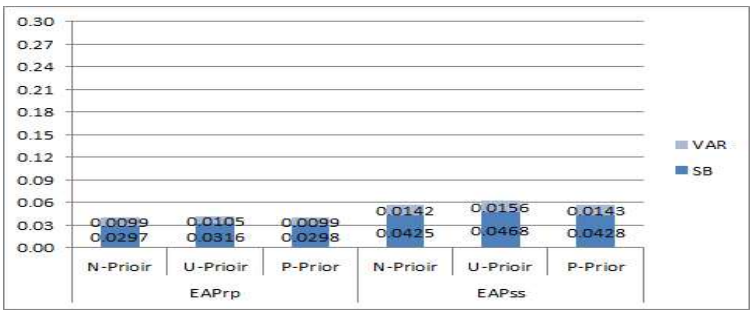
(a) N(0,1), 문항 범주:3, 검사 길이:20



(b) N(0,1), 문항 범주:3, 검사 길이:40

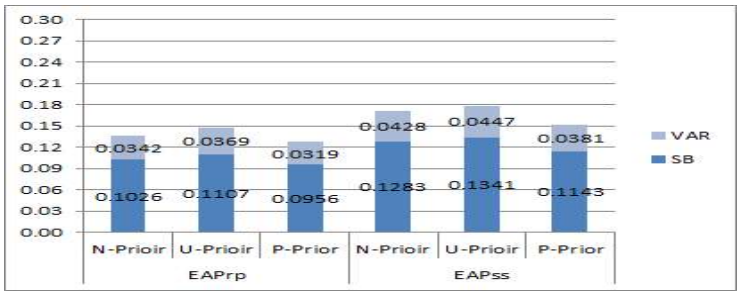


(c) N(0,1), 문항 범주:5, 검사 길이:20

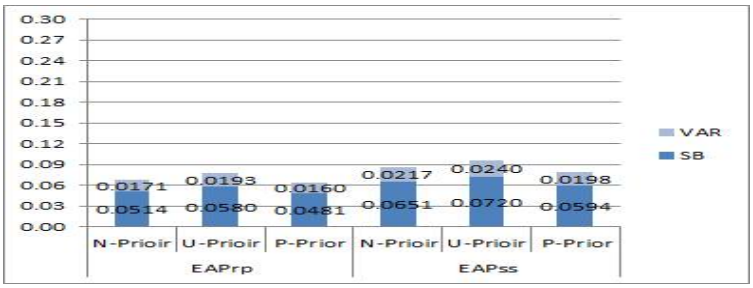


(d) N(0,1), 문항 범주:5, 검사 길이:40

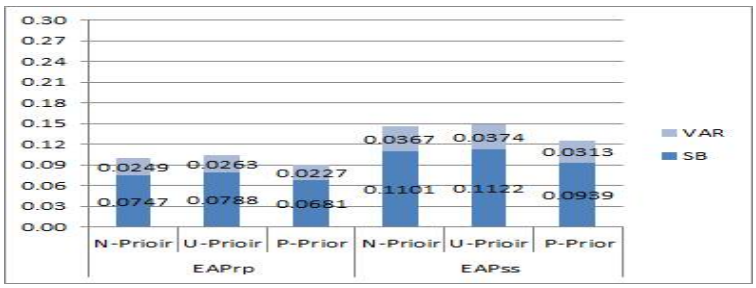
[그림 2] GRM에서 표준정규분포가 피험자 분포 일 때 각 조건에 따른 MSE



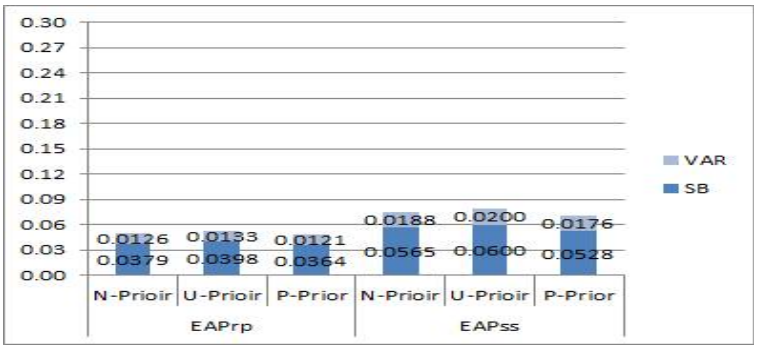
(a) U(-3,3), 문항 범주:3, 검사 길이:20



(b) U(-3,3), 문항 범주:3, 검사 길이:40

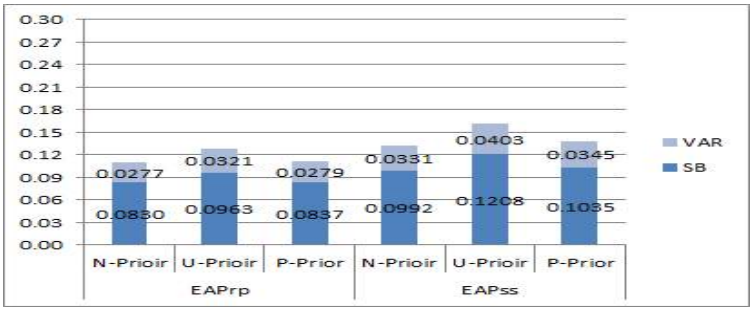


(c) U(-3,3), 문항 범주:5, 검사 길이:20

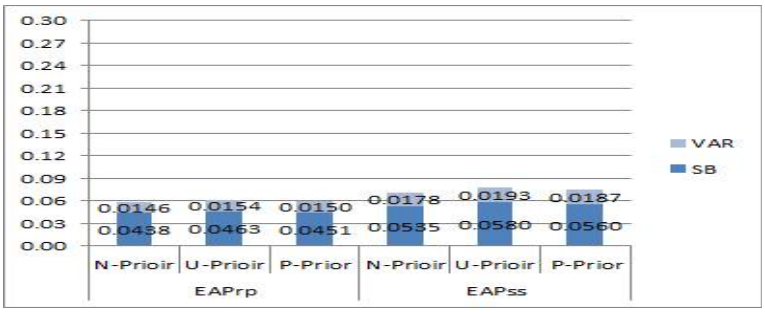


(d) U(-3,3), 문항 범주:5, 검사 길이:40

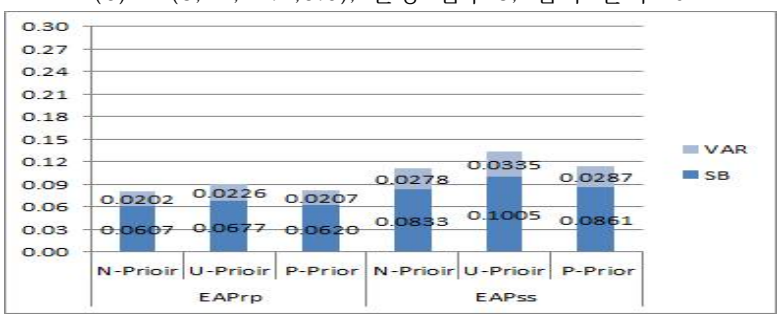
[그림 3] GRM 에서 균일분포가 피험자 분포 일 때 각 조건에 따른 MSE



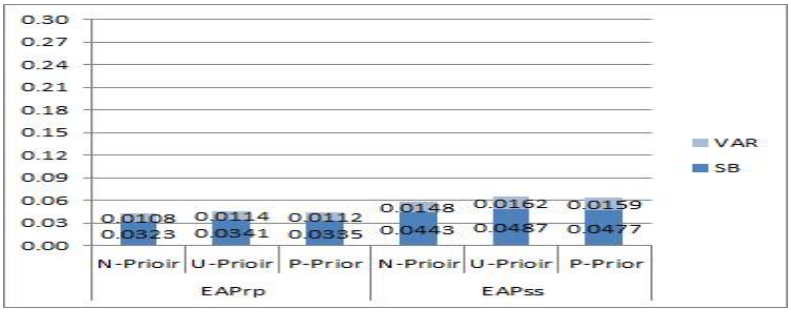
(a) B4(3,12,-2.1,5.9), 문항 범주:3, 검사 길이:20



(b) B4(3,12,-2.1,5.9), 문항 범주:3, 검사 길이:40

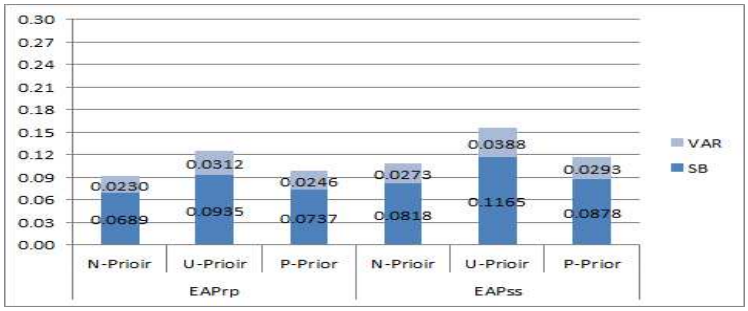


(c) B4(3,12,-2.1,5.9), 문항 범주:5, 검사 길이:20

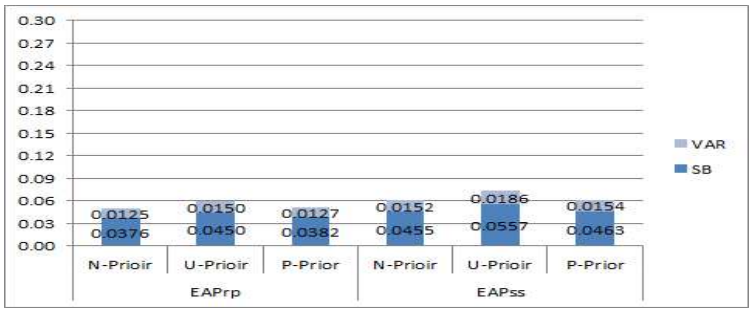


(d) B4(3,12,-2.1,5.9), 문항 범주:5, 검사 길이:40

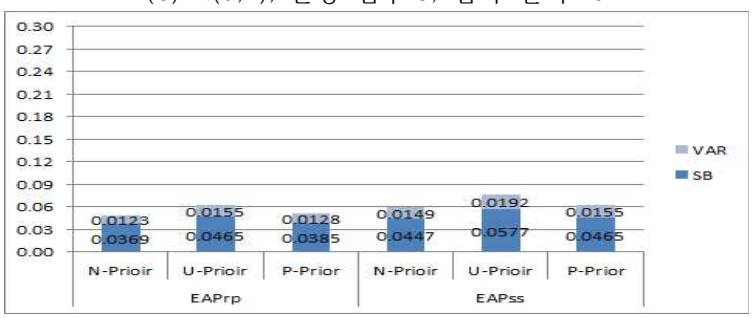
[그림 4] GRM에서 편포가 피험자 분포 일 때 각 조건에 따른 MSE



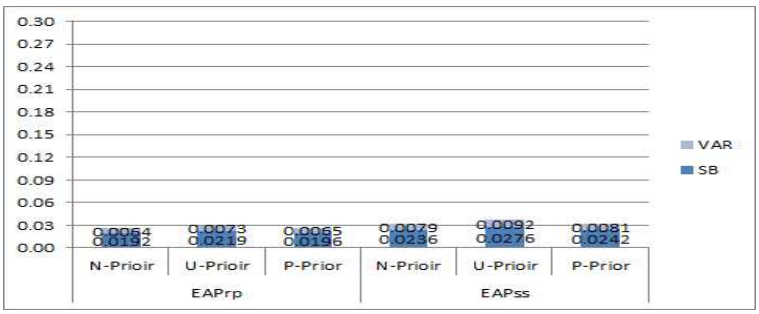
(a) N(0,1), 문항 범주:3, 검사 길이:20



(b) N(0,1), 문항 범주:3, 검사 길이:40

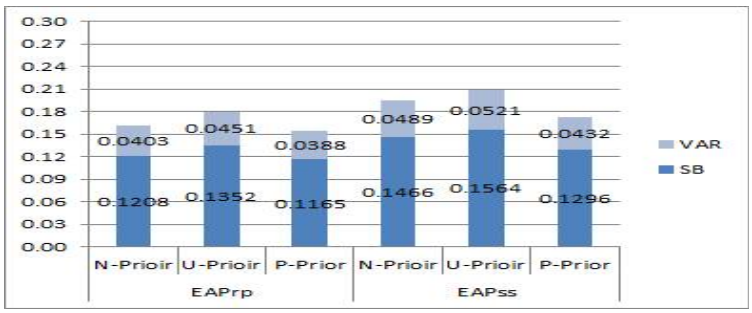


(c) N(0,1), 문항 범주:5, 검사 길이:20

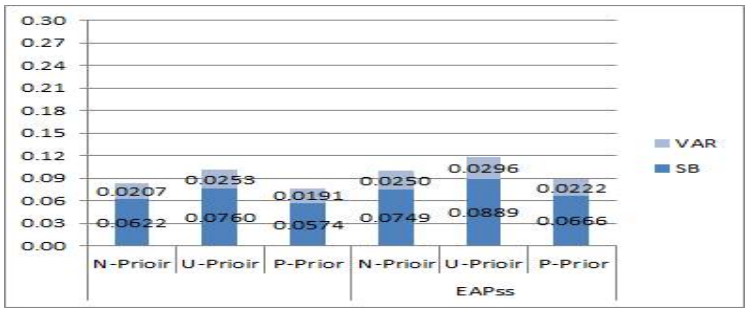


(d) N(0,1), 문항 범주:5, 검사 길이:40

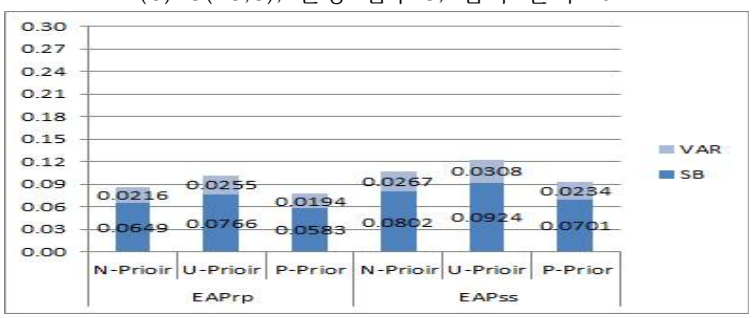
[그림 5] GPCM 모형에서 표준정규분포가 피험자 분포 일 때 각 조건에 따른 MSE



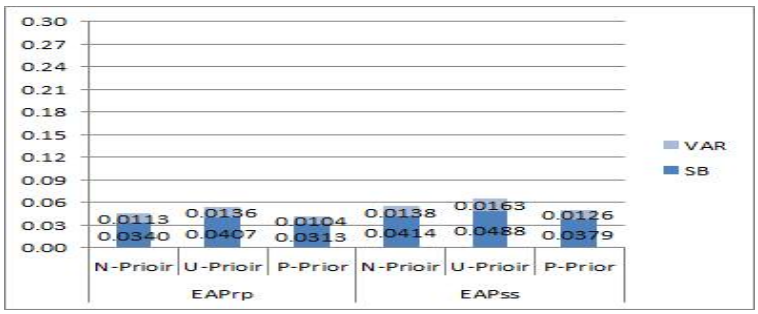
(a) U(-3,3), 문항 범주:3, 검사 길이:20



(b) U(-3,3), 문항 범주:3, 검사 길이:40

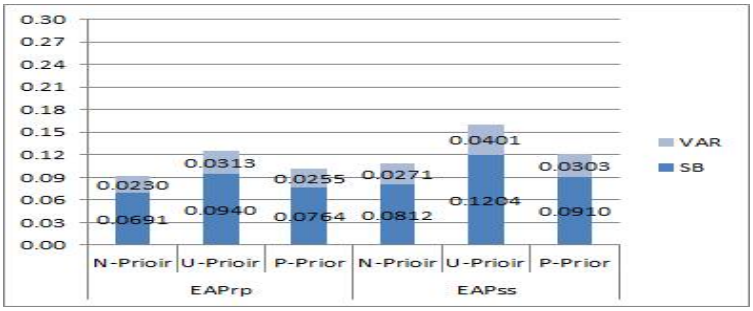


(c) U(-3,3), 문항 범주:5, 검사 길이:20

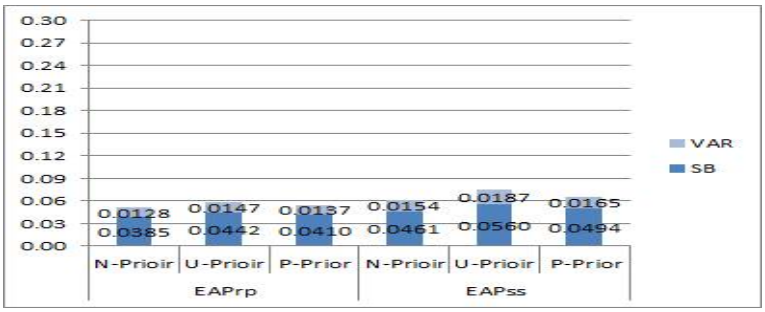


(d) U(-3,3), 문항 범주:5, 검사 길이:40

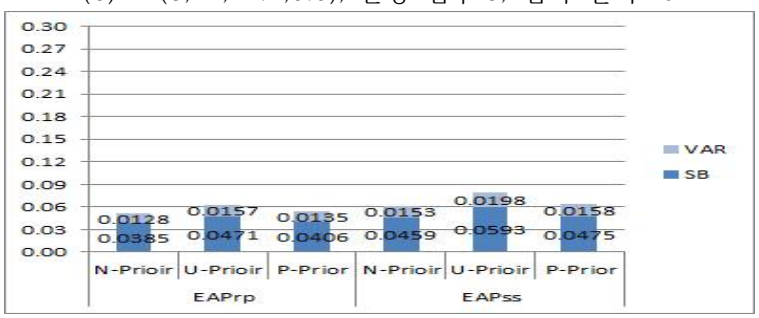
[그림 6] GPCM 모형에서 균일분포가 피험자 분포 일 때 각 조건에 따른 MSE



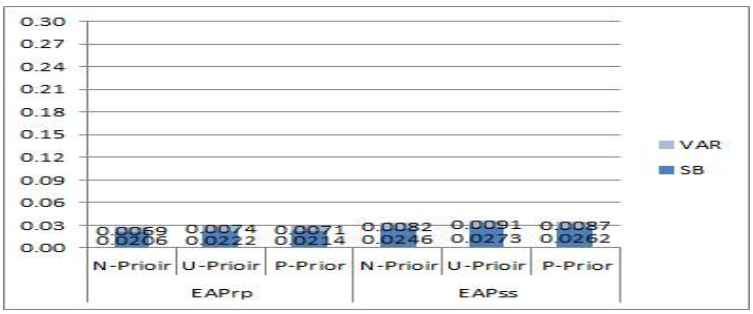
(a) B4(3,12,-2.1,5.9), 문항 범주:3, 검사 길이:20



(b) B4(3,12,-2.1,5.9), 문항 범주:3, 검사 길이:40



(c) B4(3,12,-2.1,5.9), 문항 범주:5, 검사 길이:20



(d) B4(3,12,-2.1,5.9), 문항 범주:5, 검사 길이:40

[그림 7] GPCM 모형에서 편포가 피험자 분포 일 때 각 조건에 따른 MSE

2. 능력 추정치들의 능력 수준별 분포에 대한 결과

1) 두 EAP 능력모수 추정방법의 상관계수

<표 5>에서 확인할 수 있듯이, 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력 추정 방법은 GRM과 GPCM 모형 하의 모든 조건에서 1의 상관계수를 산출하였다. 즉, 각 능력 수준별 분포를 사용하였을 때, 모형과 문항 범주, 피험자 분포, 검사길이에 대하여 EAPrp와 EAPss의 동일한 능력추정치를 가지는 경향을 보였다.

<표 5> EAPrp와 EAPss간의 상관계수

모형	문항 범주	피험자분포	검사길이			
			20		40	
			사전분포			
			N-Prior	U-Prior	N-Prior	U-Prior
GRM	3	능력 수준별 분포	1.000	1.000	1.000	1.000
	5	능력 수준별 분포	1.000	1.000	1.000	1.000
GPCM	3	능력 수준별 분포	1.000	1.000	1.000	1.000
	5	능력 수준별 분포	1.000	1.000	1.000	1.000
사전분포별 평균(표준편차)			1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
전체평균(표준편차)						1.000 (0.000)

GRM 모형에서 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력모수 추정치간의 상관계수는 모든 조건에서 .997~.998 사이의 값을 가지며 1과 매우 가까운 상관계수를 보였다. 또한 EAPrp의 적과 EAPrp의 적용에 따른 차이는 뚜렷한 반응을 보이지 않았으며 그 차이는 미미하였음을 아래의 <표 6>에서 볼 수 있다.

<표 6> GRM에서 EAPrp와 EAPss의 진능력모수와의 상관계수

능력 추정 조건	문항 범주	피험자분포	검사길이			
			20		40	
			사전분포			
			N-Prior	U-Prior	N-Prior	U-Prior
EAPrp	3	능력 수준별 분포	0.996	0.997	0.997	0.998
	5	능력 수준별 분포	0.998	0.998	0.998	0.998
EAPss	3	능력 수준별 분포	0.996	0.997	0.997	0.997
	5	능력 수준별 분포	0.997	0.997	0.998	0.998
사전분포별 평균(표준편차)			.997 (0.001)	.997 (0.000)	.997 (0.000)	.998 (0.000)
전체평균(표준편차)						.997 (0.001)

GPCM 모형에서 각기 다른 형태의 사전 분포의 사용에 따른 두 EAP 능력모수 추정치간의 상관계수는 <표 7>에 제시된 것처럼 모든 조건에서 .993~.998 사이의 값을 가지며 1과 매우 가까운 상관계수를 보였다. 또한 EAPrp의 적과 EAPrp의 적용에 따른 차이는 뚜렷한 반응을 보이지 않았으며 그 차이는 미미하였다..

<표 7> GPCM에서 EAPrp와 EAPss의 진능력모수와의 상관계수

능력 추정 조건	문항 범주	피험자분포	검사길이			
			20		40	
			사전분포			
			N-Prior	U-Prior	N-Prior	U-Prior
EAPrp	3	능력 수준별 분포	0.994	0.996	0.996	0.997
	5	능력 수준별 분포	0.996	0.997	0.997	0.998
EAPss	3	능력 수준별 분포	0.993	0.995	0.996	0.997
	5	능력 수준별 분포	0.996	0.997	0.997	0.998
사전분포별 평균(표준편차)			.995 (0.001)	.996 (0.001)	.996 (0.001)	.997 (0.000)
전체평균(표준편차)						.996 (0.001)

2) 조건부 편향

-4에서 4의 범위를 갖는 능력척도 Θ를 17개의 구간으로 나누어 각 능력 조건 구간의 편향의 변화 양상을 [그림 8]과 [그림 9]를 통해 살펴보았으며 그 특징은 다음과 같다.

첫째, 여러 조건들 속에서 EAPrp가 EAPss에 비하여 조금 더 적은 SB 값을 산출하였다. 이는 주변 부포의 RMSE 값에서 살펴보았던 것처럼 다분 문항반응모형 조건하에서 능력모수를 추정할 때, EAPss를 사용하면 피험자의 원점수와 일대일로 대응시켜서 능력모수를 추정하는 반면,

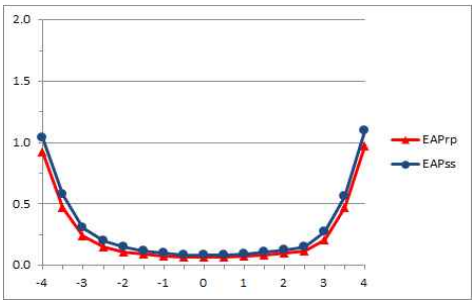
EAPrp는 피험자의 반응유형이라는 더 많은 정보를 사용하여 능력모수를 추정하기 때문에 생기는 당연한 결과이다. 그러나 표준정규 사전분포의 경우 [-2, 2]의 능력범위에서, 그리고 균일 사전분포를 사용한 경우 [-1, 1]의 능력범위에서는 EAPrp와 EAPss 방법에 의한 능력 추정치 모두 SB가 거의 0에 가까운 값을 산출하였다.

둘째, 어떤 사전정보를 사용하느냐에 따라서 조건부 편향은 독특한 변화 양상을 보였다. 표준정규분포 $N(0,1)$을 사전정보로 사용한 경우에는 모든 조건에 대하여 U 자 편향 양상을 보였다. 즉, GRM과 GPCM 하에서 EARrp와 EAPss에 의한 추정치들은 모두 [-2, 2]의 능력 구간에서는 거의 0에 가까운 SB값을 가졌으나 이 구간에서 멀어질수록 큰 SB값을 보였다. 이와는 대조적으로 균일분포 $U(-3,3)$를 사전정보로 사용한 경우는 모두 모형에서 두 EAP 능력 추정치의 편향이 W 모양을 보임을 확인할 수 있었다. 조건마다 정도의 차이는 있지만 대게 [-1, 1]의 능력 구간에서는 거의 0에 가까운 SB값을 가지며, 이 구간에서 멀어지면서 SB값이 상승하였다가 -3과 3의 능력구간에서 다시 SB값이 하락하며 이 구간에서 멀어질수록 SB값이 다시 상승하는 경향을 보였다. 따라서 표준정규분포 사전분포는 평균 근처에 위치한 중간능력의 피험자들의 능력모수를 추정하는데 상대적으로 우수한 기능을 하는 반면, 균일 사전분포는 능력모수 척도의 양 극단에 위치한 능력의 피험자들의 능력모수를 추정하는데 상대적으로 우수한 기능을 함을 시사한다.

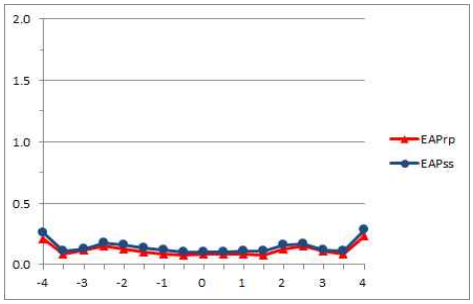
셋째, 문항 범주와 검사 길이가 증가하면 EAPrp와 EAPss를 사용하여 능력모수를 추정하였을 때, 각 능력 구간의 SB값은 줄어든다. 표준정규 사전정보를 사용한 경우, 모든 조건에 대하여 3개의 문항 범주와 20개의 검사 길이를 가진 조건 (a)와 (b)에 비하여 5개의 문항 범주와 40개의 거사 길이를 가진 조건 (g)와(h)의 SB가 눈에 띄게 감소하였음을 확인할

수 있다. 특히 동일한 검사 길이를 가졌지만 문항 범주만 증가시킨 경우와(즉, 조건(a)에서 조건(e)로, 그리고 조건(b)에서 조건(f)로 바뀐 경우이다.), 동일한 문항 범주를 가지고 검사 길이를 증가시킨 경우(즉, 조건(a)에서 조건(c)로, 그리고 조건(b)에서 조건(d)로 바뀐 경우이다.) 양 극단의 SB값이 감소하며 평평한 향상을 보이는 경향이 있었다.

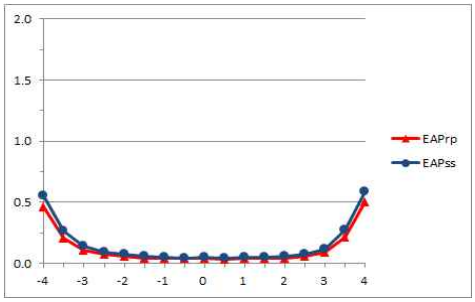
마지막으로, GRM과 GPCM 적용에 따른 두 EAP 능력추정치의 각 능력 구간에 대한 SB 값에 차이가 존재한다. 전반적으로 GPCM 하에서의 SB값이 GRM 하에서의 SB값보다 더 큰 값을 가졌다. 그러나 GRM하에서는 GRM의 조건 문항 범주와 검사 길이가 증가하는 조건 (e), (f), (g), (h)에서 EAPrp와 EAPss의 양 극단의 SB값의 차가 벌어지는 데 비하여 GPCM하에서는 모든 조건에 대하여 두 EAP 추정치의 SB값은 양 극단의 능력조건에서도 큰 간격이 벌어지지 않고 유사한 SB값을 보였다.



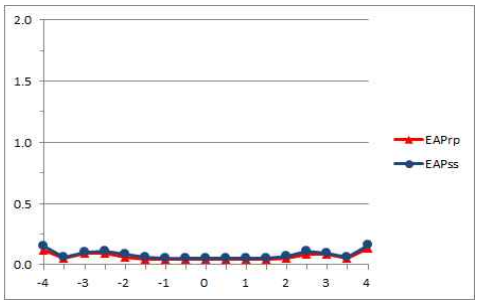
(a) N(0,1), 문항 범주:3, 검사 길이:20



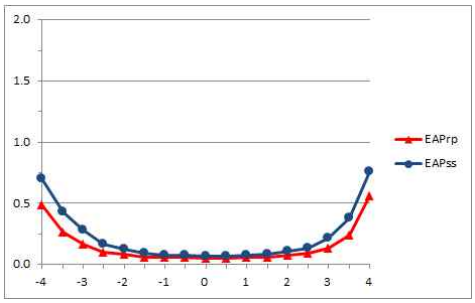
(b) U(-3,3), 문항범주:3, 검사 길이:20



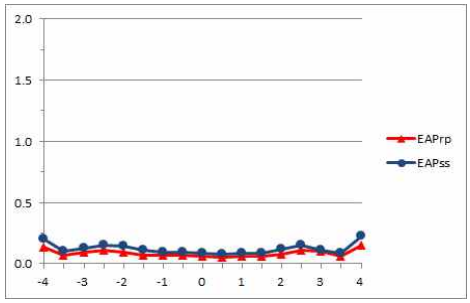
(c) N(0,1), 문항 범주:3, 검사 길이:40



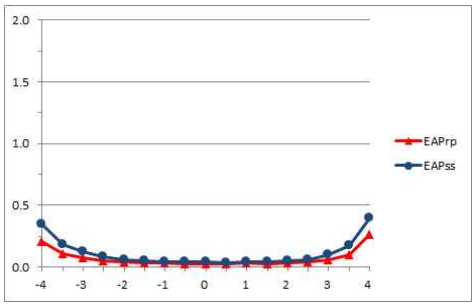
(b) U(-3,3), 문항범주:3, 검사 길이:40



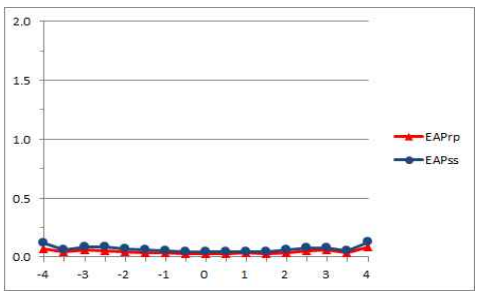
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20

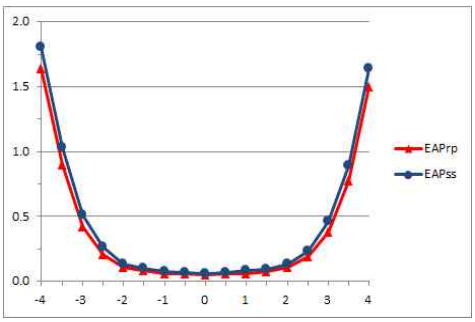


(g) N(0,1), 문항 범주:5, 검사 길이:40

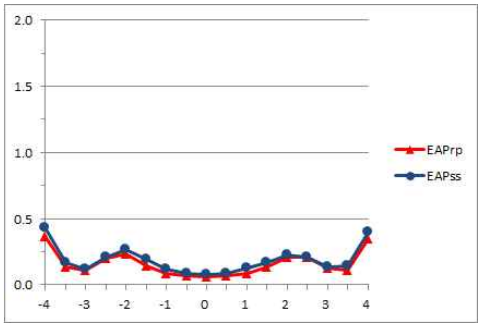


(h) U(-3,3), 문항범주:5, 검사 길이:50

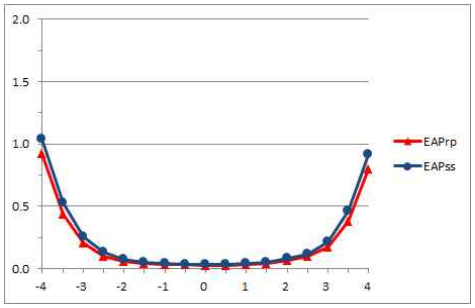
[그림 8] GRM 에서 두 피험자 분포 사용 시 각 조건에 따른 SB



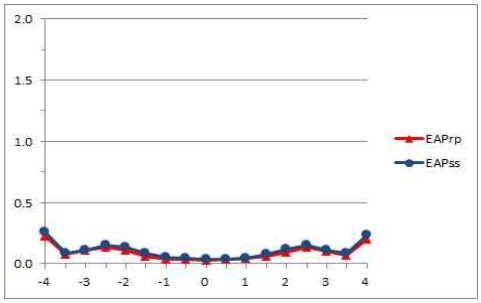
(a) N(0,1), 문항 범주:3, 검사 길이:20



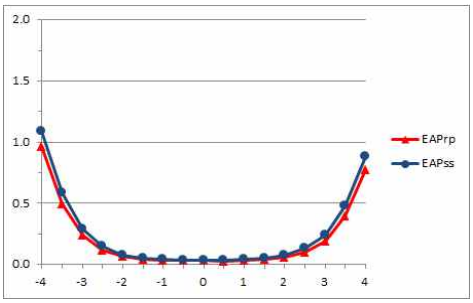
(b) U(-3,3), 문항범주:3, 검사 길이:20



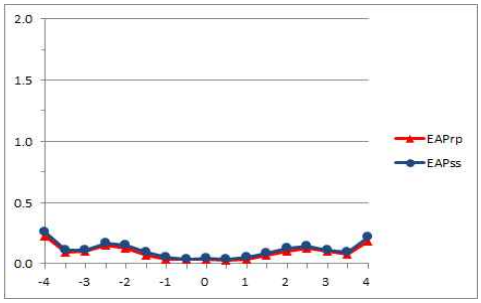
(c) N(0,1), 문항 범주:3, 검사 길이:40



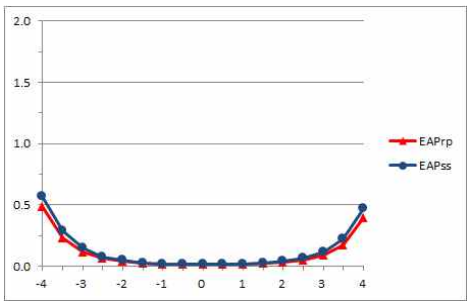
(b) U(-3,3), 문항범주:3, 검사 길이:40



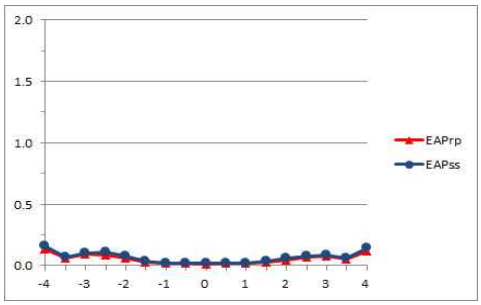
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20



(g) N(0,1), 문항 범주:5, 검사 길이:40

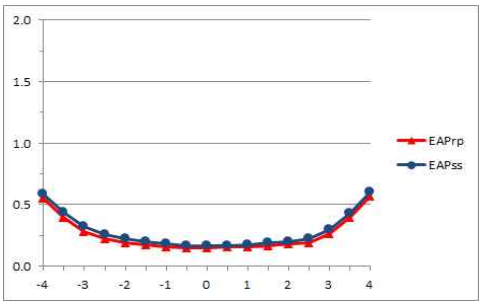


(h) U(-3,3), 문항범주:5, 검사 길이:50

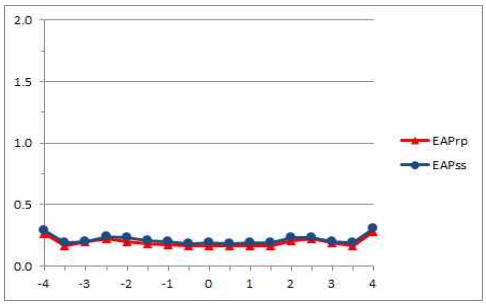
[그림 9] GPCM 에서 두 피험자 분포 사용 시 각 조건에 따른 SB

3) 조건부 표준오차

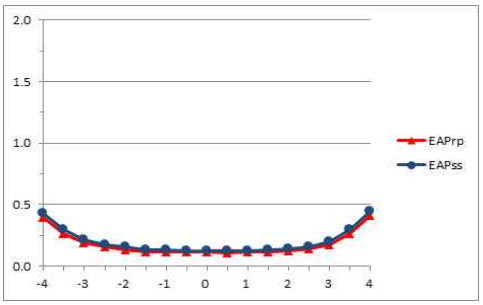
각 능력 조건 구간의 표준오차의 변화 양상을 [그림 10]과 [그림 11]을 통해 살펴보았다. 이들은 EAPrp와 EAPss의 차이, 사전정보의 사용에 따른 차이, 문항범주와 검사길에 따른 차이, 그리고 모형에 따른 차이 등에서 조건부 편향에서 살펴본 것과 매우 유사한 특징을 갖는다. 다만 조건부 표준오차는 조건부 편향과는 달리 체계적인 오차 범위를 갖는다는 것이다. 예를 들어 [그림 9]의 조건 (a)는 1.5~2.0 사이에 두 EAP 오차 값을 갖지만 조건 (g)에서는 약 0.5로 오차 값이 줄었다. 이에 반해 [그림 11]의 조건 (a)는 약 0.7~0.8의 오차 범위에 두 EAP 추정치의 SE값을 갖으며 조건 (g)에서도 약 0.4~0.5의 오차 범위를 갖으며 비교적 체계적인 모습을 보였다.



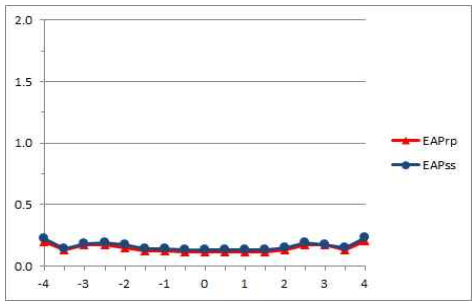
(a) N(0,1), 문항 범주:3, 검사 길이:20



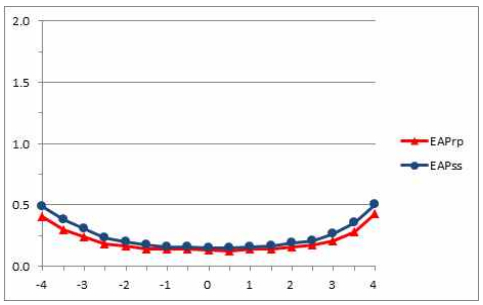
(b) U(-3,3), 문항범주:3, 검사 길이:20



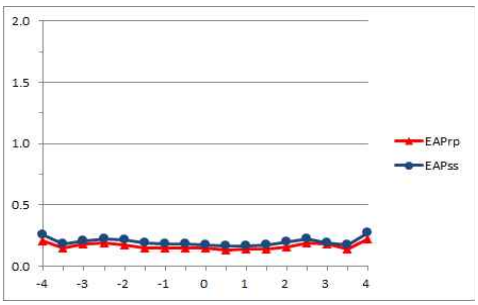
(c) N(0,1), 문항 범주:3, 검사 길이:40



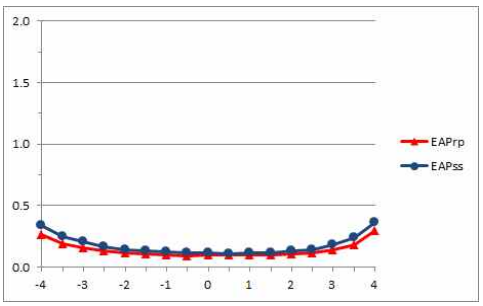
(b) U(-3,3), 문항범주:3, 검사 길이:40



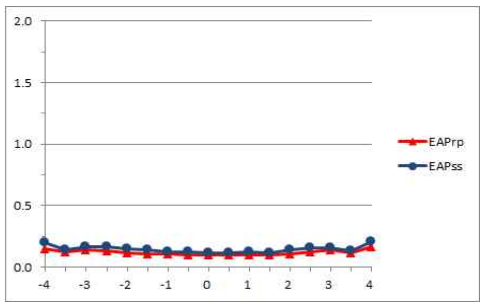
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20

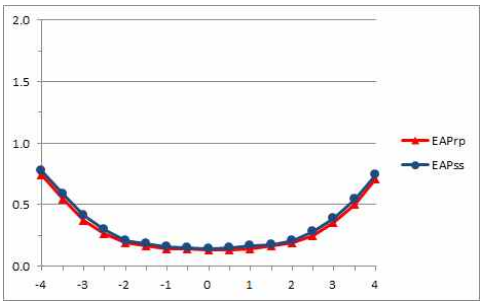


(g) N(0,1), 문항 범주:5, 검사 길이:40

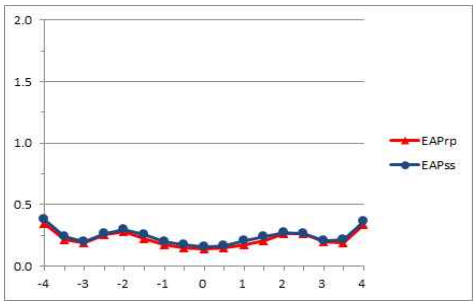


(h) U(-3,3), 문항범주:5, 검사 길이:50

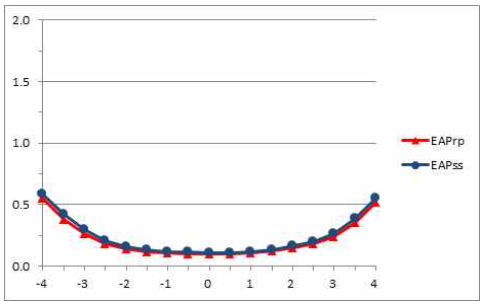
[그림 10] GRM 에서 두 피험자 분포 사용 시 각 조건에 따른 SE



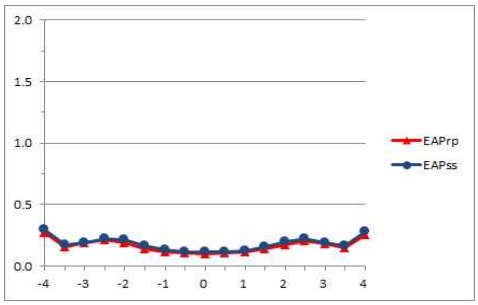
(a) N(0,1), 문항 범주:3, 검사 길이:20



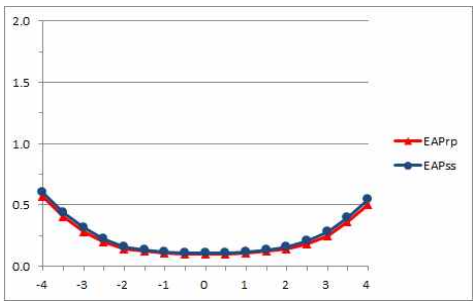
(b) U(-3,3), 문항범주:3, 검사 길이:20



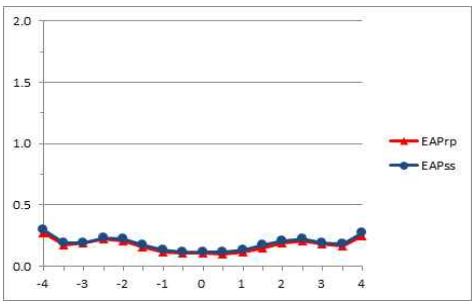
(c) N(0,1), 문항 범주:3, 검사 길이:40



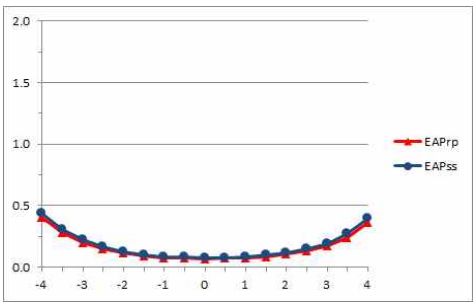
(b) U(-3,3), 문항범주:3, 검사 길이:40



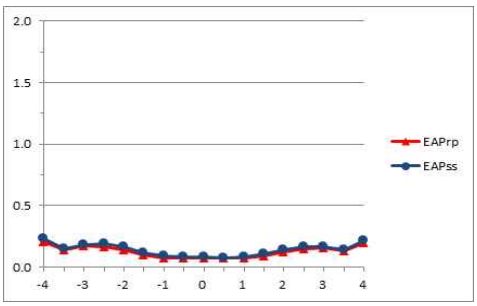
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20



(g) N(0,1), 문항 범주:5, 검사 길이:40



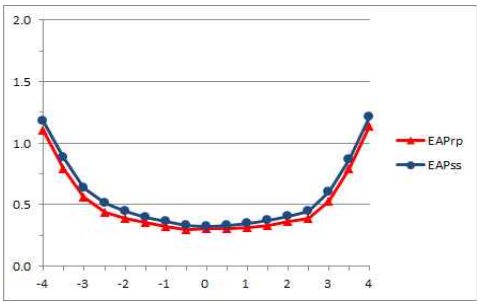
(h) U(-3,3), 문항범주:5, 검사 길이:50

[그림 11] GPCM에서 두 피험자 분포 사용 시 각 조건에 따른 SE

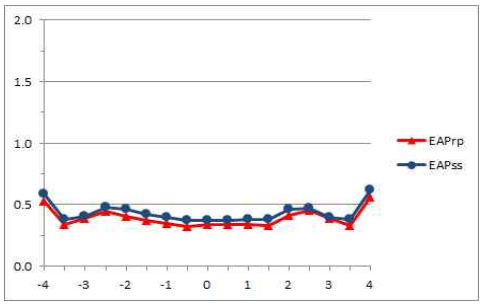
4) 조건부 RMSE

RMSE는 BS와 SE 제곱의 합으로 산출되는 통계량이기 때문에(즉, $MSE = BS + SE^2$), 각 능력 구간에 대한 조건부 RMSE의 상대적 크기와 변화 양상은 BS와 SE의 능력 수준별 평가 준거의 결과와 비슷한 양상을 보였다.

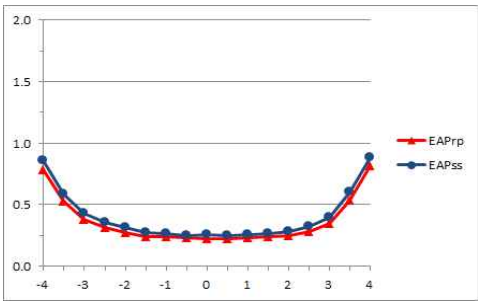
여러 조건들 속에서 EAPrp가 EAPss에 비하여 조금 더 적은 SB값을 산출하였으며, 어떤 사전정보를 사용하느냐에 따라서 조건부 편향은 독특한 변화 양상을 보였다. 즉, 표준정규분포 $N(0,1)$을 사전정보로 사용한 경우에는 모든 조건에 대하여 U 자 편향 양상을 보였으며, 균일분포 $U(-3,3)$를 사전정보로 사용한 경우는 모두 모형에서 두 EAP 능력 추정치의 편향이 W 모양을 보임을 확인할 수 있었다. 또한 문항 범주와 검사 길이가 증가하면 EAPrp와 EAPss를 사용하여 능력모수를 추정하였을 때, 특히 양 극단의 능력 구간에서 더 적은 RMSE 값을 산출하였다. 마지막으로, 전반적으로 GPCM 하에서의 RMSE값이 GRM 하에서의 RMSE값보다 더 큰 값을 가지는 경향이 있으나 GRM하에서는 GRM의 조건 문항 범주와 검사 길이가 증가하는 조건에서 EAPrp와 EAPss의 양 극단의 RMSE값의 차가 벌어지는 데 비하여 GPCM하에서는 모든 조건에 대하여 두 EAP 추정치의 RMSE값은 양 극단의 능력조건에서도 큰 간격이 벌어지지 않고 유사한 RMSE값을 보였다.



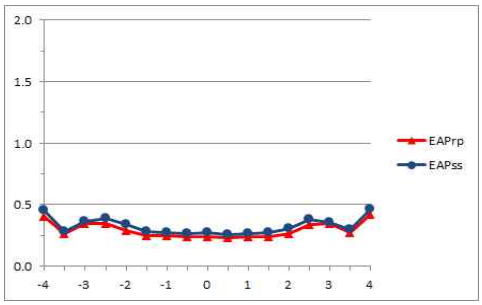
(a) N(0,1), 문항 범주:3, 검사 길이:20



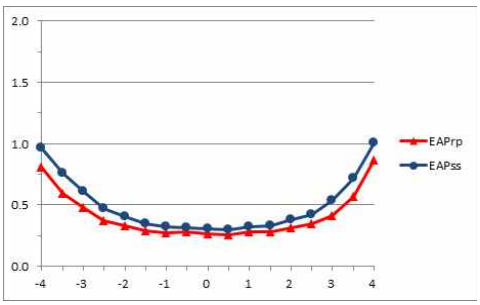
(b) U(-3,3), 문항범주:3, 검사 길이:20



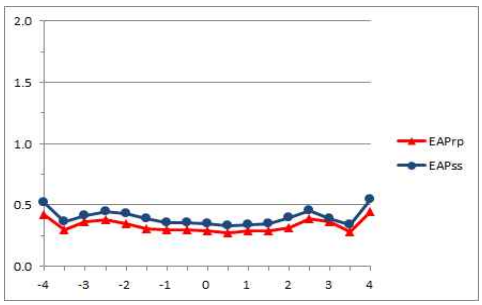
(c) N(0,1), 문항 범주:3, 검사 길이:40



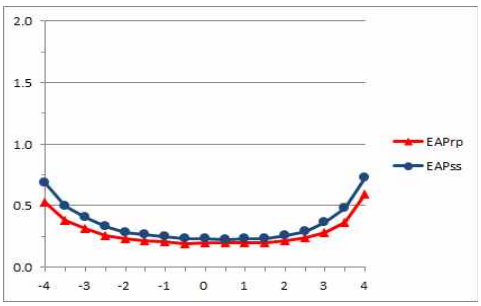
(b) U(-3,3), 문항범주:3, 검사 길이:40



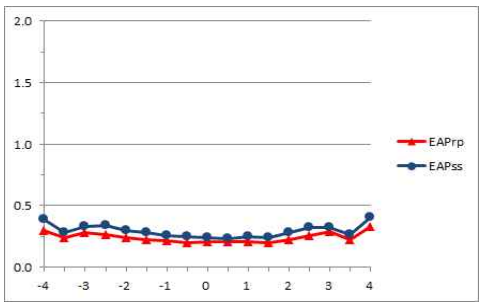
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20

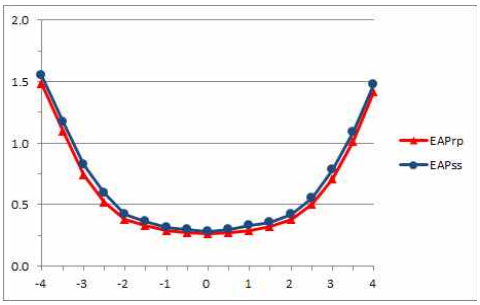


(g) N(0,1), 문항 범주:5, 검사 길이:40

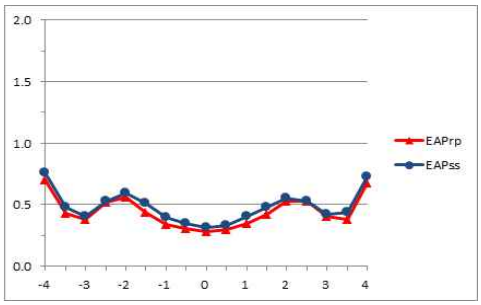


(f) U(-3,3), 문항범주:5, 검사 길이:50

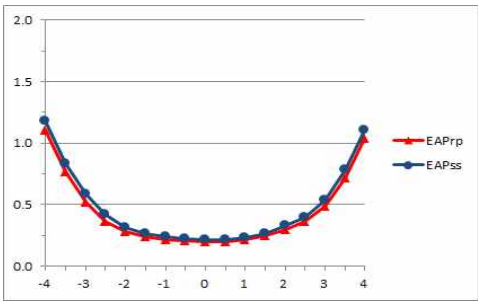
[그림 12] GRM 에서 두 피험자 분포 사용 시 각 조건에 따른 RMSE



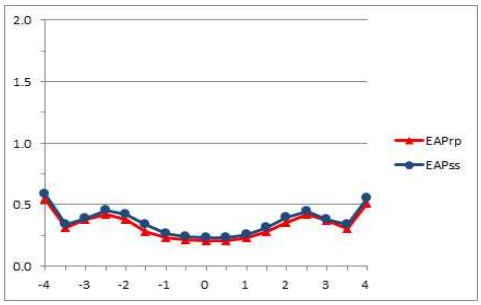
(a) N(0,1), 문항 범주:3, 검사 길이:20



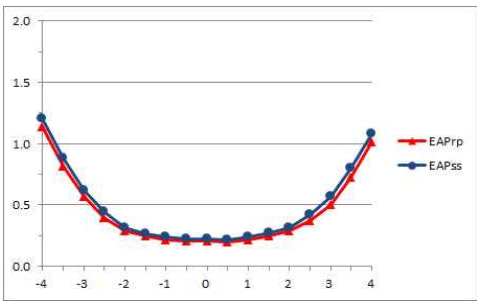
(b) U(-3,3), 문항범주:3, 검사 길이:20



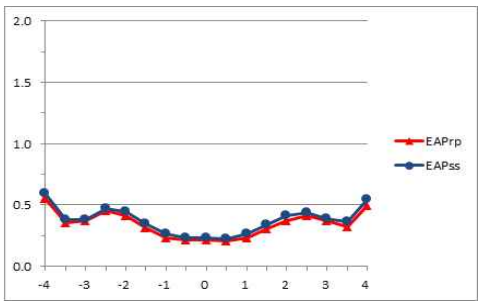
(c) N(0,1), 문항 범주:3, 검사 길이:40



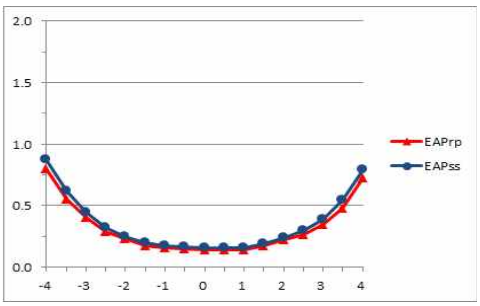
(b) U(-3,3), 문항범주:3, 검사 길이:40



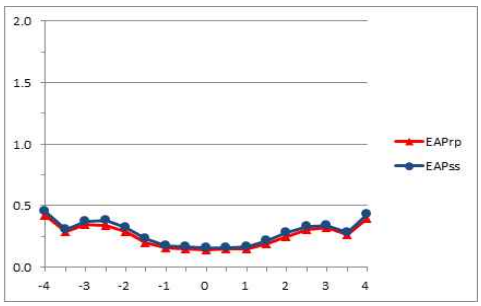
(e) N(0,1), 문항 범주:5, 검사 길이:20



(f) U(-3,3), 문항범주:5, 검사 길이:20



(g) N(0,1), 문항 범주:5, 검사 길이:40



(f) U(-3,3), 문항범주:5, 검사 길이:50

[그림 13] GPCM 에서 두 피험자 분포 사용 시 각 조건에 따른 RMSE

3. EAPrp와 EAPss의 MSE 차이의 검정

두 EAP 능력추정 방법의 차이의 유의미성을 통계적으로 검정하기 위해 검사길이, 문항범주와 사전분포의 조건별 RMSE를 종속변수로 하여 두 EAP의 RMSE를 대상으로 각 모형별로 두 대응표본 t 검정을 실시하였으며 그 결과는 아래의 <표 8>과 <표 9>에 제시된 바처럼 유의수준 0.5에서 통계적으로 유의한 차이가 있었다.

<표 8> GRM에서 두 대응표본 t 검정 결과

조건		추정방법	N	평균	표준편차	P-value
검사길이	20	EAPrp	68	0.412	0.182	.000
		EAPss	68	0.475	0.199	
	40	EAPrp	68	0.295	0.124	.000
		EAPss	68	0.343	0.140	
문항범주	3	EAPrp	68	0.390	0.189	.000
		EAPss	68	0.432	0.202	
	5	EAPrp	68	0.317	0.131	.000
		EAPss	68	0.386	0.162	
사전분포	N(0,1)	EAPrp	68	0.393	0.215	.000
		EAPss	68	0.456	0.238	
	U(-3,3)	EAPrp	68	0.314	0.079	.000
		EAPss	68	0.362	0.084	

<표 9> GPCM에서 두 대응표본 t 검정 결과

조건		추정방법	N	평균	표준편차	P-value
검사길이	20	EAPrp	68	0.460	0.277	.000
		EAPss	68	0.500	0.291	
	40	EAPrp	68	0.337	0.201	.000
		EAPss	68	0.369	0.216	
문항범주	3	EAPrp	68	0.455	0.277	.000
		EAPss	68	0.495	0.291	
	5	EAPrp	68	0.342	0.204	.000
		EAPss	68	0.374	0.219	
사전분포	N(0,1)	EAPrp	68	0.453	0.321	.000
		EAPss	68	0.494	0.341	
	U(-3,3)	EAPrp	68	0.344	0.124	.000
		EAPss	68	0.375	0.130	

위의 두 표에서처럼 EAPrp의 RMSE가 EAPss의 RMSE보다 작은 것은 EAPrp가 더 많은 정보를 이용하여 피험자의 능력을 추정한다는 면에서 당연한 결과이다. 그러나 다음의 <표 10>과 <표 11>에서 각 모형하에서 EAPss와 EAPrp의 MSE 값의 차가 능력 추정치들의 주변 분포에서는 평균 0.078이며, 능력 수준별 부포에서는 평균 0.043임을 확인할 수 있다. 이는 이분 문항반응이론에서 두 EAP 능력추정 방식의 RMSE를 연구한 강태훈(2014)의 연구에서 최대 MSE 차이가 나타나는 조건에서도 그 차이가 0.1에 불과하였음과 Lee & Ban(201)의 연구에서 EAP 추정치와 AMP 추정치의 MSE값의 차이가 평균적으로 약 0.3 정도였던 것에 비하여 매우 작은 차이임을 알 수 있다.

<표 10> 능력 추정치들의 주변 분포에서 EAPrss와 EAPrp의 MSE 값의 차

모형	문항 범주	피험자분포	검사길이					
			20			40		
			사전분포					
			N-Prior	U-Prior	P-Prior	N-Prior	U-Prior	P-Prior
GRM	3	$N(0,1)$	0.023	0.032	0.024	0.013	0.015	0.013
		$U(3,-3)$	0.308	0.326	0.280	0.155	0.173	0.143
		$B4(3.12,-2.1,5.9)$	0.022	0.033	0.026	0.013	0.016	0.015
	5	$N(0,1)$	0.031	0.042	0.031	0.017	0.020	0.017
		$U(3,-3)$	0.246	0.255	0.216	0.025	0.133	0.119
		$B4(3.12,-2.1,5.9)$	0.030	0.044	0.032	0.016	0.019	0.019
GPCM	3	$N(0,1)$	0.017	0.031	0.019	0.011	0.014	0.011
		$U(3,-3)$	0.356	0.389	0.328	0.017	0.220	0.165
		$B4(3.12,-2.1,5.9)$	0.016	0.035	0.019	0.010	0.016	0.011
	5	$N(0,1)$	0.010	0.015	0.011	0.006	0.008	0.006
		$U(3,-3)$	0.193	0.225	0.171	0.100	0.119	0.092
		$B4(3.12,-2.1,5.9)$	0.010	0.016	0.009	0.005	0.007	0.006
사전분포별 평균(표준편차)			0.105 (0.135)	0.120 (0.141)	0.097 (0.121)	0.032 (0.048)	0.063 (0.078)	0.051 (0.062)
전체평균(표준편차)								0.078 (0.103)

<표 11> 능력 추정치들의 능력 수준별 분포에서 EAPrss와 EAPrp의 MSE 값의 차

모형	문항 범주	피험자분포	검사길이			
			20		40	
			사전분포			
			N-Prior	U-Prior	N-Prior	U-Prior
GRM	3	능력 수준별 분포	0.066	0.034	0.037	0.020
	5	능력 수준별 분포	0.098	0.049	0.053	0.029
GPCM	3	능력 수준별 분포	0.076	0.037	0.049	0.022
	5	능력 수준별 분포	0.049	0.021	0.030	0.014
사전분포별 평균(표준편차)			0.072 (0.020)	0.035 (0.012)	0.042 (0.011)	0.021 (0.006)
전체평균(표준편차)						0.043 (0.023)

Ⅴ. 논의 및 결론

본 연구는 이제껏 주로 이분문항반응모형 맥락에서 연구되어 온 EAPrp와 EAPss 간의 비교를 다분 문항반응모형으로 확장하기 위하여 다분 문항반응모형하에서 EAPss의 수행력을 EAPrp의 수행력과 비교하여 EAPss를 선택하였을 때의 능력모수의 추정 정확도가 구체적으로 어느 정도인지를 체계적으로 정리함으로써, 검사 자료 분석을 실시하는 연구자가 이를 감수하고 EAPss를 사용할지 또는 EAPrp를 사용할 지에 대한 선택을 돕기 위해 수행되었다.

이를 위하여 본 연구에서는 주어진 모의실험 조건하에서 두 사후기대추정법으로 피험자 능력 모수를 추정하였을 때, 피험자의 능력에 대한 두 사후기대추정법의 진모수 복원력은 검사에 응시한 피험자 전체를 함께 고려하는 맥락에서 두 추정치 상호 간 상관계수, 진값과 추정치 간의 상관계수 및 평균제곱오차 등의 차이와 여러 수준의 피험자 능력모수 구간 각각에서 두 추정치 상호 간 상관계수, 진값과 추정치 간의 상관계수 및 조건부 편향, 조건부 표준오차, 평균제곱오차 등의 차이를 살펴보았다.

피험자의 능력을 추정하기 위하여 문항별 반응양식을 고려하는 경우와(EAPrp)와 이를 무시하고 검사 총점만을 고려한 경우(EAPss), 전자는 피험자에 대하여 더 많은 정보를 활용할 수 있다는 점에서 논리적으로나 수리적으로 후자에 비하여 더 정확한 능력을 추정 할 것이다. 그러나 후자의 선택 결과가 일반 대중이 쉽게 이해할 수 있고 거부감이 덜할 수 있다는 점을 고려하면, EAPss의 복원력이 EAPrp의 복원력과 큰 차이가 나지 않는 한 EAPss 또한 활용 가능한 선택이 될 수 있다. 따라서 본 연구에서는 두 EAP 능력추정 방법을 다양한 조건하에서 모의실험을 통하여 비교함으로써 두 방법 간의 진능력 모수 복원의 정확도를 비교 및

평가해 보았다. 연구는 주변 분포에 관한 연구와 능력 수준별 분포 연구로 두 가지 측면에서 진행되었다.

결과적으로, EAPrp는 여러 조건들 속에서 EAPss에 비하여 조금 더 적은 MSE, SB, VAR 값을 산출하였다. 이는 다분 문항반응모형 조건하에서 능력모수를 추정할 때, EAPrp는 피험자의 반응유형을 사용하여 능력모수를 추정하기 때문에 피험자의 원점수와 일대일로 대응하여 능력모수를 추정하는 EAPss에 비하여 더 많은 정보를 사용한다는 점에서 생기는 당연한 결과이다. 그러나 앞서 제시된 표에서 각 모형하에서 EAPss와 EAPrp의 MSE 값의 차가 능력 추정치들의 주변 분포에서는 평균 0.078이며, 능력 수준별 분포에서는 평균 0.043임을 확인할 수 있었다. 이는 이분 문항반응이론에서 두 EAP 능력추정 방식의 RMSE를 연구한 강태훈(2014)의 연구에서 최대 MSE 차이가 나타나는 조건에서도 그 차이가 0.1에 불과하였음과 Lee & Ban(201)의 연구에서 EAP 추정치와 AMP 추정치의 MSE값의 차이가 평균적으로 약 0.3 정도였던 것에 비하여 매우 작은 차이임을 알 수 있다. 따라서 앞서 제시된 상관계수의 결과가 모든 조건에서 .95이상으로 매우 높았던 점과, 두 EAP 방법의 RMSE의 차이 정도가 실제적으로는 소수점 자리에서만 발생하였다는 점을 감안한다면 대중의 이해를 도울 수 있다는 점과 대규모 검사를 시행할 때 피험자들의 원점수를 계산하여 기록하기 편리하다는 점 또한 점수를 보고할 때 백분율을 계산하기 쉽다는 점에서 검사 제작자의 판단 하에 두 EAP의 차이를 감안할 수 있을 것이다. 따라서 대중에 대한 설득이 중요한 여건하에서 EAPrp와 EAPss를 사용하여 피험자의 능력모수를 추정할 때는 EAPss의 사용이 충분히 고려될 수 있을 것이다.

두 EAP 능력모수 추정방법의 진능력모수 복원력의 정확성 비교 외에도 본 연구를 통해 얻을 수 있는 추가적인 결과는 다음과 같으며, 이 결

과들은 두 EAP 능력모수 추정방법에 대하여는 비슷한 결과를 산출하였다.

첫째, 베이지안 능력모수 추정을 위하여 사용할 사전정보를 선정할 때, 어떤 EAP 방법을 사용하여 능력모수를 추정하든 지에 관계없이 능력추정의 정확성을 전체 피험자 수준에서 고려할 때, 사전분포로서 와 표준정규분포는 능력수준에 따라 서로 다른 상대적 우수성을 보였다. 다양한 조건에서 두 EAP 능력추정 방법은 모두 균일분포를 사전정보로 사용하였을 때는 (-4, 2]와 [2, 4)의 극단에 위치한 능력모수를 보다 잘 추정하였다. 이와 반대로 사정정보로 표준정규분포가 사용되었을 때는 능력 수준 [-2, 2]의 지점에서 보다 정확한 두 EAP 능력모수 추정이 가증하였다. 강태훈(2014)와 김성훈, 박인심(2010)의 연구에서도 연구내에서 상정한 능력추정 방법에 관계없이 균일분포는 극단에 위치한 능력 수준을, 그리고 표준정규분포는 중간에 위치한 능력수준에 보다 작은 MSE값을 산출하였다. 따라서, 능력 수준이 평균 근처에 위치한 피험자를 평가하는 맥락에서는 표준정규분포를 사전분포로 사용하는 것이 적합하며, 능력 수준이 극단적인 피험자를 평가하는 맥락에서는 균일분포를 사전분포로 사용하는 것이 적합함을 시사한다. 더불어, 표준정규분포를 사전분포로 하여 그 결과를 이용하여 다시 피험자의 능력을 재산출하는 것이 가능한 상황에서는, 산출된 피험자의 사후분포를 사전분포로 재사용하여 능력모수를 추정하는 것도 균일분포를 사전정보로 사용하는 것보다 안전한 선택일 수 있다. 따라서 적어도 GRM과 GPCM을 적용하여 베이지안 능력모수를 추정할 때는 균일 사전분포를 사용하는 것 보다는 표준정규분포나 피험자분포를 사전분포로 재사용하는 것이 정확한 진능력모수 복원을 가능하게 해 주는 더 안전한 선택일 수 있음을 함의한다.

둘째, 베이지안 능력모수 추정을 위하여 피험자분포의 상정에 따른 능

력모수 추정의 정확성에 차이가 있었다. 두 EAP 방법과 세 가지 사전정보 조건을 상정하여 능력모수를 추정하였을 때, 전반적으로 표준정규분포를 피험자 분포로 상정한 경우에 가장 작은 MSE 값을 가졌으며, 두 번째로 큰 MSE 값은 편포를 피험자 분포로 상정한 경우였다. 가장 큰 MSE값을 갖게 하는 피험자 분포는 균일분포였다. 김성훈, 박인심(2010)의 연구에서도 정보적 사전분포(표준정규분포)에 기초한 EAP방법(이 연구에서는 EAPrp를 가리킴.)이 무정보적 사준분포(균일분포)에 기초한 방법보다 낮은 표준오차를 산출함을 확인할 수 있었다.따라서 GRM과 GPCM을 사용하여 베이지안 능력모수를 추정할 때는 편포나 균일분포로 피험자 분포를 상정하기 보다는 표준정규분포를 사용하여 피험자 분포를 상정하는 것이 더 안전한 선택이 될 수 있다.

셋째, 문항 범주와 검사 길이가 증가하면 EAPrp와 EAPss를 사용하여 능력모수를 추정하였을 때, MSE값은 줄어든다. 박정(1999b)의 연구에서 이러한 결과에 대한 이유를 유추할 수 있다. 이는 기울기 모수치는 각 문항당 공통된 모수치로서 문항에 포함된 모든 범주에 동일한 모수치이기 때문에, 범주의 개수가 증가할수록 기울기 모수치를 추정할 때 좀 더 많은 정보를 주게되어 좀 더 정확한 모수치를 추정하게 되기 때문이다.

마지막으로, GRM과 GPCM 적용에 따른 두 EAP 능력추정치의 진능력모수 정확성에는 차이가 존재한다. GPCM 하에서 표준정규분포를 피험자 분포로 사용한 조건과 부적편포를 피험자 분포로 사용한 조건을 살펴보면, 동일한 피험자 분포를 사용하였지만 GRM을 사용하여 능력모수를 추정한 조건과 비교하였을 때, 이들은 GPCM 하에서 더 적은 MSE값을 가지며 특히 문항 범주 수가 5인 경우 상대적으로 눈에 띄게 MSE가 줄었음을 확인할 수 있다. GPCM이 GRM보다 상대적으로 우월한 모형이라는 결과는 강태훈, 김명연(2012)의 모의실험 연구에서도 확인할 수 있었다.

따라서 다분 문항반응모형 하에서 베이지안 능력 추정을 위한 모형을 선택할 수 있다면 GRM 보다는 GPCM이 상대적으로 더 안정적인 선택이라 할 수 있다.

본 연구는 다음과 같은 제한점을 가지고 있으며 이에 대한 추가적 분석 및 후속연구가 필요함을 밝히는 바이다.

첫째, 본 연구는 제한적인 모의실험 조건으로 구성하였다. 기존의 연구는 피험자의 수가 증가할수록 모수치는 정확하게 추정되었음을 밝힌다(박정, 1999; De Ayala, 1995). 더불어 문항의 속성(예: 문항 곤란도 모수의 분포, 문항 변별도 모수의 분포, 문항 추측도의 전반적 크기 등)이 모의실험 조건에 포함되었다면, EAP 능력 추정 방법의 정확도에 영향을 미칠 수 있는 검사 특성들을 조금 더 다양한 조건하에서 이해할 수 있었을 것이다.

둘째, 본 연구는 피험자의 능력모수를 모의실험 자료를 통하여 추정하는 대신 진문항모수를 그대로 사용하였다. 이러한 방법은 두 EAP 능력모수 추정치 복원력의 정확성을 비교하는데 큰 지장이 되지는 않지만, 온전한 현실적 접근 방법이라고는 볼 수 있다. 본 연구에서는 동일한 표본크기를 사용하였지만, 두 EAP 능력모수 추정방법의 수행을 비교하는데 표본 크기를 달리한 강태훈(2014)의 연구는 이러한 접근방법으로 인하여, 표본 크기에 관계없이 두 EAP 능력추정 방법이 매우 비슷한 정도의 진 능력모수 복원 정확도를 보이는 결과를 얻게 되었음을 밝힌다.

셋째, 본 연구는 EAPrp와 EAPss를 추정하기 위하여 MATLAB을 이용하여 개발한 프로그램(강태훈, 2014)을 사용하였지만, 현재 EAPss의 결과를 얻기 위한 IRT 프로그램으로 IRTPRO(Cai, du Toit, & Thissen, 2011)가 있다. IRRPRO는 최근에 개발된 프로그램으로 이를 이용하여 EAPss 능력 추정치를 구한 결과에 대한 복원력 정확도 검증의 연구가 필요하다.

참 고 문 헌

강기훈, 박은성, 신기일, 신민웅, 정석오, 최대우(2012). **베이지안 통계학**. 자유아카데미

강태훈, 김명연(2012). 모의실험 연구를 통한 등급반응모형과 일반화부분점수모형 비교. **교육평가연구, 25**(3), 479~496.

강태훈(2014). IRT 능력모수 추정에 있어서 검사총점에 근거한 사후기대추정법의 정확성에 관한 연구. **교육방법연구, 26**(1), 1~19.

강태훈, 백순근(2007). 3모수 문항반응이론의 능력모수 추정 방식과 검사의 문항곤란도 구성이 능력추정의 표준오차에 미치는 영향. **교육평가연구, 20**(1), 73-97.

김경희(1993). 문항수, 문항난이도, 문항변별도 변화에 따른 신뢰도 계수와 검사정보함수의 변화. 이화여자대학교 석사학위 논문.

김석호(1998). **다분문항반응의 이론과 실제**. 황정규 편. 교육측정·평가의 새 지평, 서울: 교육과학사, 177-247.

김성훈, 박인심(2010). IRT 능력 추정에서 정보적 사전분포에 기초한 EAP 방법과 무정보적 사전분포에 기초한 방법의 기능 비교. **교육평가연**

구. 23(2), 441-463.

김성훈(2012a). IRT 능력 추정에서 정보적 사전분포에 기초한 EAP 방법과 무정보적 사전분포에 기초한 방법의 기능 비교. **교육평가연구, 23**(2), 441-463.

김성훈(2012b). 문항반응이론(IRT) 능력 추정치들의 측정학적 특성에 관한 추수 연구. **교육평가연구, 25**(4), 829-849.

박정(1999a). 다분 문항반응이론 모형의 능력모수 추정치의 편파도 감소를 위한 모수 추정방법. **교육평가연구. 12**(2), 195-218.

박정(1999b). 검사의 길이, 반응 범주의 개수, 피험자의 수 및 피험자 능력 분포에 따른 다분문항반응이론 모형의 문항모수 추정치의 정확도. **교육평가연구. 12**(1), 17-42.

박정(2001). **다분문항반응모형**: 서울: 교육과학사.

박태준, 시기자, 신동광, 김성혜, 이용상, 윤지환, 박지선, 민호기, 박용효, 김준식, 정채관, 임수연, 주헌우, 김미지, 박찬호, 반재천, 조동완(2013). 대학 입시 전형 간소화 정책과 연계한 NEAT(2,3급) 개선 방안 연구. RRE 2013-1, 한국교육과정평가원.

선은주(2011). 작은 샘플 크기의 One-shot device를 위한 베이지안 신뢰도 추정, 한양대학교 : 산업공학과.

성태제(1998). 다분문항반응이론(등급반응모형)에 의한 학구적 실패내성척도의 문항분석과 피험자 특성추정. **교육심리연구, 12**(2), 203-218.

송정무(2013). 베이지안 기법을 이용한 보증데이터 부석방법 연구. 성균관대학교 석사학위 논문.

임미경(2001). 등급반응모형, 평정척도모형, 부분점수모형의 문항모수와 피험자 모수 추정치 비교분석. 이화여자대학교 교육학과 석사학위논문.

한국성인교육학회(1998). 교육평가 용어사전. 학지사.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-57

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques*. NY: Marcel Dekker Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Cai, L., du Toit, S. H. C., Thisse, D. (2011). *IRTPRO: Fleible, multidimensional, multiple categorical IRT modeling*. Chicago, IL: Scientific Software International.

Childs, R, A. & & Chen, W. (1999). Obtaining comparable item parameter estimation in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 23*(4), 371-379.

De Ayala, R. J. (1995). *Itme parameter recovery for the nominal response model*. Paper presented at the annul meeting of the American Education Research Association, San Franciscon, CA.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. Applied *Measurement in Education, 11*(4), 17-34.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with plytomous items. *Applied Measurement in Education, 19*(1). 5-22.

Garner, W. R. (1960). Rating scales, discmininability, and information transmission. *Psychological Review, 67*(6), 343.

Guilford, J. P. (1954). *Psychometric methods*. New York, NY; US:

McGraw-Hill. Nunnally, 1978Finn, 1972

Hamada, M. S. (2008). *Bayesian Reliability*, Springer Series in Statistics.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice. 12*(3). 38-47.

Kang, T. H., Cohen, A. S., & Sung, H. J. (2005). IRT Model Selection Methods for Polytomous Items. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kass, R. E., & Raftery. (1995). Bayes Factors. *Journal of the American Statistical Association. Vol. 90*, No. 430, pp. 773-795.

Kim, J. (2007). A comparison of calibration methods and proficiency estimators for creating IRT vertical scales. (Doctoral dissertation). Retrieved from http://ir.uiowa.edu/etd/163

Kolen, M. J., & Brennan, R. L. (2004). *Test equating methods and practice*. New York: Springer-Verlag.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT

proficiency estimates. *Educational MEasurement: Issues and Practice, 29*(3), 8-14.

Kolen, M. J. (2012). *Scores and sclaes: considerations for PARCC assessments*. Retrieved from http://www.parcconline.org/sites/parcc/files/KolenPARCCScoresandScales.pdf

Lee, W. -C., & Ban, J. -C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*, 23-48.

Linacre, J. M. (1989). Many-faceted Rasch model. Chicago: MESA Press.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometirika, 47*(2), 149-174.

Mckelvie, S. J. (1978). Graphic rating scales ?How many categories? *British Jornal of Pshchology, 29*(2), 185-202.

Muraki, E. & Wang, M. (1992). *Issues relating to the marginal maximum likelihood estimation of the partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muraki, E., & Bock, R. D. (1998). *PARSCALE: IRT based test coring and item analysis for graded open-ended exercises and performance tasks*. Chicago, IL: Scientific software.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded score. *Psychometrika Monograph Supplement, No.17*.

Stroud, A. H. (1974). *Numerical quadrature and solution of ordinary differential equations*. New York: Springer-Verlag.

Stucky D. Brian. (2009). Item Response Theory for Weighted Summed Scores. University of North Carolina at Chapel Hill. A thesis for the degree of Master of Arts in the Department of Psychology.

Thissen, D. M. (1976). Information in wrong responses to the Raven progressive mattrices. *Journal of Educational Measurement, 13*, 201-214.

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]*. Chicago: Scientific Software International.

Thissen, D., & Orland, M. (2001). *Item response theory for items scored In two categories*. In D. Thissen and H. Wainer (Eds.), Test scoring (pp. 73-140). Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc., Publishers.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses, *Applied Psychological Measurement. 19*(1), 39-49.

Tissen, D. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*(4), 501-519.

Thissen, D.. & Steinberg, L. (1986). A taxonomy of item response theory, *Psychometrika, 51*(4), 567-577.

Tong, Y., & Kolen, M. J. (2010). *IRT proficiency estimators and their impact*. Paper presented at the annual conference at the National Council on Measurement in Education, Denver, CO.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Jorunal of Mathematical and Statistical Psychology. 43*(1). 39-55.

Walsh, J. E. (1963). *Corrections to two papers concerned with binomial*

events. Sankhya, 25, Series A, 427.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450;

Yen, W. M. (1984). Obtaining maximum likelihood trait estimaties from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93-111.

ABSTRACT

A Study on the Accuracy of IRT Ability Parameter Estimates based on EAP method under the Polytomous IRT model

SHIM, Hye-jin

Dept. of Education

The graduate school of Sungshin Women's University

This study investigates the accuracy of IRT ability parameter estimate based on summed score EAP method(EAPss) under the polytomous IRT model through several simulation conditions, comparing traditional EAP method based on item response pattern(EAPrp). The former has an advantage in that it can produce a convincing scale score to the public by estimating one-to-one ability estimates with the number correct scores. However, EAPss also has an disadvantage, ignoring the information which the item response pattern has. Even though EAPss has the problem of losing information, this study wants to show the accuracy of EAPss is similar to EAPrp. For this, the study compares the accuracy of recovery of true ability parameter under the several simulation conditions, and presents the results with the value of MSE, SB, and VAR. The result shows that EAPrp produces a little bit smaller value of MSE, SB, and VAR than EAPss, however the difference happens at decimal places. Moreover, the results of both two EAP have the close value of 0 on the ability parameter

scale from −2 to 2. Therefore, the results imply that the use of EAPss can be the promising alternative under the operation of an actual testing program.

부 록 1: 3개의 범주를 가진 실제 자료의 문항 모수

	GRM			GPCM		
Item	a	$b1$	$b2$	a	b	$\tau 1$
1	1.189	-1.211	1.766	1.159	-0.420	1.260
2	0.965	-1.324	1.218	0.513	-0.244	0.664
3	1.517	-0.364	1.835	1.429	0.614	1.472
4	2.483	-0.620	1.824	2.252	-0.370	0.737
5	0.585	-1.485	0.225	0.708	0.163	1.229
6	1.133	-2.955	0.589	1.536	0.597	0.762
7	1.634	0.237	2.213	1.872	0.113	0.516
8	0.823	-2.406	0.806	0.451	-0.400	0.650
9	1.972	-2.382	0.460	0.487	-0.376	1.566
10	1.212	-2.078	1.170	1.331	0.150	0.719
11	1.098	-1.776	1.035	0.819	-0.187	0.912
12	0.798	0.678	2.427	1.412	-0.030	0.668
13	2.021	-2.098	0.925	1.504	0.364	1.177
14	1.848	-0.211	1.418	1.428	0.346	0.525
15	1.476	-1.000	1.688	1.906	-0.292	1.028
16	1.404	-1.971	0.152	1.397	-0.345	0.972
17	2.467	-1.512	1.909	1.814	0.160	0.789
18	0.935	-1.354	0.851	0.548	-0.246	0.979
19	1.244	-1.138	2.254	0.990	0.208	0.373
20	1.654	-1.105	1.312	0.925	0.195	1.272
21	1.189	-1.211	1.766	1.159	-0.420	1.260
22	0.965	-1.324	1.218	0.513	-0.244	0.664
23	1.517	-0.364	1.835	1.429	0.614	1.472
24	2.483	-0.620	1.824	2.252	-0.370	0.737
25	0.585	-1.485	0.225	0.708	0.163	1.229
26	1.133	-2.955	0.589	1.536	0.597	0.762
27	1.634	0.237	2.213	1.872	0.113	0.516
28	0.823	-2.406	0.806	0.451	-0.400	0.650
29	1.972	-2.382	0.460	0.487	-0.376	1.566
30	1.212	-2.078	1.170	1.331	0.150	0.719
31	1.098	-1.776	1.035	0.819	-0.187	0.912
32	0.798	0.678	2.427	1.412	-0.030	0.668
33	2.021	-2.098	0.925	1.504	0.364	1.177
34	1.848	-0.211	1.418	1.428	0.346	0.525
35	1.476	-1.000	1.688	1.906	-0.292	1.028
36	1.404	-1.971	0.152	1.397	-0.345	0.972
37	2.467	-1.512	1.909	1.814	0.160	0.789
38	0.935	-1.354	0.851	0.548	-0.246	0.979
39	1.244	-1.138	2.254	0.990	0.208	0.373
40	1.654	-1.105	1.312	0.925	0.195	1.272
평균	1.423	-1.304	1.304	1.224	0.000	0.913
표준편차	0.526	0.908	0.668	0.522	0.330	0.326

부 록 2: 5개의 범주를 가진 실제 자료의 문항 모수

	GRM					GPCM				
Item	a	$b1$	$b2$	$b3$	$b4$	a	b	$\tau 1$	$\tau 2$	$\tau 3$
1	1.189	-1.593	-0.828	1.250	2.281	1.159	-0.420	2.562	-0.042	-1.665
2	0.965	-2.353	-0.295	0.601	1.836	0.513	-0.244	0.877	0.452	-1.665
3	1.517	-0.671	-0.058	1.285	2.386	1.429	0.614	3.048	-0.104	-0.950
4	2.483	-1.198	-0.043	1.223	2.425	2.252	-0.370	-0.408	1.882	0.000
5	0.585	-1.836	-1.134	-0.172	0.621	0.708	0.163	2.346	0.112	-0.671
6	1.133	-3.681	-2.230	-0.302	1.480	1.536	0.597	1.445	0.078	-0.264
7	1.634	-0.584	1.058	1.811	2.616	1.872	0.113	1.272	-0.240	0.499
8	0.823	-3.831	-0.980	0.487	1.125	0.451	-0.400	1.899	-0.600	-0.281
9	1.972	-3.508	-1.256	0.130	0.790	0.487	-0.376	3.172	-0.039	-2.081
10	1.212	-2.506	-1.650	0.720	1.620	1.331	0.150	1.587	-0.149	-0.341
11	1.098	-2.152	-1.400	0.594	1.476	0.819	-0.187	2.201	-0.378	-1.199
12	0.798	0.212	1.144	2.042	2.811	1.412	-0.030	0.733	0.603	-0.740
13	2.021	-3.070	-1.125	0.335	1.516	1.504	0.364	1.234	1.119	0.377
14	1.848	-0.637	0.216	1.002	1.834	1.428	0.346	0.031	1.019	0.283
15	1.476	-1.973	-0.026	0.964	2.412	1.906	-0.292	0.494	1.561	-1.358
16	1.404	-2.637	-1.304	-0.328	0.632	1.397	-0.345	1.676	0.267	-0.017
17	2.467	-2.090	-0.935	1.419	2.399	1.814	0.160	1.158	0.420	-1.243
18	0.935	-1.913	-0.795	0.439	1.263	0.548	-0.246	2.141	-0.184	-1.438
19	1.244	-1.613	-0.664	1.662	2.846	0.990	0.208	1.605	-0.858	0.407
20	1.654	-2.047	-0.162	0.667	1.958	0.925	0.195	1.620	0.923	-0.162
21	1.189	-1.593	-0.828	1.250	2.281	1.159	-0.420	2.562	-0.042	-1.665
22	0.965	-2.353	-0.295	0.601	1.836	0.513	-0.244	0.877	0.452	-1.665
23	1.517	-0.671	-0.058	1.285	2.386	1.429	0.614	3.048	-0.104	-0.950
24	2.483	-1.198	-0.043	1.223	2.425	2.252	-0.370	-0.408	1.882	0.000
25	0.585	-1.836	-1.134	-0.172	0.621	0.708	0.163	2.346	0.112	-0.671
26	1.133	-3.681	-2.230	-0.302	1.480	1.536	0.597	1.445	0.078	-0.264
27	1.634	-0.584	1.058	1.811	2.616	1.872	0.113	1.272	-0.240	0.499
28	0.823	-3.831	-0.980	0.487	1.125	0.451	-0.400	1.899	-0.600	-0.281
29	1.972	-3.508	-1.256	0.130	0.790	0.487	-0.376	3.172	-0.039	-2.081
30	1.212	-2.506	-1.650	0.720	1.620	1.331	0.150	1.587	-0.149	-0.341
31	1.098	-2.152	-1.400	0.594	1.476	0.819	-0.187	2.201	-0.378	-1.199
32	0.798	0.212	1.144	2.042	2.811	1.412	-0.030	0.733	0.603	-0.740
33	2.021	-3.070	-1.125	0.335	1.516	1.504	0.364	1.234	1.119	0.377
34	1.848	-0.637	0.216	1.002	1.834	1.428	0.346	0.031	1.019	0.283
35	1.476	-1.973	-0.026	0.964	2.412	1.906	-0.292	0.494	1.561	-1.358
36	1.404	-2.637	-1.304	-0.328	0.632	1.397	-0.345	1.676	0.267	-0.017
37	2.467	-2.090	-0.935	1.419	2.399	1.814	0.160	1.158	0.420	-1.243
38	0.935	-1.913	-0.795	0.439	1.263	0.548	-0.246	2.141	-0.184	-1.438
39	1.244	-1.613	-0.664	1.662	2.846	0.990	0.208	1.605	-0.858	0.407
40	1.654	-2.047	-0.162	0.667	1.958	0.925	0.195	1.620	0.923	-0.162
평균	1.423	-1.984	-0.623	0.791	1.816	1.224	0.000	1.535	0.292	-0.625
표준편차	0.526	1.060	0.844	0.672	0.692	0.522	0.330	0.910	0.702	0.777

부 록 3: EAPrp를 구하기 위한 MATLAB 코드

```
%%% ability parameter estimation method: EAPrp
%%% three kinds of priors
%%% 1) N(0,1)
%%% 2) U(-4,4)
%%% 3) posterior (after applying N(0,1))
clear all;

%%% GPCM
n = 10000; % number of examinees (simulated)
T=20; nc=3;
nos = 41; % number of scores when 40 is maximum test score <-
(0,1,2) for each item
dfn ='GPCM1.txt';  % data file name
% read data
cadata = dlmread(dfn); % cadata = cadata - 1;
% item parameter estimates
load 'paraGPCM3_20.txt'; itempara = paraGPCM3_20;
% Quadrature Points to develop Lx (NOS by K)
theta=-4:.2:4; N=length(theta);

% prior 1): N(0,1)
eapall = zeros(n,2);
tempwei=normpdf(theta,0,1); abilwei=tempwei/sum(tempwei);
```

```
for j=1:n
    eaprp1 = zeros(2,1); resp=zeros(1,T); resp=cadata(j,:);
    eaprp1 = ind_eap(itempara,T, nc,resp,theta, N, abilwei);      % EAPrp
when prior is N(0,1)
    eapall(j,1)=eaprp1(1);
    eapall(j,2)=eaprp1(2);
end
dlmwrite('EAPrp1.sco', eapall);
eaprp1_all = eapall; % this is for the eaprp3 below

% prior 2): U(-4,4)
eapall = zeros(n,2);
abilwei = ones(1,41)*(1/41);
for j=1:n
    eaprp2 = zeros(2,1); resp=zeros(1,T); resp=cadata(j,:);
    eaprp2 = ind_eap(itempara,T, nc,resp,theta, N, abilwei);      % EAPrp
when prior is N(0,1)
    eapall(j,1)=eaprp2(1);
    eapall(j,2)=eaprp2(2);
end
dlmwrite('EAPrp2.sco', eapall);

% prior 3): posterior (after applying N(0,1))
eapall = zeros(n,2);
[cc,bb] = hist(eaprp1_all(:,1),N); cc = cc/sum(cc); % the posterior
```

```
distribution
abilwei = cc;
for j=1:n
      eaprp3 = zeros(2,1); resp=zeros(1,T); resp=cadata(j,:);
      eaprp3 = ind_eap(itempara,T, nc,resp,theta, N, abilwei);      % EAPrp
when prior is N(0,1)
      eapall(j,1)=eaprp3(1);
      eapall(j,2)=eaprp3(2);
end
dlmwrite('EAPrp3.sco', eapall);
```

부 록 4: EAPss를 구하기 위한 MATLAB 코드

```
%%% ability parameter estimation method: EAPss
%%% three kinds of priors
%%% 1) N(0,1)
%%% 2) U(-4,4)
%%% 3) posterior (after applying N(0,1))
clear all;

%%% GP111
n = 10000; % number of examinees (simulated)
T=20; nc=3;
nos = 41; % number of scores when 40 is maximum test score <-
(0,1,2) for each item
dfn ='GP111.txt';  % data file name
% read data
cadata = dlmread(dfn); cadata = cadata - 1;
% item parameter estimates
load 'paraGPCM3_20.txt'; itempara = paraGPCM3_20;
% Quadrature Points to develop Lx (NOS by K)
theta=-4:.2:4; N=length(theta);

% prior 1): N(0,1)
eapss1 = zeros(n,2);
tempwei=normpdf(theta,0,1); abilwei=tempwei/sum(tempwei);
```

```
eapss1 = ind_eapss(itempara,cadata,nc,nos,theta, N, abilwei);
dlmwrite('EAPss1.sco', eapss1);

% prior 2): U(-4,4)
eapss2 = zeros(n,2);
abilwei = ones(1,41)*(1/41);
eapss2 = ind_eapss(itempara,cadata,nc,nos,theta, N, abilwei);
dlmwrite('EAPss2.sco', eapss2);

% prior 3): posterior (after applying N(0,1))
eapss3 = zeros(n,2);
[cc,bb] =hist(eapss1(:,1),N); cc = cc/sum(cc); % the posterior distribution
abilwei = cc;
eapss3 = ind_eapss(itempara,cadata,nc,nos,theta, N, abilwei);
dlmwrite('EAPss3.sco', eapss3);
```