강 태 훈 교수지도

석사학위 청구논문

시뮬레이션 연구를 통한 일반화부분점수모형과 등급반응모형의 적합도 비교

2012

성신여자대학교 대학원

교육학과

김 명 연

시뮬레이션 연구를 통한 일반화부분점수모형과 등급반응모형의 적합도 비교

강 태 훈 교수지도

이 논문을 석사학위논문으로 제출함

2011년 11월

성신여자대학교 대학원

교육학과

김 명 연

인 준 서

김명연의 석사학위 논문으로 인준함.

심사위원_____(인)

심사위원_____(인)

심사위원_____(인)

성신여자대학교 대학원

논 문 개 요

본 연구는 다분적으로 채점되는 문항으로 이루어진 학업성취도 및 심리 검사 결과를 분석 할 때 주로 사용되고 있는 GPCM과 GRM 두 모형이 실질적으로 자료를 설명하는 힘 즉, 적합도에 있어서 어떤 차이를 가지고 있는지 밝히는 데 있다. 이를 위하여 시뮬레이션 연구를 통하여 두 모형 간의 모형 적합도에 있어서의 차이점을 살펴보았다. 각 모형을 이용하여 다양한 수준의 조건에서 자료들을 생성하고, 동일 자료에 대하여 두 모형 모두를 이용하여 분석하였다. 이러한 결과를 이용하여 다음과 같은 연구가 수행되었다.

1. 문항 적합도 지수(item fit index)를 사용하여 각 모형이 자신에 의해 생성된 자료들을 문항 수준에서 다른 모형에 비해 보다 잘 적합할 수 있는지를 살펴보는데 있어서 최근 제1종 오류를 통제하는 측면과 통계적 검정력 측면에서 가장 우수한 문항 수준 적합도 지수인 $S-X^2$지수를 사용하였다.
2. 전체 검사 수준에서의 모형-자료 적합도(model-data overall fit)를 구하여, 두 모형이 자료의 적합에 있어서 유의미한 차이가 있는지 살펴보고, 각 자료에 대한 관찰된 원점수 분포와 적용된 모형에 의해 재생산된(reproduced) 원점수 분포를 비교하여 CS_1의 값과 $S-X^2$를 활용한 CS_2를 살펴보았다.

본 연구를 위해 GRM과 GPCM으로 다분 문항 자료를 생성하였다. 시뮬레이션 연구를 위한 시뮬레이션 조건은 두 개의 검사 길이(I=10, 20)와 두 개의 표본 크기(N=500, 1,000)를 고려하였다. 각 조건에서 50개의 자료를 반복적으로 생성하여 두 모형을 각기 적용하여 문항 모수를 추정하였다. 각 조건의 개별 자료에 대하여 두 모형을 적합한 결과를 바탕으로 $S-X^2$와 검사 전체 적합도 지

수인 CS_1과 CS_2를 계산하여 그 결과를 상호 비교하였다.

본 연구의 결과를 요약하면 다음과 같다.

1. $S-X^2$를 통해 살펴 본 결과 문항 적합도 수준에서 생성모형과 분석모형이 동일할 때 경험적 제1종오류가 유의수준 .05와 비슷하게 나타났다.
2. CS_1과 CS_2를 통하여 카이자승 검정을 실시한 결과 제1종 오류의 통제라는 측면에서 좋지 않은 결과가 산출되었다. 이를 통해 CS_1 통계지표의 개선이 필요함을 알 수 있었다.
3. CS_2의 결과가 .05와 비슷하지는 않으나 CS_1의 결과보다는 더 합리적인 결과를 도출하였다.
4. 문항 적합도 지수를 살펴보았을 때 GPCM이 GRM보다 상대적으로 주어진 자료가 어떤 모형으로 생성되었는가에 관계없이 보다, 설명력 및 적합도가 더 높았다.

주요어 : 등급반응모형, 일반화부분점수모형, 제1종오류, $S-X^2$, CS_1, CS_2

목 차

논문개요

표 목 차

그 림 목 차

Ⅰ. 서 론

1. 연구의 필요성 및 목적

교육현장에서 검사를 통하여 피험자의 특성을 잘 파악하는 것은 무엇보다 중요하다. 더욱이 학생의 수학적인 능력이나 지능과 같은 인간의 심리적 특성들은 눈에 보이는 물리적인 특성과는 달리 직접 관찰할 수 없어 그 측정이 어렵다. 그러므로 정의적·인지적 특성들을 외현화하기 위한 타당하고 신뢰로운 검사를 제작할 필요가 있다.

또한, 검사 문항에 대한 피험자의 반응으로부터 직접 관찰이 불가능한 개인의 잠재적 특성과 능력을 보다 정확하게 추정하는 것이 중요하며 이에 관련한 학문적 영역을 측정 및 검사이론이라고 부른다.

사회과학 분야에서 사용하고 있는 가장 대표적인 검사이론은 고전검사이론(classical test theory, 이하 CTT)과 문항반응이론(item response theory, 이하 IRT)이다. 고전검사이론은 대부분의 검사 자료에 쉽게 만족될 수 있어서 광범위하게 적용될 수 있다는 장점이 있는 반면에, 검사점수 자체를 분석 대상으로 하고 피험자의 능력 추정이 사용 검사에 좌우되는 등의 단점을 지닌다. 또한 모든 피험자에 대하여 동일한 측정 오차를 가정하여 실제로 존재할 수밖에 없는 능력 추정 정확도에서의 개인 간 차이를 무시하는 문제를 가지고 있다. IRT는 검사 문항을 분석 단위로 하며 각 문항이 지니고 있는 고유의 특성이 문항에 응답한 피험자 집단의 특성에 의하여 변화되지 않는다는 문항모수불변성 개념(the invariance concept of item parameter)과 피험자의 능력이 검사의 특성에 의하여 달리 추정되지 않음을 말하는 능력모수불변성 개념(the invariance concept of ability parameter)을 바탕으로 하기 때문에 기존 CTT의

단점을 상당 부분 해결해 주는 것으로 알려져 있다. 또한 한 문항의 모든 양적 특성을 문항특성곡선(item characteristic curve, 이하 ICC)이라는 그림으로 표현할 수 있으며, 피험자의 능력 모수와 문항의 모수를 같은 척도 상에 놓을 수 있다는 장점을 가지기 때문에 컴퓨터화 검사라든가 맞춤형 검사 제작 등에 있어서 필수적인 역할을 하고 있다.

문항반응이론의 모형 중 간단한 형태는 이분 문항반응이론(dichotomous item response theory)[1]이다. 나아가 교육 및 심리검사 문항이 두 개 이상의 범주로 채점될 때 이를 다분 문항(polytomous item)이라고 한다. 다분문항반응은 범주 수가 맞고 틀리고의 경우처럼 2개인 이분문항반응을 확장한 경우라 할 수 있으며, 다분 문항반응모형은 이러한 다분 문항으로 이루어진 검사 자료를 다루기 위한 방법론이다.

1970년대까지는 문항반응이론의 적용이 이분문항에 집중하여 논의되고 발전되었으나, 점차 개방형 질문(open-ended question), 논술형 검사, 수행평가의 확산으로 인하여 다분화 된 문항 반응 결과에 대한 분석을 어떻게 할 것인가에 대한 관심이 고조되었다(구슬기, 2010). 달리 말하여, 이분문항반응이론을 확장시킨 형태가 다분문항반응이론(polytomous IRT, 이하 PIRT)이며 이는 피험자의 반응이 보통 정답과 오답이 아닌, 3개 이상의 범주로 나타나는 경우를 다룬다. 예를 들어, 선다형 문항에서 맞으면 1점, 틀리면 0점의 점수를 매기는 것이 대표적 이분 문항 반응 자료라면, 서술형 평가에서는 0점부터 특정 점수까지 부분점수를 통해 수험자의 응답 형태에 따른 점수를 부여할 수 있다. 이렇듯 다분화된 점수부여를 통하여 피험자들의 반응을 단계적으로 세분화하여 이분반응모형이 간과하는 중간 단계에서의 능력에 대한 정보를 활용할 수 있다.

다분문항반응이론에서는 피험자들의 반응을 문항반응범주(item response

1) 피험자의 반응이 두 가지인 경우에 사용된다. 즉, 선다형 문항의 맞고, 틀리고의 두 종류의 점수를 줄 때 사용되어진다. 이분 문항에 대한 문항반응모형에서 능력 수준 θ의 능력을 가진 피험자가 문항의 답을 맞힐 확률은 $P(\theta)$이다.

category)로 분류하며, 문항반응범주는 한 문항에 대하여 피험자의 반응이 생길 수 있는 가능 한 종류를 말하는 것으로서 연속적인 범주들이 연속적인 숫자로 점수화될 수 있다(한국교육평가학회, 2004). 대표적인 예는 태도(attitued)나 인성검사(personality)에서 주로 사용되는 Likert 척도이다. 문항반응이론에서는 이러한 검사의 자료를 분석하기 위하여 여러 종류의 다분 문항반응 모형들이 개발되어 왔다.

Samejima(1969)의 등급반응모형(graded response model, 이하 GRM)을 시작으로 정답과 오답사이에 부분점수를 부여하는 경우나 Likert 척도와 같은 서열된 척도에 적용 가능한 모형이 개발되었다. 또한 Muraki(1992)의 일반화부분점수모형(generalized partical credit model, 이하 GPCM), Master(1982)의 부분점수모형(partial credit model, 이하 PCM), 비서열적인 척도에 적용가능한 모형인 Bock(1972)의 명명반응모형(nominal response model, 이하 RNM) 등이 개발되었다. 그 밖에도 Andrich(1978)의 평정척도모형(rating scale model, 이하 RSM) 등이 사용되고 있다. PCM은 Rasch 모형 계열에 속하며 GPCM의 특수한 경우로서 모든 문항의 변별도를 같다고 본다. RSM 모형은 Likert와 같은 평정척도 또는 똑같이 정렬된 범주를 가지고 있는 의미변별척도와 관련된 문항에 주로 적용된다. 즉 RSM은 점수 수준 사이에 거리가 모든 문항을 위해 일정할 것으로 가정하며 앞에서 소개한 PCM의 특수한 경우이다.

Thissen(1986)은 모형이 전개되는 과정에 따라 여러 일차원성 가정 하의 문항반응모형들을 차이모형(difference models)과 나눔모형(divide by total)으로 분류하였는데, GRM은 전자에 해당하며 GPCM은 후자에 속한다. 이렇듯 현재까지 다양한 형태의 다분문항반응을 위한 모형들이 개발되어 왔으며, 각 모형은 이론적 가정과 모수의 개수 등에서 차이를 가진다. 특히, 정답만을 확인하여 점수를 부여하는 선다형 위주의 검사에서 심동적, 인지적, 정의적인 모든 능력에 걸친 평가를 강조하는 수행평가가 점차 대두됨에 따라서 이러한 다분적

문항 자료 분석을 위해 활용될 수 있는 모형에 관한 연구들이 활발하게 진행되어 왔다.

교육 및 심리 검사의 결과를 IRT를 통하여 분석하려 할 때, 일차원성[2] 가정의 검증 같은 기본적 검증만큼 중요한 요소로서 주어진 자료들을 분석하기 위하여 사용할 모형을 선택하고 검토하는 일을 들 수 있다. 앞서 소개한 다분문항 모형 중에서 가장 흔히 쓰이는 다분 문항반응은 Samejima의 GRM과 Murakii의 GPCM이다(Kang, Cohen, & Sung, 2009). 두 모형은 같은 수의 문항 모수를 사용하며, 문항특성곡선(item characteristic curve; ICC)[3] 상에서 문항마다 다른 기울기가 존재한다는 공통점을 가지고 있다.

또한, Maydu-Olivares, Drasgow, & Mead(1994)는 이 두 모형을 같은 자료에 적용하였을 때 대부분의 경우 같은 정도의 적합성을 보일 것이라고 주장하기도 하였다. 하지만 최근 Kang, Cohen, & Sung(2009)이 지적하였듯이 이 두 모형이 별다른 차별성 없이 여느 다분 문항 자료에 적용되어도 그 분석 결과가 같다고 결론짓는 것은 다소 성급한 것으로 보인다. 그들은 모형 선택 지수(model selection index)[4]를 사용한 시뮬레이션 연구를 통하여, 각 모형이 자신을 생성모형 혹은 진모형(generating model or true model)으로 하는 자료들을 위해 더 나은 모형으로 기능할 수 있다는 것을 확인하였고, 이러한 점에서 두 모형이 상호 구별 없이 사용되기에는 무리가 있다고 언급하였다.

---

2) 검사가 인간의 잠재적 특성을 측정할 때 유일한 특성(예를 들어, 수학검사에서 수학능력만을 측정)을 측정하여야 함을 의미한다.

3) 각 검사 문항에 대하여 정답 확률과 능력 척도 사이의 관계를 나타내는 것이다. 각 문항이 얼마나 잘 달성하고 있는지에 대하여 문항특성곡선을 보면 알 수 있으며, 문항특성곡선의 기울기가 클수록(가파른 곡선) 그 문항은 능력의 상하집단을 더 잘 변별하는 문항이다.

4) 자료의 적합에 뛰어난 하나의 모형을 찾기 위하여 모형 선택 지수들이 사용되어왔다. 다양한 다분문항반응 모형들을 상호 비교할 때 매우 효과적인 도구로 알려져 있다.

2. 연구의 목적 및 개요

이러한 선행 연구들에 있어서, 두 모형이 실제 모형 적합도의 측면에서 차별성이 있는 지 즉, 주어진 자료를 설명하는 데에 있어서 그 설명력에 근본적인 차이를 가지는 지는 탐구된 바가 없다. 따라서 본 연구의 주된 목적은 자료 분석에 있어서 두 모형의 적합도 차이를 시뮬레이션 연구를 통하여 밝히는 데에 있다. 이를 통하여 일반 연구자나 검사자료 분석자들은 다분문항 자료를 분석할 때 과연 Maydu-Olivares, Drasgow, & Mead(1994)의 주장처럼 두 모형 중 아무 모형이나 사용할 수 있는 것인 지, 혹은 자신의 자료에 보다 더 적절한 모형을 선택하여 분석할 필요성이 있는 지에 대한 시사점을 얻을 수 있을 것이다.

본 연구의 목적과 방법을 요약하면 다음과 같다. 주된 연구 목적은 다분적으로 채점되는 문항으로 이루어진 학업성취도 및 심리 검사 결과를 분석할 때 주로 사용되고 있는 GPCM과 GRM 두 모형이 실질적으로 자료를 설명하는 힘 즉, 적합도에 있어서 어떤 차이를 가지고 있는지 혹은 같다고 볼 수 있는지를 밝히는 것이다.

이를 밝히기 위하여, 시뮬레이션 연구를 통하여 두 모형 간의 모형 적합도에 있어서의 차이점을 살펴보고자 한다. 각 모형을 이용하여 다양한 수준의 조건에서 자료들을 생성하고, 동일 자료에 대하여 두 모형 모두를 이용하여 분석하였다. 이러한 결과를 이용하여 다음과 같은 두 가지 연구가 수행되었다.

첫째, 문항 적합도 지수(item fit index)를 사용하여 각 모형이 자신에 의해 생성된 자료들을 문항 수준에서 다른 모형에 비해 보다 잘 적합할 수 있는지를 살펴보았다(연구Ⅰ). 이를 위하여 최근 제1종 오류를 통제하는 측면과 통계적 검정력 측면에서 가장 우수한 문항 수준 적합도 지수로서 각광받고 있는 $S\text{-}X^2$ 지수(Orlando & Thissen, 2000, 2003)를 사용하였다.

둘째, 전체 검사 수준에서의 모형-자료 적합도(model-data overall fit)를 구해본다. 두 모형이 자료의 적합에 있어서 유의미한 차이가 있다면, 각 모형은 자신에 의해 생성된 자료를 보다 잘 설명 및 적합해야 할 것이다. 이를 위하여 각 자료에 대하여 관찰된 원점수 분포와 적용된 모형에 의하여 재생산된(reproduced) 원점수 분포를 비교하여 카이자승 검정을 실시해 보았다. 이하에서는 이를 CS_1로 부르기로 한다. 또한 일종의 카이자승 분포를 따르는 문항 수준의 $S\text{-}X^2$ 통계치와 관련 자유도(degree of freedom)들을 모두 합하여 검사 전체 수준에서의 카이자승 검정을 실시하였다. 이는 카이자승 통계치의 가법성[5]을 활용한 것이다. 이하에서는 이를 CS_2라고 부르기로 한다. 본 연구에서는 전체 검사 수준에서의 모형 적합도를 이들 두 방법을 사용하여 살펴보았다 (연구 Ⅱ).

5) 카이자승의 가법성이란 한 변수 X1이 자유도 df1인 카이자승 분포를 따르고 다른 변수 X2가 자유도 df2인 카이자승 분포를 따르면, 각 변수에서 계산된 카이자승 통계치의 합은 자유도가 df1+df2인 카이자승 분포를 따른다는 의미이다.

Ⅱ. 이론적 배경

IRT 모형들이 많은 연구자들에 의해서 다양하게 개발되었고, 심리 척도의 제작이나 교육측정 등의 분야에 폭넓게 적용되어 왔다. IRT에서 자료 분석은 대개 하나의 심리측정학적 모형을 상정하여 선택된 모형이 자료를 제대로 설명할 것이라는 기대에서 시작된다. 주어진 자료를 제대로 요약하지 못하는 모형을 사용하면, IRT의 적용으로 인해 얻을 수 있는 이점들에 대한 적용이 제대로 실시 될 수 없을 뿐만 아니라 궁극적으로 IRT의 최대 장점 중 하나인 모수불변성이 유지될 수 없다. 이러한 이유로 모형과 자료의 적합도를 살피는 과정이 매우 중요시 되며, 수많은 관련 참고문헌에서는 전체 검사 자료 수준에서의 적합도 뿐만 아닌, 각각의 문항 수준에서의 모형-자료 적합도를 검토하기 위한 여러 가지 방법을 소개하였으며, 그 수행 정도를 평가하였다.

이러한 방법들은 우선 도표를 이용한 문항 수준의 모형 적합도를 눈으로 살펴보는 접근법과 Hambleton & Swaminathan(1985)이 피험자를 동질적인 몇 개의 집단으로 묶은 뒤에 이들의 실제적 문항 수행 결과를 모형에 의해 예측되는 비율에 대비하여 그림으로 살펴보는 방법을 제안하였으며, Mislevy & Bock(1990)도 이와 유사한 도표 사용하는 법을 예시하였고, Kingston & Dorans(1985)는 문항 수준의 적합도를 검증하는 도구로서 문항-능력 회귀분석 기법을 사용할 것을 제안하였다.

위와 같은 노력에 더하여, X^2를 기반으로 해서 유의도 검증을 통하여 문항 수준의 모형-자료 적합도를 보기 위한 다양한 기법들이 나타났다. 변별도와 추측도 등을 고려하는 모형들에서는 비록 상황이 조금 더 복잡해지나, 문항 적합도를 검증하기 위해 사용될 수 있는 여러 지수가 꾸준히 개발되어 왔으며, 이러한 종류의 적합도 지수로는 Yen(1981)의 Q_1, Bock(1972)의 X^2, McKinley &

Mill(1985)의 G^2, 그리고 Orlando & Thissen(2000, 2003)의 $S-G^2$와 $S-X^2$ 등을 들 수 있다. X^2은 관찰빈도와 기대빈도가 비슷하면 통계량도 적어지며 귀무가설을 긍정할 가능성이 커진다. 결론적으로 관찰치 전체의 수와 각 칸의 빈도에 의해 좌우되는 특성을 가지고 있다. X^2과 비슷하게, G^2 역시 검사의 길이와 피험자 집단의 크기에 매우 민감하게 반응하여 조건에 따라 문항 적합도를 올바르게 파악하는 수행력에 있어서 큰 차이가 나는 것으로 밝혀졌다.

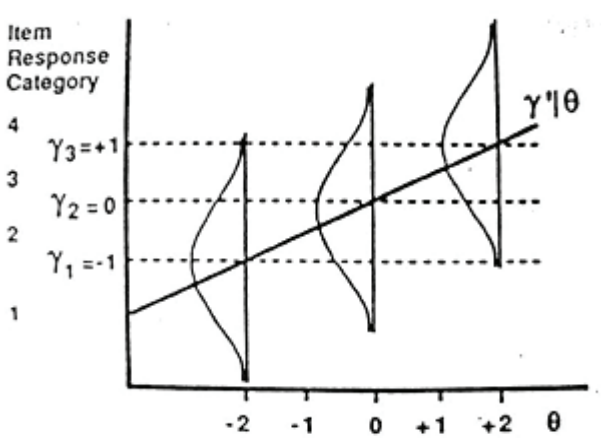
이와는 다르게 $S-G^2$와 $S-X^2$ 은 Orlando & Thissen(2000, 2003)과 Kang & Chen(2008)의 연구에서 보고된 것과 같이 검사에 있어서 문항의 수에 민감하게 영향을 받지 않으며, 피험자의 수가 500명에서 5,000명에 이를 때 까지 안정된 수준의 경험적 제1종 오류(empirical type I error rates)를 보였다. 또한 부적합 문항을 찾아내는 통계적 검정력(statistical power) 측면에 있어서도 다양한 연구 조건에서 우수한 수행을 보이는 것으로 나타났다. 이러한 결과에 고무되어 Kang & Chen(2011)은 Samejima(1969)의 GRM 하에서, Zhang & Stone(2008)은 다차원 문항반응모형 하에서, 그리고 Chun, Dunbar, & Lee (2009)는 한 검사 속에 혼합 유형의 문항들이 사용된 경우에 S-X^2의 수행이 어떠한 지를 연구하였다.

따라서 본 연구에서는 문항반응모형의 자료 적합도를 살펴보는 데에 있어서 가장 우수한 적합도 지수로 밝혀진 S-X^2를 중심으로 GPCM과 GRM 두 모형 간의 문항 수준의 적합도를 상호 비교하였고, 이를 확장하여 전체 검사 수준에서의 적합도 또한 살펴보았다.

1. 등급반응모형(Graded Response Model: GRM)

Samejima(1969)의 GRM은 교육현장의 성취검사 이외에 태도검사나 선호도 검사와 같은 심리학 분야에서도 문항반응이론을 폭 넓게 응용하는데 기여하였다(박정, 2001). 예를 들어 미술 시험에서 창의성, 독창성과 색의 사용 등의 여러 문항에 대하여 1점부터 5점까지 5개의 범주로 점수가 주어진 경우, GRM을 적용할 수 있을 것이다. GRM은 각 범주가 순서화된 가중치를 갖는 경우에 사용할 수 있는데, 즉, Likert-type 문항으로 예를 들면, 하나의 태도검사 문항의 범주가 '매우 동의함', '동의함', '관심 없음', '반대함', 그리고 '매우 반대함' 등의 5개 범주로 이루어지는 경우가 될 수 있다.

GRM은 정답과 오답 사이의 중간 능력을 고려하여 피험자들의 반응을 상이한 등급인 m개의 범주로 분류하여 점수화 한다(성태제, 1998). 여기서 m번째 범주가 가장 높은 등급이 되고, 첫 번째 범주가 가장 낮은 등급이 된다. 이런 방법으로 채점할 때 피험자의 반응을 등급반응(graded response)이라 하며, 그 문항을 등급문항(graded item)이라 한다(김보연, 2005).



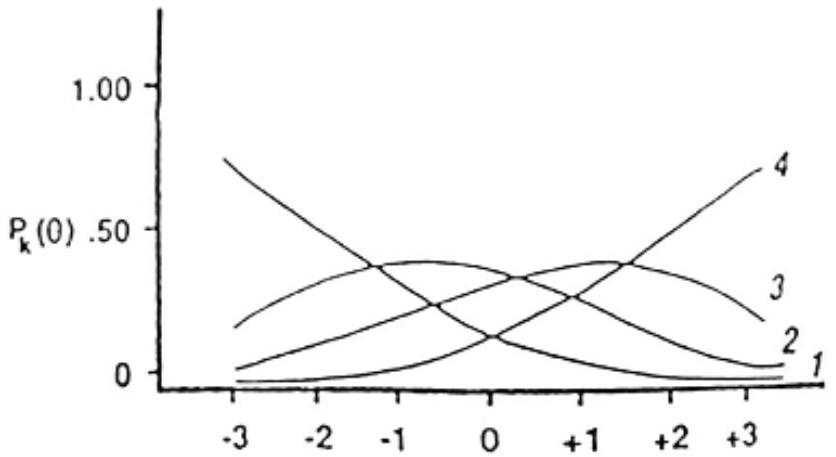
[그림 Ⅱ-1] 등급문항에 대한 문항반응모형의 원리

Samejima는 GRM을 만들기 위하여 2-모수[6] 이분반응 모형을 설명했던 Lord와 Novick(1968)의 정규오자이브 모형을 그대로 확장하였다(박정, 2001). 피험자의 반응이 여러 개 이므로 분류하는 경계 기준 값을 여러 개로 확장하였다. 한 피험자의 문항변수 Γ_i가 한 문항에 대한 피험자의 반응경향을 나타내는 가상적인 연속변수라 하고, γ_k가 문항반응범주를 구별하는 기준이라 간주 할 때, 능력 수준 θ_j 를 가지는 피험자의 반응이 k번째 범주에 속할 확률은 γ_{k-1}과 γ_k사이에 있는 Γ_i에 대한 조건 분포의 면적, $P_k(\theta_j)$이다. 각 능력수준에서 각각의 범주에 속할 확률을 모두 합하면 $\sum_{k=1}^{m} P_k(\theta_j) = 1$이 되며, [그림 Ⅱ-1]과 같다.

능력수준의 $P_k(\theta_j)$를 연결하면 [그림 Ⅱ-2]와 같은 문항반응범주특성곡선(item response category characteristic curve: IRCC)이 된다.

Samejima(1969)는 이 곡선을 문항반응범주의 기능특성곡선(operating characteristic curve)이라고 했으며, 문항범주특성곡선(item category characteristic curve) 또는 문항범주흔적곡선(trace curve)이라고 부르기도 한다(박정, 2001).

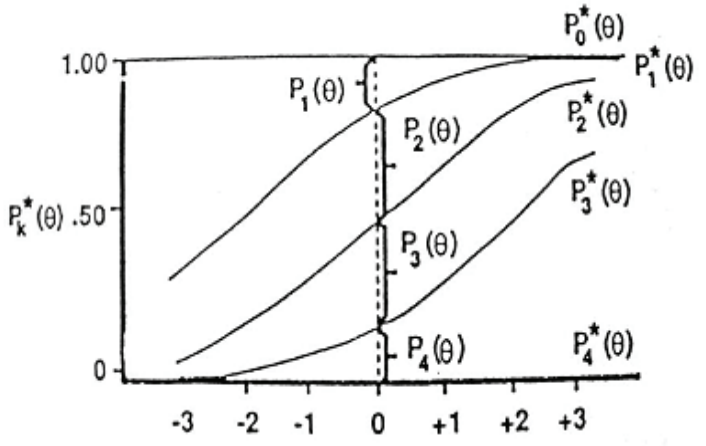
6) 2-모수 문항반응모형은 모든 문항의 추측도를 고려하지 않고 문항난이도와 문항변별도만 고려하는 문항이다. 모든 문항의 문항추측도는 0이고 각 문항들의 난이도와 변별도가 다를 것이라는 가정을 전제한다.



[그림 Ⅱ-2] 문항반응범주특성곡선

이분 문항으로 채점하여 3, 4범주에 속한 반응을 정답으로 하고, 1, 2범주에 속한 반응을 오답으로 하면, 정답률은 $P_3(\theta_j)+P_4(\theta_j)$이다. 또한 2, 3, 4범주에 속한 반응을 정답으로 하고 그 외의 반응을 오답으로 하면 이 때 정답률은 $P_2(\theta_j)+P_3(\theta_j)+P_4(\theta_j)$이다.

피험자의 반응을 이분적으로 채점하는 방법으로 정답률을 나타내면 곡선은 [그림 Ⅱ-3]과 같고 이것을 경계특성곡선(boundary characteristic curve)이라 하고, 각 문항반응범주를 구분하는 경계가 된다.



[그림 Ⅱ-3] 네 개의 범주를 가진 등급반응에 대한 경계특성곡선

경계특성곡선의 수리적 형태는 이분반응모형의 문항특성곡선의 수리적 형태와 같으며, k번째 범주와 k+1번째 범주를 구분하는 경계특성곡선인 $P_k^*(\theta_j)$와 피험자의 반응이 k번째 범주에 속할 확률인 $P_k(\theta_j)$사이에는 다음과 같은 관계가 성립한다(Samejima, 1969).

$$P_k{}^*(\theta_j) = \sum_{k+1}^{m} P_k(\theta_j)$$

$$P_0{}^*(\theta_j) = 1 = \sum_{k=1}^{m} P_K(\theta_j)$$

$$P_k(\theta_j) = P^*{}_{k-1}(\theta_j) - P_k{}^*(\theta_j) \qquad (Ⅱ.1)$$

정규오자이브와 로지스틱 계산법에 의해 변별도와 난이도를 모두 고려하는 2모수 모형으로 기술하면 다음과 같다. 정규오자이브 모형은 공식(Ⅱ.2), 로지스틱모형은 공식(Ⅱ.2)와 같다.

$$P_{ik}^{*}(\theta) = \int_{-\infty}^{a_i(\theta - b_{ik})}, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}) \qquad (\text{Ⅱ}.2)$$

$$P_{ik}^{*}(\theta) = \frac{\exp[-Da_i(\theta - b_{ik})]}{1 + \exp[-Da_i(\theta - b_{ik})]},\ P_{ik}^{*}(0) = 1,\ P_{ik}^{*}(k) = 0 \qquad (\text{Ⅱ}.3)$$

i : 문항

k : 범주

D : 1.72, 로지스틱 함수를 사용할 경우 정규오자이브 함수를 사용하는 경우에 가장 근접한 확률값을 갖도록 만드는 상수값

a_i : 문항 i에 대한 변별도

b_{ik} : 문항 i에서 k번째 범주난이도

경계특성곡선의 수리적 형태에는 정규 오자이브 모형과 로지스틱 모형이 있다. 정규 오자이브 모형과 로지스틱 모형이란 문항반응모형을 유도하기 위하여 함수를 이용한 수학적 문항반응모형을 말한다. 경계특성곡선의 속성은 변별도 모수치[7] a와 위치모수[8] b_k로 규명할 수 있다. 각 경계특성곡선은 변별도 모수를 공유하고 문항반응범주의 순위 상 위치모수는 서열이 있기 때문에 경계특성곡선들은 겹치지 않는다. 경계특성곡선은 문항반응범주확률의 누적이기 때문에 경계특성곡선의 위치모수들은 문항반응범주의 개수보다 하나 적으며 서열적이다. 다시 말하면, 모든 문항반응범주의 변별도는 동일하며, 가장 높은 범주의 위치모수가 가장 큰 값을 갖고 가장 낮은 범주의 위치모수가 가장 작은 값을 갖으며, 다음과 같은 관계가 있다.

---

7) 문항이 지닌 특성으로서, 문항이 피험자를 능력에 따라 얼마나 잘 변별하느냐 하는 정도를 나타내는 지수를 말한다. 문항반응이론에 의한 문항변별도는 문항특성곡선상의 문항난이도를 나타내는 점에서의 기울기를 말하며, a 혹은 α로 표기한다.

8) 문항이 능력수준의 어느 지점에서 기능하는가를 말한다.

$$b_m > b_{m-1} > \cdots b_k > b_2 > b_1$$

범주난이도의 위치는 문항범주특성곡선의 모양과 치우침을 결정하며, 일반적으로 점수들은 동일한 간격을 가정하고 있다. 경계특성곡선에 의해 추정되는 위치모수의 개수는 문항반응범주의 개수보다 하나가 작으므로 Samejima(1969)는 경계특성곡선의 정의에 기초하여 다음과 같은 방법으로 각 문항반응범주에서의 위치모수를 구체화하였다.

범주	위치모수
1	$b_{1'} = b_1$
2	$b_{2'} = (b_2 + b_1)/2$
3	$b_{2'} = (b_2 + b_1)/2$
.	.
.	.
k	$b_{k'} = (b_k + b_{k-1})/2$
.	.
.	.
m	$b_{m'} = b_{m-1}$

피험자가 선택 가능한 반응이 3개일 경우에 피험자의 반응 확률을 결정하도록 하는 경계 기준값은 2개(선택 가능한 반응지에서 1을 뺀 갯수)가 되는 형태를 말한다.

모든 문항반응범주에서 동일한 값을 갖는 문항변별도 모수 a는 이웃하는 문항반응범주에 속하는 피험자들을 변별하여 준다. 위치모수 bk는 문항반응범주에 속할 확률이 .5에 해당하는 능력척도 상의 점이며, 문항반응범주가 능력척도상의 어느 지점에서 가능한가를 나타낸다.

다분문항검사에서 문항에 대한 평가는 문항변별도와 경계특성곡선의 위치모수 또는 각 범주에 해당하는 위치모수에 의해 실시하며, 다분반응모형에서 문항범주난이도의 간격이 동일한 문항이 이상적이다(Baker, 1992).

GRM은 특별한 범주에서 수험자 반응에 대한 조건부 확률을 계산하도록 2단계 과정을 요구하는 사실에 의해 일반화부분점수모형(Generalized Partial Credit Model; GPCM)과 구별된다. 그 결과, 그것은 Embretson과 Reise(2000)에 의해 "간접적(indirect)" IRT 모형으로서 언급된다. 더욱이, GRM에 의해 가정된 득점과정은 개념적으로 PCM(부분점수모형)과 GPCM(일반화부분점수모형)에 의해 가정된 것과 다르다.

2. 일반화부분점수모형(Generalized partial credit model: GPCM)

Muraki(1992)는 부분점수모형에서 문항 변별도 지수를 사용할 수 있도록 하여 GPCM으로 확장하였다. GPCM은 GRM처럼 문항변별도(item discrimination)[9] 지수를 사용할 수 있으나 GRM처럼 범주난이도 지수가 서열화 되어 있다는 가정이 필요 없기 때문에 현실에서의 실용 가능성이 높은 모형이다.

모형의 수리적 전개는 부분점수모형(partial credit model: PCM)과 동일하고, 부분점수모형에 변별도 모수(a_i)를 추가시킨 형태이다. 즉, 부분점수모형은 Mater(1982)가 이분 반응 모형의 하나인 1-모수 모형(Rasch 모형)의 확률 원리를 그대로 다분문항반응 모형으로 확장한 것이다(박정, 2001). GPCM은 PCM과 같은 Rasch 계열의 모형이 제공하는 것보다 더 많은 정보를 제공하여 준다. 부분점수모형은 문항 i에서 h-1의 반응범주 대신 h의 반응범주를 선택할 확률이

9) 문항이 피험자를 능력에 따라 변별하는 정도를 나타내는 것으로, 문항특성곡선상의 문항난이도를 나타내는 점에서의 기울기를 말한다. 문항변별도는 α 또는 a로 표기하며, 문항변별도의 이론적 범위는 $-\infty$에서 $+\infty$를 지닌다.

로지스틱 이분반응모형에 의하여 결정된다는 가정하에 공식(Ⅱ.12)를 만든다.

$$C_{ih} = P_{ih|h-1,h}(\theta) = \frac{P_{ih}(\theta)}{P_{i,h-1}(\theta)+P_{ih}(\theta)} = \frac{\exp[Z_{ih}(\theta)]}{1+\exp[Z_{ih}(\theta)} \qquad (Ⅱ.12)$$

공식(Ⅱ.12)의 가정에 따라 P_{ih}는 공식(Ⅱ.13)으로 계산된다.

$$P_{ih}(\theta) = \frac{C_{ih}}{1-C_{ih}} P_{i,h-1}(\theta) = \exp[Z_{ih}(\theta)]P_{i,h-1}(\theta) \qquad (Ⅱ.13)$$

공식(Ⅱ.13)에서 $C_{ih}/(1-C_{ih})$는 h-1 범주 중에서 h범주를 선택하는 승산비이며, 이것의 로그변형인 $Z_{ih}(\theta)$는 로짓이라고 한다.

여기에서 각 $P_{ih}(\theta)$를 $\Sigma P_{ih}(\theta)=1$로 정규화하면 GPCM이 되고, GPCM은 문항변별도 모수(a_i)를 추가시킨 형태 이므로 문항 i의 k번째 범주에 반응할 확률은 공식(Ⅱ.14)의 가정하에 공식(Ⅱ.15)와 같이 계산된다.

$$P_{ik}(\theta) = P_{ik|k,k-1}(\theta) = \frac{P_{ik-1}(\theta)}{P_{ik}(\theta)+P_{ik-1}(\theta)} = \frac{\exp[Da_i(\theta-b_{ik})]}{1+\exp[Da_i(\theta-b_{ik})]} \qquad (Ⅱ.14)$$

$$P_{ik}(\theta) = \frac{\exp[\sum_{u=1}^{k} Da_i(\theta-b_{iu})]}{\sum_{c=1}^{m_i}\exp[\sum_{u=1}^{c} Da_i(\theta-b_{iu})]},\ k=1,2,...,m_i \qquad (Ⅱ.15)$$

GPCM에서 b_{ih}는 두 개의 선택지 중, 하나를 선택 하는 것을 표현하는데, 하나의 단계에서 다음 단계로 넘어가는 것을 의미하므로, 문항 범주난이도라고 부르기보다는 단계난이도라고 부른다(박정, 2001).

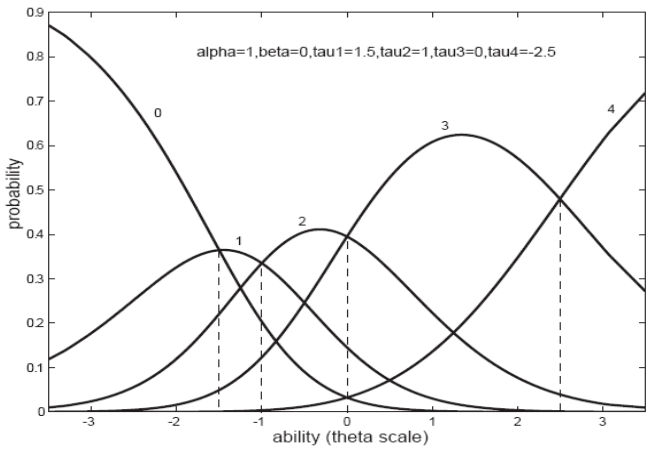
GPCM에서 문항변별도 즉, 모든 문항의 기울기를 1이라고 하면, 부분점수모형

이 되고, b_{ih}를 문항 위치 모수인 b_i와 문항 범주 모수인 d_h로 구분하여 $b_{ih}=b_i-d_h$로 표현하게 되면, GPCM은 평정척도모형이 된다. 평정척도모형의 원리를 부분점수모형에도 적용하고 변별도를 추가하여 GPCM으로 확장하였기에 여러 상황에 사용할 수 있는 일반적인 형태의 모형이다.

GPCM의 형성과정은 부분점수 모형의 형성과정과 같고, GRM처럼 문항 변별도 지수를 사용할 수 있으나, 등급반응 모형처럼 문항난이도의 서열화를 가정하지 않으며 문항을 제작할 때 범주난이도 지수에 대한 제한점을 가지지 않기 때문에 실용성이 높다. 일반화부분점수 모형에 의한 문항 i에 범주 x에 수험자 j의 점수에 의한 공식은 다음과 같다.

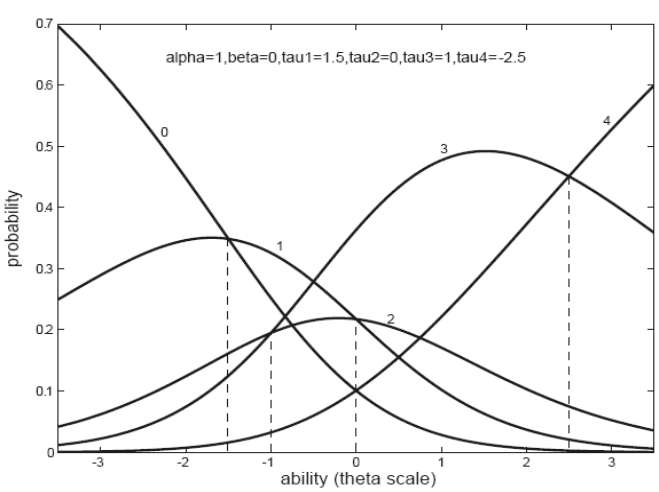
$$P(X_{ij}=x|\theta_j,\alpha_i,\beta_i,\tau_{ki})=\frac{\exp\sum_{k=0}^{x}\alpha_i[\theta_j-(\beta_i-\tau_{ki})]}{\sum_{y=0}^{m}\exp\sum_{k=0}^{y}\alpha_i[\theta_j-(\beta_i-\tau_{ki})]} \qquad (\text{Ⅱ}.16)$$

이때, j=1,...,N, i=1,...,T, 이고, x=0,...,m이다. 이 모형에서 α_i는 문항 i에 대한 변별도를 β_i는 문항 i의 난이도, τ_k는 문항 i에 대한 범주 k의 위치 모수가 된다.



[그림 Ⅱ-4] 5개의 범주를 가진 GPCM에 대한 반응범주곡선

[그림 Ⅱ-4]는 GPCM에 따른 5개의 범주(0, 1, 2, 3, 4)와 다분 문항의 반응범주곡선(category response curve)을 보여준다. [그림 Ⅱ-4]에서 $\alpha=1, \beta=0, \tau1=1.5, \tau2=1, \tau3=0, \tau4=-2.5$이다. $\beta_i-\tau_1$은 $\beta_i-\tau_4$를 통해서 반응범주곡선은 잠재-특성 척도에 교차되는 위치를 나타낸다. [그림 Ⅱ-5]는 $\alpha=1, \beta=0, \tau1=1.5, \tau2=0, \tau3=1, \tau4=-2.5$ 일 때, 반응범주곡선을 나타낸다.



[그림 Ⅱ-5] 5개의 범주를 가진 GPCM에 대한 반응범주곡선

3. 등급반응모형과 일반화부분점수모형의 비교

다양한 선행연구를 통하여 다분화된 반응 모형이 이분화된 모형에 비해 보다 객관적이고 신뢰로운 평가를 가능하게 해준다는 것을 알 수 있는데, Thissen, Nelson, Rosa, & McLeod(2001)의 연구에 따르면 다분문항점수는 많은 교육, 심리학 검사 프로그램에서 사용하고 있으며, 그 중 본 연구에서 사용된 등급반응모형과 일반화부분점수모형의 교육, 심리학적 분야에서의 사용은 대체로 등급반응모형은 심리검사에서, 일반화부분점수모형은 국가수준학업성취도검사 등에서 사용되고 있다.

등급반응모형은 특별한 범주에서 피험자 응답을 위한 조건확률을 계산하도록 요구 된다는 점에서 일반화부분점수모형과 구별되며, 등급반응모형과 일반화부분점수모형이 각각의 문항 적합도를 위해 같은 수의 모수가 필요하여도 두 모형의 함수 형태가 매우 다르기 때문에 동일한 모형의 복잡성을 가지고 있다고 말할 수 없다(Kang, 2006). 또한 본 연구와 관련하여 두 모형의 모형 적합도 관련 측면에서 살펴보면, 모형이 본 자료에 얼마나 잘 적합하느냐를 검증하는 모형 적합도를 비교하는 결과에서, 자료에 따라 더 적합한 모형에 대하여 상반되는 결과가 나오기도 한다. Cook, Dodd, & Fitzpatrick(1999)에 따르면 자료에 있어서 모형에 따른 적합도의 차이를 볼 수 있는데, 경험적 자료를 사용할 경우 부분점수모형이 더 적합한 것으로 나타났다. 반대로, 시뮬레이션 자료에 대해서는 등급반응모형과 일반화부분점수모형에서 더 높은 적합도를 보이는 것으로 나타났다.

De Ayala, Dodd & Koch(1992)의 연구에서 시뮬레이션 자료에 관한 연구를 살펴보면, 피험자 1,000명에 대한 컴퓨터능력적응검사(computer adaptive test)에서 등급반응모형과 부분점수모형에 의한 피험자 능력추정 결과 비교 연구에서도 등급반응모형이 자료를 더 잘 적합하는 것으로 나타났다. 이러한 선행연

구를 토대로 본 연구에서 시뮬레이션 자료를 사용하여 등급반응모형과 일반화 부분점수모형의 문항수준의 모형 적합도 지수와 전체 검사 수준의 모형 적합도 지수를 살펴보고 어떤 모형이 설명력에 있어서 좋은 결과를 보이는지 알아볼 필요가 있다.

Ⅲ. 연구방법

1. 시뮬레이션 연구의 설계

본 연구에서는, 우선 GRM과 GPCM으로 다분 문항 자료를 생성하였다. 즉, 자료 생성을 위한 모형(generating model, 이하 GM)으로 두 모형을 사용하였다. 시뮬레이션 연구를 위한 시뮬레이션 조건은 주로 검사길이와 표본크기가 사용되었는데, 두 개의 검사길이(I=10, 20 문항)와 두 개의 표본크기(N=500, 1,000)를 고려하였다. 다분 문항에서의 두 개의 검사 길이는 10개의 문항은 일반적인 적당한 검사 길이이고, 20개의 문항은 큰 수의 문항이다. 표본 크기에 있어서는 500명일 경우 작은 수를 나타내고, 1,000명의 경우 적당한 표본크기를 나타낸다(Kang, 2006). 정리하자면 본 시뮬레이션 연구의 조건은 총 8가지(2 GM × 2 검사길이 × 2 표본크기)이다.

각 조건에서 50개의 자료를 반복적으로 생성하였고, 각 자료에 대하여 GRM과 GPCM 두 모형을 각기 적용하여 문항 모수를 추정하였다. 즉, 자료 분석을 위하여 사용된 모형 (calibration model, 이하 CM) 역시 이들 두 모형이었다. 각 조건의 개별 자료에 대하여 두 모형을 적합한 결과를 바탕으로 문항 적합도 지수 $S\text{-}X^2$와 검사 전체 적합도 지수인 CS_1과 CS_2를 계산하여 그 결과를 상호 비교하고자 하였다.

2. 자료 생성

본 연구를 위한, 일반적인 방법으로 문항 반응 생성을 위해 사용하였다. 자

료 생성을 위한 단계를 살펴보면 다음과 같다(Kang & Chen, 2008).

(1)문항모수와 능력모수를 생성한다.

(2)선택한 IRT 모형(GRM, GPCM)에서, $P(z|\alpha_j, \beta_i, \tau_{ci})$의 확률을 계산하는데, 응답에 대한 문항과 능력 모수치를 계산한다.

(3)0의 경우 $u \leq P^*(0)$, z의 경우 $z = 1, 2, ...,$또는 Z_i에 대한 $P^*(z-1) < u \leq P^*(z)$로 할당 된, 난수 분포 U(0,1)로부터 그린 난수 u가 표시된다.

본 연구는, 모든 문항들 I에 대하여 반복적인 절차를 통하여 문항모수를 생성한다. 문항과 능력모수치와 GPCM, GRM에 하에서 응답 dataset를 생성한다.

3. 문항 모수의 추정

일차원적 및 다차원적 IRT 모형은 보다 복잡해지고, 모수 추정은 더욱 표준 추정 방법을 이용하여 구현시키기 어렵게 된다. 마르코프 연쇄 몬테칼로 방법(Markov chain Monte Carlo: MCMC)은 이러한 환경에서 추정을 매개변수화 하기 위해 대안적 접근법으로 제기되었다(Yen & Fitzpatrick, 2006).

이렇듯 본 연구에서는 MCMC 방법을 사용한 문항 및 피험자 모수 능력 추정법이 개발, 사용되었다. MCMC 방법은 접근이 다차원적, 복잡한 종속성을 가지고 있거나 다중수준구조를 반영하는 문항 응답을 포함할지라도 모든 복잡한 데이터 구성을 쉽게 수용한다(강태훈, 김동일, 2010). MCMC 방법의 장점 중 하나는 탄력적이라는데 있다. 접근이 다차원적 또는 복잡한 종속성을 가지고 있거

나, 다중수준구조를 반영한 문항 반응을 포함할지라도 모든 복잡한 자료 구성을 쉽게 수용한다(Yen & Fitzpatrick, 2006). MCMC 방법론은 일련의 모수를 위해 제안된 이론적 분포와 이론적 분포의 특성을 추론하기 위해 표본 정보를 이용하는 것으로부터 그려지는 표본을 포함한다.

MCMC 방법은 보통 베이시안 모형을 사용하여 실행된다(Patz & Junker, 1999b). 본 연구에서도 베이시안 방법을 적용하여, GRM과 GPCM의 모수를 추정하였다. 마찬가지로, GRM과 GPCM, 두 모형의 관련 모수 추정을 위하여 동일한 MCMC 방법을 사용함으로써 추정된 모수로 계산되는 문항선택지수나 문항적합도 지수 등을 이용하여 GRM과 GPCM 모형 모두를 공정하게 비교하도록 하였으며, GRM과 GPCM을 다루기 위하여 WinBUGS(Spiegelhalter, Thomas, Best, & Lunn, 2003)를 통하여 실시하였고, GRM과 GPCM을 위한 WinBUGS 코드는 부록1에 제시되어 있다.

4. 문항 수준의 모형 적합도 지수: $S\text{-}X^2$

IRT 모형 하에서 적합도에 관한 연구는 수 많은 IRT 문헌에서 보고 되었다(Bock, 1972; Douglas &Cohen, 2001; Glas &Suarez-Falcon, 2003; Liang &Wells, 2007; McKinley &Mills, 1985; Orlando &Thissen, 2000, 2003; Sinharay, 2003, 2005; Stone, 2000; Stone &Zhang, 2003; Suarez-Falcon &Glas, 2003; Wells, 2004; Yen, 1981).

선행 연구에 의하면, X^2, G^2은 검사의 길이와 피험자 집단의 크기에 매우 민감하게 반응하여 조건에 따라 문항 적합도를 올바르게 파악하는 수행력에 있어서 큰 차이가 나는 것으로 밝혀졌으나 이와 달리 $S-G^2$와 $S-X^2$은 문항의 수에 민감하게 영향을 받지 않으며, 안정된 수준의 경험적 제1종 오류

(empirical type I error rates)를 보이는 것으로 나타났다. 이를 통해 본 논문에서는 PIRT(polytomous item response theory)와 관련하여 가장 신뢰할 수 있는 문항적합도 분석 방법으로 알려진 $S-X^2$ 문항 모형 적합도 지수를 사용한다(Kang & Chen, 2008).

하나의 문항 i를 위한 적합도 지수 $S-X^2$를 구하기 위한 계산 공식은 다음 (Ⅲ.1)과 같다.

$$S-X^2=\sum_{K=Z_i z_{i=0}}^{F-Z_i}\sum^{Z_i} N_k\frac{(O_{ikz}-E_{ikz})^2}{E_{ikz}} \qquad (Ⅲ.1)$$

(Ⅲ.1)에서 F는 각 검사의 문항 범주가 $z=0,1,...,Z_i$일 때 검사의 최고 점수인 $\sum Z_i$를 뜻하며, O_{ikz}는 문항 i에서 k번째 검사 점수 집단에서 문항범주 z에 속하는 관찰된 반응빈도를, E_{ikz}는 같은 경우에 적합도를 검증하려는 문항반응모형을 통해 계산된 기대빈도를 의미한다.

$S-X^2$ 지수는 카이자승 검증을 통하여 문항의 적합도에 대한 영가설(H_o: 문항 자료가 해당 모형에 의해서 잘 설명 및 적합된다)을 검정하기 때문에 계산된 자유도에 비하여 그 값이 지나치게 클 경우 영가설을 부정하게 된다. 제1종 오류의 통제와 관련하여 얼마나 가깝게 영가설 하에서 문항적합도 지수가 표집분포가 이론적 분포에 있는지에 관해 고려되는데, 경험적 제1종 오류 및 통계적 검증력에서도 다른 문항적합도 지수에 비하여 매우 우수한 결과를 산출하는 것으로 알려졌다(Kang & Chen, 2008).

5. 전체 검사 수준의 모형 적합도 지수: CS_1과 CS_2

검사 실시 후 모든 학생들의 빈도분석을 통하여 검사점수 분포를 작성하면 최소점부터 최고점까지 각 점수를 획득한 학생들의 빈도 내지는 도수에 따라서 그 모양이 결정되며, 이를 '관찰된 검사점수 빈도분포=$f(X)$'로 볼 수 있을 것이다. 또한 자료 분석을 위해 사용된 특정문항반응모형이 주어진 자료를 잘 설명 및 적합한다는 전제 하에서 추정된 문항모수를 이용하여 검사점수의 빈도분포를 예측하는 것이 가능하며 이를 '기대된 검사점수 빈도분포=$F(X)$'로 볼 수 있을 것이다. 이를 위해서는 주어진 능력 추정치 값에서 가능한 점수빈도의 분포 즉, $F(X|\theta)$를 우선 구할 필요가 있으며 이 때 사용되는 공식이 잘 알려진 순환 공식(recursive formula)이다(Lord & Wingersky, 1984; Thissen, Pommerich, Billeaud & Williams, 1995). 일단 $F(X|\theta)$를 성공적으로 구하고 나면 이를 능력모수의 분포를 곱한 뒤에 적분하는 과정을 통하여, 문항반응모형에 의하여 기대된 검사점수 빈도분포 $F(X)$를 계산할 수 있다. 이 때 검사전체 수준에서의 적합도 검정을 위한 CS_1은 다음과 같은 공식을 통하여 계산할 수 있다. 이 때 통계적 검정을 위하여 사용되는 카이자승 분포의 자유도는 기본적으로 범위(=최고 검사점수-최저 검사점수)로서 계산하였으며, 만약 $F(X)$가 5보다 작을 경우에는 옆의 빈도에 통합(collapsing)되도록 하였으므로 추가적으로 자유도가 감소할 여지가 존재한다.

$$CS_1 = \Sigma \frac{[f(X) - F(X)]^2}{F(X)} \qquad (\text{Ⅲ}.2)$$

CS_2의 경우에는 앞서 밝힌 대로 카이자승 통계치의 가법성을 이용하는 것이다. 각 문항에 대하여 계산된 $S\text{-}X^2$값과 해당 자유도들을 각각 합하여 전체 검

사 수준에서의 카이자승 통계치와 자유도로 간주하였으며, 이러한 정보에 기반하여 전체 검사 수준에서 적용된 문항반응모형이 제대로 설명하는지를 알기 위하여 카이자승 검정을 실시하였다.

Ⅳ. 연구결과

1. 모형 적합도 지수: S-X^2

<표 Ⅳ-1> GPCM으로 자료 분석 시, $S-X^2$ 결과 ($\alpha = .05$)

GM	Test length (n)	Sample size (N)	부적합 판정 문항 비율
GPCM	10	500	0.09
		1000	0.05
	20	500	0.044
		1000	0.06
GRM	10	500	0.07
		1000	0.096
	20	500	0.034
		1000	0.067

<표 Ⅳ-1>의 결과를 보면, GPCM으로 생성된 자료를 GPCM에 적합하였을 때, 10문항, 500명의 문항수와 사례수가 적은 경우를 제외하고, 생성모형과 분석모형이 같은 경우 경험적 제1종 오류가 이론적 제1종 오류라고 할 수 있는 $\alpha = .05$에 가까운 값을 얻었다. 보다 구체적으로 살펴보면 생성모형이 GPCM일 때 이 자료를 같은 모형 즉, GPCM으로 분석하면 검사길이가 짧고(i.e., n=10) 표본 크기가 작은 경우(i.e., N=500)를 제외하고 경험적 제1종 오류가 0.05, 0.044, 그리고 0.06으로 유의수준에 매우 근접함을 볼 수 있었다. 그러나 생성모형과 분석모형이 다른 경우, 즉, GRM으로 생성 된 자료를 GPCM에 적합하였을 경우 $\alpha = .05$와 가깝지 않은 값이 산출되었다.

<표 Ⅳ-2> GRM으로 자료 분석 시, $S-X^2$ 결과($\alpha=.05$)

GM	Test length (n)	Sample size (N)	부적합 판정 문항 비율
GPCM	10	500	0.104
		1000	0.128
	20	500	0.069
		1000	0.084
GRM	10	500	0.088
		1000	0.058
	20	500	0.051
		1000	0.062

<표 Ⅳ-2>의 결과 역시, GRM으로 생성 된 자료를 GRM으로 적합할 때 $\alpha=.05$에서 가까운 값이 나옴을 알 수 있다. 반면에 GPCM으로 생성된 자료를 다른 모형인 GRM으로 분석 시, $\alpha=.05$에서 상대적으로 먼 값이 나온 것을 발견할 수 있었다. 구체적으로 보면, 이론적 제1종오류 라고 할 수 있는 $\alpha=.05$에 가까운 값을 얻었다. 보다 자세히 살펴보면, 생성모형이 GRM일 때 이 자료를 같은 모형 즉, GRM으로 분석하면 검사길이가 짧고(i.e., n=10) 표본 크기가 작은 경우(i.e., N=500)를 제외하고 경험적 제1종 오류가 0.058, 0.051, 그리고 0.062로 유의수준에 매우 근접함을 볼 수 있었다. 또한 <표 Ⅳ-1>과 <표 Ⅳ-2>의 값을 비교해 보면, GRM으로 생성된 자료를 GPCM으로 분석한 경우에 비해 GPCM으로 생성된 자료를 GRM으로 분석했을 경우에 비하여 제1종 오류가 더 잘 통제되고 있음을 즉, 경험적 제1종 오류가 유의수준과 가깝게 산출되고 있음을 볼 수 있었다. 이러한 결과가 함의하는 것은, 두 모형 모두 다른 모형으로 산출된 자료에 적용될 경우, 그 설명 및 적합도 측면에서 GPCM이 GRM 보다 좀 더 나은 모형일 수 있다는 점이다.

<표 Ⅳ-3> $S-X^2$ 결과 분석 예시 :
10문항, 500명, 50개 dataset 중 하나의 경우
CM=GPCM, GM=GPCM

문항	$S-X^2$	자유도	p-value
1	28.064	23	0.213
2	24.387	28	0.661
3	15.277	21	0.809
4	22.996	18	0.191
5	34.331	26	0.127
6	18.403	20	0.561
7	18.301	19	0.502
8	41.417	27	0.038*
9	13.67	27	0.984
10	34.182	23	0.063

* p-value < 0.05

<표 Ⅳ-3>은 생성모형과 분석모형 모두 GPCM 일 때, $S-X^2$결과 중 10개의 문항, 500명, 50개의 자료 중 한 경우의 분석 예시를 보여준다. 위의 표에서 볼 수 있는 바와 같이 이 자료에서는 1번에서 10번 문항의 $S-X^2$ 분석 결과 중 8번 문항에서 유의수준 .05에서 GPCM에 의해 제대로 적합 되지 않는 것으로 검정되었다. 각각의 자유도는 기대빈도가 1을 넘지 않는 경우가 발생하지 않도록 인접 점수 집단을 통합하는 정도에 따라서 달리 계산되었고 이러한 이유로 각 문항마다 그 값이 다를 수 있다.

2. 전체 검사 수준의 모형 적합도 지수: CS_1과 CS_2

<표 Ⅳ-4> GPCM으로 자료 분석 시, CS_1 결과($\alpha = .05$)

GM	Test length (n)	Sample size (N)	경험적 제1종오류
GPCM	10	500	0.16
		1000	0.34
	20	500	0.16
		1000	0.28
GRM	10	500	0.36
		1000	0.26
	20	500	0.12
		1000	0.36

<표 Ⅳ-5> GRM으로 자료 분석 시, CS_1 결과($\alpha = .05$)

GM	Test length (n)	Sample size (N)	경험적 제1종오류
GPCM	10	500	0.62
		1000	0.8
	20	500	0.9
		1000	0.94
GRM	10	500	0.72
		1000	0.94
	20	500	0.9
		1000	0.94

<표 Ⅳ-4>와 <표 Ⅳ-5>는 GPCM과 GRM으로 분석 시의 CS_1의 결과를 보여준다. 즉 시뮬레이션 조건에서 자료의 수가 50개이므로 전체 검사 수준에서 몇 개의 자료가 부적합으로 산출되었는지를 계산하고 이를 50으로 나눈 것을 경험적 제1종 오류라고 규정하였다. 사례수가 500인 경우, $\alpha = .05$에서 .05에 가깝게 나왔지만, 사례수가 1,000인 경우는 그 값이 크게 나옴을 알 수 있다. 즉, 경험적 제1종 오류의 값이 .05와 매우 다른 값으로 나왔는데, 이는 시뮬레이션 자료의 특성상 자료의 문제라기보다는 CS_1이라는 적합도 지수 자체가 적합도 검정이라는 측면에서 어떤 문제를 가지고 있는 것으로 생각되었다.

이러한 문제점들을 무시하고 앞의 표를 해석한다면, 경험적 제1종오류가 GRM, GPCM 그 어느 모형으로 분석하여도 비슷하기 때문에, Maydu-Olivares, Drasgow, & Mead(1994)의 연구에서와 같이 어떤 모형을 선택하여 사용하여도 그 적합도에 차이가 없다는 주장을 받아들일 수도 있을 것이다. 하지만, CS_1 자체를 신뢰할 수 없을 정도의 좋지 않은 결과가 산출되었기 때문에 이러한 결론은 받아들이기 어려워 보였다.

반면, 아래의 <표 Ⅳ-6>과 <표 Ⅳ-7>의 CS_2의 결과를 살펴보면, CS_1 보다는 경험적 제1종 오류 결과에 있어서 보다 나은 모형 적합도 지수라고 할 수 있다. 또한, CS_2의 결과에서도 앞서 <표 Ⅳ-1>과 <표 Ⅳ-2>에서 본 것과 같이 GPCM으로 자료를 분석 시 GRM보다 경험적 제1종류가 유의수준 .05와는 상대적으로 가까운 값으로 산출되었다.

<표 Ⅳ-6> GPCM으로 자료 분석 시, CS_2 결과($\alpha = .05$)

GM	Test length (n)	Sample size (N)	경험적 제1종오류
GPCM	10	500	0.2
		1000	0.14
	20	500	0.1
		1000	0.16
GRM	10	500	0.14
		1000	0.16
	20	500	0.06
		1000	0.16

<표 Ⅳ-7> GRM으로 자료 분석 시, CS_2 결과($\alpha = .05$)

GM	Test length (n)	Sample size (N)	경험적 제1종오류
GPCM	10	500	0.28
		1000	0.4
	20	500	0.26
		1000	0.26
GRM	10	500	0.24
		1000	0.2
	20	500	0.1
		1000	0.16

Ⅴ. 논의 및 결론

본 연구에서는 GRM과 GPCM을 소개하고 시뮬레이션연구를 통하여 두 모형 간의 모형 적합도에 있어서의 차이점을 살펴보았다. 각 모형을 이용하여 다양한 수준의 조건에서 자료들을 생성하고, 동일 자료에 대하여 두 모형 모두를 이용하여 분석하여 문항적합도지수를 사용해서 주어진 자료를 얼마나 잘 설명할 수 있는지를 살펴보았다. 또한, 전체 검사 수준에서의 모형-자료 적합도(model-data overall fit)를 구하여, 두 모형이 자료의 적합에 있어서 유의미한 차이가 있는지 살펴보고, 각 자료에 대하여 관찰된 원점수 분포와 적용된 모형에 의하여 재생산된(reproduced) 원점수 분포를 비교하여 CS_1의 값과 $S-X^2$를 활용한 CS_2를 살펴보았다.

모형 적합도 지수($S-X^2$)를 통해 살펴 본 결과 문항 적합도 수준에서 생성모형과 분석모형이 동일할 때 경험적 제1종오류가 유의수준 .05와 비슷하게 나타났으며, GRM보다 GPCM이 생성모형과 분석모형이 다를지라도 보다 더 주어진 자료를 잘 설명하는 것으로 나타났다. 하지만, GRM으로 생성 된 자료를 GPCM으로 분석할 때에도 여전히 경험적 제1종 오류가 $\alpha = .05$과는 상당히 다른 값으로 나와 어떤 자료든 GPCM을 사용할 수 있다는 결론을 유도하기는 어려웠다. 즉, 주어진 자료에 가장 적합하고 타당한 모형을 골라 적용하려는 노력이 여전히 요구된다고 볼 수 있다.

반면, CS_1과 CS_2를 통한 시사점을 살펴보면 다음과 같다. 첫째, 관찰된 검사점수 빈도분포와 적용된 모형에 의해 재생산된, 혹은 기대된 검사점수 빈도분포를 비교하여 카이자승 검정을 실시한 결과 제1종 오류의 통제라는 측면에서 좋지 않은 결과가 산출되었다. 이를 통하여 CS_1 통계 지표의 개선이 필요하며, 제대로 작동 되지 않는 이유에 대한 추후 연구가 필요하다고 생각되었다.

둘째, 본 연구에서는 각 조건에서 50개의 자료를 사용하였는데 문항 수준의 연구를 위해서는 꽤 많은 수의 문항들을 다룰 수 있기 때문에 별다른 문제가 없었으나 전체 검사 수준의 적합도를 살펴보고 경험적 제1종 오류를 계산하기 위해서는 그 수가 충분하지 않았던 것으로 생각된다. 따라서 보다 일반화된 결과를 얻기 위해서는 그 이상의 자료 수 즉, 100 혹은 1,000개 정도의 반복 자료를 고려할 필요가 있다. 세 번째는 CS_2의 결과가 .05와 비슷하지는 않으나 CS_1의 결과 보다는 더 합리적인 결과를 도출하였다는 점이다. 두 적합도 중 하나를 택하여 추가적 개선 연구를 실시한다면 CS_1보다는 CS_2에 집중할 필요가 있을 것이다. 마지막으로, 네 번째는 문항 적합도 지수와 살펴보았을 때 GPCM이 GRM보다 상대적으로 주어진 자료가 어떤 모형으로 생성되었는가에 관계없이 보다 설명력 및 적합도가 더 높았다고 말할 수 있다는 점이다. 따라서 보다 나은 모형을 위한 연구를 충분히 할 수 없는 상황 속에서 두 모형 중 하나를 택해야 한다면 GPCM을 택하는 것이 보다 합리적인 선택이 될 수 있는 것으로 보인다.

위와 같은 본 연구의 시사점 및 개선점에 비추어서 앞으로 추가적으로 연구를 할 때 고려해야 할 사항들을 열거하면 다음과 같다.

첫째, 전체 검사 수준에서 적합도를 보기에는 본 연구에서 사용한 조건당 50개의 반복 자료가 부족한 것으로 보였다. Kang, Cohen, & Sung(2009)의 연구에서도 자료의 수가 클수록 적합도 지수에 있어서 일관되게 좋은 연구 결과가 나온다고 보고되었다. 이렇듯 향후 관련 시뮬레이션 연구에서는 조건 당 보다 많은 자료, 100 혹은 1,000개 정도의 자료를 생성하여 사용할 필요가 있다.

둘째, CS_1이 이론적으로는 별다른 하자가 없는 것으로 보였으나 실제 시뮬레이션 결과 많은 문제를 나타내었으므로 이 적합도 지수를 사용함에 있어서 주의가 요구된다. 이 경우, 보다 단순한 문항반응모형을 적용하여 그 통계적 속성을 차근차근 살피는 접근이나 혹은 여러 가지 계산상의 개선점을 도출하는

과정 등이 필요할 것이다.

셋째, CS_2의 문항 적합도는 $S-X^2$의 결과와 유사하나 경험적 제1종 오류가 .05 수준과는 조금 차이가 나는 것으로 나타났기 때문에 그 원인을 밝히는 작업 등과 함께 개선이 필요함을 알 수 있었다. 이를 위하여 기존의 우도비 검증이나 G^2값 등과 같은 전통적 적합도 지수와 계산 과정 및 결과를 체계적으로 비교하는 연구를 수행하여 시사점을 얻을 필요가 있다.

참 고 문 헌

강태훈(2010). 정의적 특성 검사 자료 분석을 위한 문항반응모형 탐색: 일반화등급전개모형과 일반화부분점수모형의 적용을 중심으로. **교육평가연구**, **23**(1), 149-170.

강태훈(2011). 문항모수 추정치에 기반한 문항 및 검사 신뢰도의 추정 : 혼합유형 문항검사 자료를 중심으로. **아시아교육연구**, **12**(1), 119-140.

구슬기(2011). 일반화부분점수모형에 의한 대인관계능력 진단도구의 문항 특성분석. 석사학위논문, 이화여자대학교.

김경아(1999). 다분문항반응모형에 의한 자아개념 척도의 양호도 분석. 석사학위논문, 숙명여자대학교.

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

김보연(2005). 다분문항반응이론의 등급반응모형에 의한 초등학생용 자아개념 진단검사의 양호도 분석. 석사학위논문, 대구교육대학교.

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

김성훈(2008). 등급반응모형을 위한 검사특성곡선 방법 및 적률 방법의 척도변환계수의 표준오차. **교육평가연구**, **21**(1), 227-247.

김주학, 이기봉(2000). 스텝검사에서 부분점수모형과 일반화부분점수모형에 의해 추정된 모수의 비교분석. **체육학연구**, **39**(1), 679-689.

박 정(1999). 검사의 길이, 반응범주의 개수, 피험자의수 및 피험자 능력분포에 따른 다분문항반응이론 모형의 문항모수 추정치의 정확도. **교육평가연구**, **12**(1), 17-42.

박 정(2001). **다분 문항반응이론 모형**. 서울: 교육과학사.

박찬호, 강태훈(2011). 전문가 판정에 의한 차등 배점을 활용한 제한적 일반화부분점수 모형의 적용. **교육평가연구**, **24**(3), 781-797.

백순근(1995). 부분점수모형을 이용한 다단계형 자료의 분석 연구. **교육평가연구**, **12**(1), 59-75.

백순근, 채선희(1998). **컴퓨터를 이용한 개별적응검사: 교육 및 심리검사를**

위한 새로운 방법. 서울: 원미사.
백순근, 김혜숙(2004). 부분점수모형에 근거한 적합도의 활용 가능성에 대한 탐색-교사의 전문적 판단과의 상관-. **교육평가연구**, **17**(1), 103-120.
설현수(2000). 차별문항기능에 대한 IRT에 기초한 문항적합도 지수의 민감도 분석. **한국교육문제연구소**, **15**, 93-109.
성태제(1989). 사전능력분석의 특성에 따라 문항과 능력의 모수치를 추정하는 주변최대우도추정법의 민감도. **교육평가연구**, **3**(1), 87-117.
성태제(1998). 다분문항반응이론(등급반응모형)에 의한 학구적 실패내성척도의 문항분석과 피험자 특성추정. **교육평가연구**, **12**(2), 203-218.
성태제(2009). **문항반응이론의 이해와 적용**. 서울: 교육과학사.
송미영(1994). 이분반응모형과 등급반응모형에 의한 문항특성과 피험자 능력 모수 추정의 정확성. **교육평가연구**, **7**(2), 241-261.
임미경(2001). 등급반응모형, 평정척도모형, 부분점수모형의 문항모수와 피험자모수 추정치 비교분석. 석사학위논문, 이화여자대학교.
지은림, 채선희 (2000). **Rasch 모형의 이론과 실제**. 서울: 교육과학사.
한국교육평가학회 (2004). **교육평가 용어사전**. 서울: 학지사.
Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.
Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.
Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Psychological Measurement*, *1*(4), 363-378.
Baker, F. B. (1997). Estimation of graded response model parameters using multilog. *Applied Psychological Measurement*, *21*(1), 89-90.
Baker, F. B. & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
Bock, R. D. & Lieerman, M. (1970). Fitting a response model for dichotomously scored tems. *Psychometrika*, *35*, 179-197.
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. & Aitkin, M. (1981). Marginalmaximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443-459.

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.

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3*(1), 1-20.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, 28(3), 165-185.

De Ayala, R. J., Dodd, B. G., & Koach, W. R. (1992). A comparison of the partial credit and graded response models in computerized anaptive testing. *Applied Measurement in Education*, *15*(1), 17-34.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement*, *65*, 42-50.

Dodd, B. G. & Koch, W. R. (1987). Effectsofvariationsin item step valueson item and testinformationin thepartialcreditmodel. *Applied Measurement in Education*, *11*(4), 17-34.

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

onoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of educational measurement*, *31*(4), 295.

Glas, C. A. W. (1995). Testing the generalized partial credit model. *Objective Measurement*, *4*, 237-260.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Hambleton, R. K. & Swaminathan, H. & Jane Rogers, H. (1991). *Fundamentals of item response theory*. Newbury: SAGE.

Johnson, M. A. (2006). An investigation of stratification exposure

control procedures in CATs using the generalized partial credit model. Unpublished doctoral dissertation, University of Texas at Austin.

Kang, T. H. (2006). Model selection methods for unidimensional and multidimensional IRT models. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Kang, T. H. & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Kang, T. H. & Chen, T. T. (2008). performance of the generalized $S-X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391-406.

Kang, T. H., Cohen, A.S., & Sung, H.-J. (2009). Model Selection Indices for Polytomous Items. *Applied Psychological Measurement*, *33*. 499-518.

Kendall, M. G. & Stuart, A. (1973). *The Advanced Theory of Statistics*, vol.2. New York: Hanfner Publishing Company

Koch, W. R. (1983). Likert scaling using the graded response model. *Applied Psychological Measurement*, *7*, 15-32

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodnes-of-fit statistics. *Journal of the American Statistical Association*, *73*, 253-263.

Lee, K. J. (1992). Quasi-likelihood ratio test and related statistical computations. Unpublished doctoral dissertation, University of Brunswick.

Lee, K. O. (1995). Application of the graeded response model to the revised tennessee self-concept scale: unidimensionality, parameter invariance, and differential item functioning. Unpublished doctoral dissertation, University of Southern california.

Lord, F. M. (1997). Practical application of item characteristic curve theory. *Journal of Educational Measurement*, *14*, 177-138.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score

and equipercentile observed-score “equatings." *Applied Psychological Measurement*, *8*, 453-461.

Master, G. N. (1982). A rascha model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics, *Applied Psychological Measurement*, *9*, 49-57.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices, *Psychometrika*, *55*, 75-106.

Muraki, E . (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71.

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

Muraki, E . (1993). Information function of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351-363.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

rlando, M., & Thissen, D. (2003). Further investigation of the performance of $S-X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Pastor, D. A., Dodd, B. G. & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26*(2), 147-163.

Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement*, *32*, 407-423.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*(4).

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement*, *60*, 974-991.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352.

Thissen, 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.

Thissen, D., Pommerich, M., Billeaued, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39-49.

von Davier, M. (2004). Parially observed mixtures of IRT models: an extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*(6), 389-406.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. Brennan, L.L.(EDT.) *Educational Measurement*. Connecticut: Greenwood Publishing Group.

ABSTRACT

A simulation study: a comparison of fit in generalized partial credit model versus graded response model

Kim, Myung Yeon
Department of Edutation
The Graduate School
Sungshin Women's University

This study set out to investigate differences, if any, in fit or power to practically explain data between GPCM and GRM, which are the two models usually used to analyze the results of academic achievement tests or psychological tests comprised of items to be scored polytomously. For that purpose, the investigator examined differences in model fit between the two models through simulations. The two models were used to generate data under conditions of various levels and analyze the same data. Based on the results, the following research was conducted:

1. The $S-X^2$ index was employed, which is regarded as the most excellent item fit index in terms of control of Type 1 error

and statistical power in recent years, to examine if each model would record better fit for their own data at the item level than other models.

2. The model-data overall fit was obtained at the overall test level to see whether there were any significant differences in data fit between the two models. The raw scores distribution observed in each data was compared with the reproduced raw scores distribution by model application to obtain the CS_1 value and CS_2 using $S-X^2$.

The investigator also produced polytomous items data with GRM and GPCM. Considering while setting simulation conditions for the simulation study were the two test lengths(I=10 and 20) and two sample sizes(N=500 and 1,000). In each condition, 50 data were repeatedly generated, to which each of the two models was applied to estimate item parameters. Based on the two models' fit results, CS_1 and CS_2, which were the overall fit indexes, along with a test $S-X^2$ were calculated for individual data in each condition. The results were then compared with one another.

The research findings can be summarized as follows:

1. According to the $S-X^2$ results, the experiential Type 1 error was similar to a significance level of .05 when the production model was the same as the analysis model at the

item fit level.

2. According to the results of a chi square test using CS_1 and CS_2, the results were very bad in terms of control of Type 1 error, which suggests a need to improve the CS_1 statistical index.
3. The CS_2 results were not similar to .05 but more rational that the CS_1 results.
4. When considering the item fit index, GPCM had more explanatory power or fit than GRM regardless of which model was used to generate data.

Key words: Graded Response Model, Generalized Partial Credit Model, Type 1 error, $S-X^2$, CS_1, CS_2

부 록

부 록 1 : GPCM과 GRM을 위한 WinBUGS 코드

부 록 2 : S-X^2 계산을 위한 MATLAB 코드

부 록 3 : CS_1과 CS_2 계산을 위한 MATLAB 코드

부록 1: GPCM과 GRM을 위한 WinBUGS 코드

- GPCM을 위한 WinBUGS 코드

```
-----------------------------------------------------
# WinBUGS code for calibrating Generalized Partical Credit Model

model
{
  for (j in 1:N) {
   for (i in 1:T) {
          r[j,i]<-resp[j,i];
}}

for (j in 1:N){
    for (i in 1:T) {
            denom[j,i,1] <- 1;
            numer[j,i,1] <- 0;
            enumer[j,i,1] <- 1;
    }}

    # GPCM
    for (j in 1:N) {
        for (i in 1:T) {
            for (k in 2:mI[i]) {
                numer[j,i,k] <- a[i]*(theta[j] - b[i] + tau[i,k]   ) +
numer[j,i,k-1];
                enumer[j,i,k] <- exp(numer[j,i,k]);
                denom[j,i,k] <- enumer[j,i,k] + denom[j,i,k-1];
            }
                denom2[j,i,1] <- denom[j,i,mI[i]];
      }
  }

    for (j in 1:N) {
        for (i in 1:T) {
            for (k in 1:mI[i]){
                p[j,i,k] <- enumer[j,i,k]/denom2[j,i,1];
                r[j,i] ~ dcat(p[j,i,1:mI[i]]);
        }
```

```
    }
  theta[j] ~ dnorm(mu,1);
  }
mu ~ dnorm(0,1);

# Priors
    # item discrimination
    for (i in 1:T) {
         a[i] ~ dlnorm(0.,1.) ;      }

    # item difficulty
    #   for (i in 1:T){
    #         b.pre[i]~dnorm(0.,1.) ;      }
    #    for (i in 1:T){
    #         b[i] <- b.pre[i]-mean(b.pre[1:T]) ;      }
        for (i in 1:T){
            b[i]~dnorm(0.,1.);     }

    # The first Category of every item
    for (i in 1:T) {
         tau[i,1] <- 0;     }

    # From the second to (ml-1)th Categories of each item
    for (i in 1:T){
         for (k in 2:(mI[i]-1)) {
            tau[i,k]~dnorm(0,.1);         }}

    # The last Catetory of every item: It makes the sum of all catetory
parameters be 0
      for (i in 1:T){
            tau[i,mI[i]] <- -sum(tau[i,2:(mI[i]-1)]);        }

# If you are interested in parameters about the categories without item
difficulty
# for (i in 1:T) {
#       step[i,1] <- 0;          }
# for (i in 1:T){
#   for (k in 2:mI[i]){
#      step[i,k] <-  b[i] - tau[i,k] ;       }}
}
```

- GRM을 위한 WinBUGS 코드

```
-----------------------------------------------
# Graded Response Model

model
{

    for (j in 1:N)  {
        for (i in 1:T)  {
        r[j,i]<-resp[j,i]
}}

# GRM
    for (j in 1:N)  {
        for (i in 1:T)  {
            for (k in 1: (mI[i]-1)) {
            p[j,i,k] <- 1 / (1+exp(-a[i]*(theta[j]-b[i,k])));
    }}}

    for (j in 1:N)  {
        for (i in 1:T)  {
        pcat[j,i,1] <- 1-p[j,i,1];
            for (k in 2: (mI[i]-1)) {
            pcat[j,i,k] <- p[j,i,k-1]-p[j,i,k];
            }
        pcat[j,i,mI[i]] <- p[j,i,(mI[i]-1)];
    }}

    for (j in 1:N)  {
        for (i in 1:T)  {
            for (k in 1:mI[i]) {
            pc[j,i,k] <- pcat[j,i,k]  / sum( pcat[j,i, 1:mI[i]] ) ;
            }
        r[j,i] ~ dcat(pc[j,i,1:mI[i]]);
        }
    theta[j] ~ dnorm(mu,1);
```

```
    }
mu ~ dnorm(0,1);

#  Priors
    for (i in 1:T) {
        a[i] ~ dlnorm(0, 1.);
        b[i,1] ~ dnorm(0,.1);
        for (k in 2: (mI[i]-1)) {
            b[i,k] ~ dnorm(0, .1) I(b[i,k-1], );
            }
       }

# code below is very useful because they fix the mean of item
difficulty as zero
# But, when you use this, there can be 'trapping' often
# So, I didn't use this for this study

#    for (i in 1:T) {
#       a[i] ~ dlnorm(0, 1.);
#       b.pre[i,1] ~ dnorm(0,.1);
#       for (k in 2: (mI[i]-1)) {
#          b.pre[i,k] ~ dnorm(0, .1) I(b[i,k-1], );
#         }
#    }
#    for (i in 1:T){
#        for (k in 1: (mI[i]-1)) {
#               b[i,k] <- b.pre[i,k]-mean(b.pre[1:T,1:(mI[i]-1)]) ;
  } }
}
-------------------------------------------------
```

부록 2 : S-X^2 계산을 위한 MATLAB 코드

- GM=GPCM, CM=GPCM, 검사 문항=20, 피험자 수=1,000인 경우

gpgp_SX2_22.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Codes for item fit analysis using s-X^2
%%% of Polytomous items (GPCM)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear all;  % CONDITION 22, GPGP

% SX2 for GPCM
I=20; N=1000; nd = 50;
% Item Parameter
load 'c:/PROJ2011/KMY/parabu/gpgp2212all.txt'; itemest=gpgp2212all;
% ability distribution
k=-4:.2:4; den = normpdf(k,0,1); den=den/sum(den); J=length(k);
abilpara = ones(J,2); abilpara(:,1) = k'; abilpara(:,2) = den';
gquad=abilpara(:,1); densi = abilpara(:,2);
m=3; % Calibration Model = GPCM: 1 a, 1 b and 1 taus, # of parameters per item
% 123 data -> 012 data
nc=3; C=2; nos=I*C+1; x=[0:I*C]; NG=I*C-2*C+1;
TNG = I*C-1; % total number of groups that exclude zero(0) and full(I*C) scores.
nl=2*I + C*I; % number of lines per dataset
% original DF (before being midified with cell-collapsing)
oriDF=NG*C-m;

% result collects:
resultSX2=zeros(I*nd,7); resultSX2(:,1)=2212;

for v = 1:nd;
    itempara=zeros(nl,1); itempara = itemest(v*nl-(nl-1):v*nl);
    a=itempara(1:I); b=itempara(I+1:I*2); tau = zeros(I,C);
    for u=1:I
       tau(u,1:C) = itempara(2*I+u*C-(C-1):2*I+u*C)';
    end
    itemesti=zeros(I,C+2); itemesti(:,1)=a; itemesti(:,2)=b; itemesti(:,3:(C+2))=tau;
    % read data
    dfilen =  ['c:/proj2011/KMY/gp_ori/2212/gp2212c', int2str(v), '.dat'];
    datc = load (dfilen); totdat = datc - 1; % to make 012 data with 123 data, missing is now "8"
    % item index
    ti=(1:I); ii=zeros(I,I-1);
    ii(1,:)=ti(2:I); ii(I,:)=ti(1:I-1);
    for j=2:I-1
        ii(j,:)=ti([1:(j-1) (j+1):I]);
    end
    % calculating P_jiw for each item's category
    pofc=zeros(I,nc,J);
    pofc=pofcateGPCM3(abilpara,itemesti);  % P(item, category, examinee) = P_iwj
    % Conditional Distribution of Number-Correct Raw score for each theta value , Pr(X=x|theta)
    % Recursion Formula
    allitemSk=ones(J,nos); % Sk(theta)
    for j=1:J
        allitemSk(j,:)=recursive3(pofc(:,:,j));
    end
    % each item E_k,w,i: group, category, item
    % homogeneous groups: from C to (I-1)C: (I-1)*C-C+1=NG
    % categories: 0, 1,2,...C
    Ekwi=zeros(TNG,C+1,I);
    for i=1:I
        ipofc=zeros(I-1,nc,J); % except item i.  P(item, category, examinee) = P_iwj
        ipofc=pofcateGPCM3(abilpara,itemesti(ii(i,:),:));  % K thetas and I-1 items (no item i was
```

```
included)
        % Recursion Formula
        Skwoi=ones(J,nos-C); % Sk(theta) without item i
        for j=1:J
            Skwoi(j,:)=recursive3(ipofc(:,:,j));
        end
        % C+1 numerators and 1 denominators for each group: #groups by C+1 or 1
        nEkwi=zeros(TNG,C+1); dEkwi=zeros(TNG,1);

        % low extreme scores from 1 to (C-1)
        for gr=1:(C-1)
            for j=1:J
                for imC=1:gr+1   % temporary valid category numbers
                    nEkwi(gr,imC) = nEkwi(gr,imC) + pofc(i,imC,j)*Skwoi(j,gr+1-(imC-1))*densi(j);
                end
                dEkwi(gr)   = dEkwi(gr)    + allitemSk(j,gr+1)*densi(j);
            end
        end
        % middle scores from C to (I-1)*C
        for gr=C:(I-1)*C
            for j=1:J
                nEkwi(gr,1) = nEkwi(gr,1) + pofc(i,1,j)*Skwoi(j,gr+1-0)*densi(j);
                nEkwi(gr,2) = nEkwi(gr,2) + pofc(i,2,j)*Skwoi(j,gr+1-1)*densi(j);
                nEkwi(gr,3) = nEkwi(gr,3) + pofc(i,3,j)*Skwoi(j,gr+1-2)*densi(j);
                %nEkwi(gr,4) = nEkwi(gr,4) + pofc(i,4,j)*Skwoi(j,gr+1-3)*densi(j);
                dEkwi(gr)   = dEkwi(gr)    + allitemSk(j,gr+1)*densi(j);
            end
        end
        % high exreme scores from (I-1)*C+1 to (I*C-1)
         for gr=(I-1)*C+1:(I*C-1)
            for j=1:J
                NVC= I*C-gr+1;  % number of valid category numbers for each high score group
                for imC=(C+1)-(NVC-1):(C+1);   % temporary valid category numbers
                    nEkwi(gr,imC) = nEkwi(gr,imC) + pofc(i,imC,j)*Skwoi(j,gr+1-(imC-1))*densi(j);
                end
                dEkwi(gr)   = dEkwi(gr)    + allitemSk(j,gr+1)*densi(j);
            end
        end
        Ekwi(:,1,i) = nEkwi(:,1) ./dEkwi(:);
        Ekwi(:,2,i) = nEkwi(:,2) ./dEkwi(:);
        Ekwi(:,3,i) = nEkwi(:,3) ./dEkwi(:);
        %Ekwi(:,4,i) = nEkwi(:,4) ./dEkwi(:);
    end

    %plot(C:(I-1)*C,Ekwi(:,1,1)); hold on;
    %plot(C:(I-1)*C,Ekwi(:,2,1));
    %plot(C:(I-1)*C,Ekwi(:,3,1));
    %plot(C:(I-1)*C,Ekwi(:,4,1));
    %plot(C:(I-1)*C,Ekwi(:,5,1));

    %% Pkwi (=the observed proportions) P_k,w,i: group, category, item
    Pkwi=zeros(TNG,C+1,I);  % observed...
    % number of people belong to a group
    denP=zeros(TNG,1);  denPP=zeros(TNG,1);
    newtotd=totdat; % newtotdat: for missing treatment
    for j=1:N
        for i=1:I
            if totdat(j,i)==8
                newtotd(j,i)=100;
            end
        end
    end
    div= sum(newtotd');  % total test scores: 0  to  I*C
    for gr=1:I*C-1  % necessary groups :  from 1 to I*C-1
        for j=1:N
            if div(j)==gr
                denP(gr)=denP(gr)+1;
            end
        end
```

```
end
denPP=denP;
for gr=1:I*C-1
    if denP(gr)==0;
        denPP(gr)=999;
    end
end
numkwi=zeros(TNG,C+1,I);
for i=1:I
    for gr=1:I*C-1      %  from 1 to I*C-1
        for j=1:N
            if  (div(j)==gr) & (totdat(j,i)==0);
                numkwi(gr, 1,i) = numkwi(gr, 1,i) +1;
            elseif (div(j)==gr) & (totdat(j,i)==1);
                numkwi(gr, 2,i) = numkwi(gr, 2,i) +1;
            elseif (div(j)==gr) & (totdat(j,i)==2);
                numkwi(gr, 3,i) = numkwi(gr, 3,i) +1;
         %   elseif (div(j)==gr) & (totdat(j,i)==3);
         %       numkwi(gr, 4,i) = numkwi(gr, 4,i) +1;
            end
        end
    end
end
for i=1:I
    for w=1:C+1
        Pkwi(:,w,i) = numkwi(:,w,i) ./ denPP;
    end
end

%plot(C:(I-1)*C,Pkwi(:,1,1)); hold on;
%plot(C:(I-1)*C,Pkwi(:,2,1));
%plot(C:(I-1)*C,Pkwi(:,3,1));
%plot(C:(I-1)*C,Pkwi(:,4,1));
%plot(C:(I-1)*C,Pkwi(:,5,1));

% make pik and eik as a counts rather than proportion
ekwi=zeros(TNG,C+1,I); pkwi=zeros(TNG,C+1,I);
for i=1:I
    for w=1:C+1
        ekwi(:,w,i) = Ekwi(:,w,i) .* denP;
        pkwi(:,w,i) = Pkwi(:,w,i) .* denP;
    end
end

% final results of S-X^2 calculations:
% columns
% 1) condition name
% 2) data set number
% 3) item number
% 4) S-X^2
% 5) DF having -m in the calculaiton of DF, i.e. K(J-1)-m, considering loss of DF by collapsing
% 6) P-value with  I(NC-1)-m, considering loss of DF by collapsing
% 7) MISFIT item
% resultSX2=zeros(I*nd,7);
resultSX2((v-1)*I+1:v*I,2)=v;
% collapsing them to maintain a minimum Expected frequency equal to 1
for i=1:I
    eee=zeros(NG,C+1); ooo=zeros(NG,C+1);
    eee(1,:)=sum(ekwi(1:C,:,i)); ooo(1,:)=sum(pkwi(1:C,:,i));
    eee(NG,:)=sum(ekwi(TNG-(C-1):TNG,:,i)); ooo(NG,:)=sum(pkwi(TNG-(C-1):TNG,:,i));
    for gr=2:NG-1
        for w=1:C+1
            eee(gr,w)=ekwi(gr+(C-1),w,i);
            ooo(gr,w)=pkwi(gr+(C-1),w,i);
        end
    end
    % S-X2 and modified DF after collapsing
    imsisx2df=zeros(2,1); imsisx2df=sx2df(eee,ooo,oriDF);
    resultSX2((v-1)*I+i,3)=i;
```

```
        resultSX2((v-1)*I+i,4)=imsisx2df(1); %SX2(i);
        resultSX2((v-1)*I+i,5)=imsisx2df(2); %modified DF(i);
        resultSX2((v-1)*I+i,6)=1- chi2cdf(imsisx2df(1),imsisx2df(2));
        resultSX2((v-1)*I+i,7)= (resultSX2((v-1)*I+i,6) < .05);
    end
end
dlmwrite('SX2_gpgp2212.txt',resultSX2)
```

pofcateGPCM3.m

```
function[newpofc]=pofcateGPCM3(abilpe,itempe);
imsiN=size(abilpe); N=imsiN(1);
imsiI = size(itempe); I=imsiI(1);
nc=imsiI(2)-1;
gquad=abilpe(:,1); %densi=abilpe(:,2);
a=zeros(I,1); b=zeros(I,1); tau2=zeros(I,1); tau3=zeros(I,1); %tau4=zeros(I,1); tau5=zeros(I,1);
a=itempe(1:I,1);   b=itempe(1:I,2);   tau2=itempe(1:I,3);   tau3=itempe(1:I,4);   %tau4=itempe(1:I,5);
tau5=itempe(1:I,6);

pofc=zeros(N,I,nc); tt=zeros(N,I,nc); denom=zeros(N,I);
for j=1:N
    for i=1:I
        tt(j,i,1) = 1;
        tt(j,i,2) = exp(a(i)*(gquad(j)-b(i)+tau2(i)));
        tt(j,i,3) = exp(a(i)*(gquad(j)-b(i)+tau2(i) + gquad(j)-b(i)+tau3(i)));
        %tt(j,i,4)       =       exp(a(i)*(gquad(j)-b(i)+tau2(i)       +       gquad(j)-b(i)+tau3(i)+
gquad(j)-b(i)+tau4(i)));
        %tt(j,i,5) = exp(a(i)*(gquad(j)-b(i)+tau2(i) + gquad(j)-b(i)+tau3(i)+ gquad(j)-b(i)+tau4(i)+
gquad(j)-b(i)+tau5(i)));
        denom(j,i) = 1 + tt(j,i,2) + tt(j,i,3) ;% + tt(j,i,4) + tt(j,i,5);
    end
end
for j=1:N
    for i=1:I
        for w=1:nc
            pofc(j,i,w)=tt(j,i,w)/denom(j,i);
        end
    end
end

% to make it sure that the sum of pofc's(prob. of category) is the unity
for j=1:N
     for i=1:I
        totpofc(j,i)=pofc(j,i,1)+pofc(j,i,2)+pofc(j,i,3); %+pofc(j,i,4) +pofc(j,i,5);
        pofc(j,i,1) = pofc(j,i,1) / totpofc(j,i);
        pofc(j,i,2) = pofc(j,i,2) / totpofc(j,i);
        pofc(j,i,3) = pofc(j,i,3) / totpofc(j,i);
        % pofc(j,i,4) = pofc(j,i,4) / totpofc(j,i);
        % pofc(j,i,5) = pofc(j,i,5) / totpofc(j,i);
    end
end

% P(item, category, examinee) = P_iwj
newpofc=zeros(I,nc,N);
for j=1:N
    for i=1:I
        for w=1:nc
            newpofc(i,w,j)=pofc(j,i,w);
        end
    end
end
```

recursive3.m

```
function[xnew]=recursive3(temppro);

% input: probabilities of each examinee: item by category
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Recursive Formula: pp.219-221 in Kolen and Brennan (2004)%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
par=temppro;
itembyca=size(par); T=itembyca(1);
nc=ones(T,1)*3;
%nc=zeros(T,1);
%for ii=1:T
%    nc(ii)=sum(par(ii,:)<9);  % recognize if each item has different number of categories
%end

% Recursion Formula to get p(x|theta of the examinee j), x=T through T*ncat
xold(1) = par(1,1);              % Pr(X=1|theta) at the first item
xold(2) = par(1,2);              % Pr(X=2|theta) at the first item
xold(3) = par(1,3);              % Pr(X=3|theta) at the first item
%xold(4) = par(1,4);              % Pr(X=4|theta) at the first item
maxn=3;

for ii=2:T
    minn = ii  ; maxn=maxn + nc(ii);
    mino = ii-1; maxo=maxn - nc(ii);
    if ii>=3
        xold=xnew;
    end
    for h = minn:maxn
        in=h-minn+1;
        xnew(in)=0;
        for s=1:nc(ii)
            io=h-s-mino+1;
            if (io >= 1) & (io <= maxo-mino+1);
                xnew(in) = xnew(in) + xold(io)*par(ii,s);
            end
        end
    end
end
```

부록 3 : CS_1과 CS_2 계산을 위한 MATLAB 코드

- GM=GPCM, CM=GPCM, 검사 문항=20, 피험자 수=1,000인 경우

gpgp_FIT_22.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% codes for model-data fit analysis
%%% of Polytomous items (GPCM)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear all;  % CONDITION 2212, GPGP

% Data-Model FIT for GPCM
I=20; N=1000; nd = 50;
% Item Parameter
load 'c:/PROJ2011/KMY/parabu/gpgp2212all.txt'; itemest=gpgp2212all;
load c:₩PROJ2011₩KMY₩SX2₩SX2_gpgp2212.txt;
% ability distribution
k=-4:.2:4; den = normpdf(k,0,1); den=den/sum(den); J=length(k);
abilpara = ones(J,2); abilpara(:,1) = k'; abilpara(:,2) = den';
gquad=abilpara(:,1); densi = abilpara(:,2);
m=3; % Calibration Model = GPCM: 1 a, 1 b and 1 taus, # of parameters per item
% 123 data -> 012 data
nc=3; C=2; nos=I*C+1; x=[0:I*C]; NG=I*C-2*C+1;
TNG = I*C-1; % total number of groups that exclude zero(0) and full(I*C) scores.
nl=2*I + C*I; % number of lines per dataset
% original DF (before being midified with cell-collapsing)
oriDF=NG*C-m;

% result collects:
resultFIT=zeros(nd,9); resultFIT(:,1)=2212;
   % 1) condition name
   % 2) expected score distribution - observed score distribution: Chi-square
   % 3) d.f = #scores - 1
   % 4) p-value
   % 5) fit or misfit
   % 6) S-X2 based FIT
   % 7) d.f = #scores - 1
   % 8) p-value
   % 9) fit or misfit


for v = 1:nd;
    itempara=zeros(nl,1); itempara = itemest(v*nl-(nl-1):v*nl);
    a=itempara(1:I); b=itempara(I+1:I*2); tau = zeros(I,C);
    for u=1:I
       tau(u,1:C) = itempara(2*I+u*C-(C-1):2*I+u*C)';
    end
    itemesti=zeros(I,C+2); itemesti(:,1)=a; itemesti(:,2)=b; itemesti(:,3:(C+2))=tau;
    % read data
    dfilen =  ['c:/proj2011/KMY/gp_ori/2212/gp2212c', int2str(v), '.dat'];
    datc = load (dfilen); totdat = datc - 1; % to make 012 data with 123 data, missing is now "8"
    % calculating P_jiw for each item's category
    pofc=zeros(I,nc,J);
    pofc=pofcateGPCM3(abilpara,itemesti);  % P(item, category, examinee) = P_iwj
    % Conditional Distribution of Number-Correct Raw score for each theta value , Pr(X=x|theta)
    % Recursion Formula
```

```
    allitemSk=ones(J,nos); % Sk(theta)
    for j=1:J
        allitemSk(j,:)=recursive3(pofc(:,:,j));
    end
    % Distribution of Number-Correct Raw score , Pro(X=x)
    Pro=zeros(nos,2); Pro(:,1)=[0:1:I*C]';
    for z=1:nos
        for j=1:J
            Pro(z,2)=Pro(z,2) + allitemSk(j,z)*densi(j);
        end
    end
    % Expected Frequency Distribution of the raw scores
    EPro = Pro; EPro(:,2)= Pro(:,2)*N;
    % observed Frequency Distribution of the raw scores
    rs = sum(totdat'); OPro = zeros(nos,2); OPro(:,1)= [0:1:I*C]';
    for z=1:nos
        for j=1:N
            if rs(j) == (z-1)
                OPro(z,2) = OPro(z,2) + 1;
            end
        end
    end
    %dlmwrite('imsi1.txt',rs')
    % Chi-Square
    CS1 = 0; DF1= nos-1;
    CS1res = zeros(2,1); % chi-sqare value and modified df
    CS1res = CS1cal(EPro(:,2), OPro(:,2),DF1);
    resultFIT(v,2)= CS1res(1);
    resultFIT(v,3)= CS1res(2);
    resultFIT(v,4)= 1- chi2cdf(CS1res(1),CS1res(2));
    resultFIT(v,5)= (resultFIT(v,4) < .05);
        % S-X2 based FIT
    iFIT = SX2_gpgp2212(v*I-(I-1):v*I,4:5);
    resultFIT(v,6)= sum(iFIT(:,1));
    resultFIT(v,7)= sum(iFIT(:,2));
    resultFIT(v,8)= 1- chi2cdf(resultFIT(v,6), resultFIT(v,7));
    resultFIT(v,9)= (resultFIT(v,8) < .05);
end
dlmwrite('FIT_gpgp2212.txt',resultFIT)
```