creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

교육학박사 학위논문

베이지안 추정을 활용한 위계적 채점자 모형의 확장 및 적용 연구

2021년 8월

서울대학교 대학원

교육학과 교육학전공

박 민 호

베이지안 추정을 활용한 위계적 채점자 모형의 확장 및 적용 연구

지도교수 박 현 정

이 논문을 교육학박사 학위논문으로 제출함
2021년 6월

서울대학교 대학원
교육학과 교육학전공
박 민 호

박민호의 박사 학위논문을 인준함
2021년 8월

위 원 장 백 순 근

부위원장 강 태 훈

위 원 박 찬 호

위 원 함 은 혜

위 원 박 현 정

국문초록

이 연구의 목적은 대규모 검사의 구성형 문항 채점에서 나타나는 채점자 특성에 따른 영향이 문항에 따라 다르게 나타나는 것을 정확하게 파악하기 위해서 기존의 위계적 채점자 모형(HRM)을 확장한 모형을 제안하고, 이 모형의 정확성과 효율성을 검증하는 것이다. 최근에 평가에서 널리 활용되고 있는 수행평가의 단순한 형태인 구성형 문항은 채점자의 평정을 거치기 때문에 측정학적 측면에서 채점 신뢰성이 확보되어야한다. 이를 위해서는 채점자의 전문성, 엄격성, 일관성 등이 갖춰져야 하므로 채점 기준 개발에 많은 노력을 기울이고, 채점에 앞서 채점자 훈련을 실시하지만 완전한 통제가 어려운 것이 사실이다. 때문에 통계적 분석 방법으로 채점자 특성의 영향을 제거하고 피험자의 능력을 정확히 추정하기 위한 문항반응이론 모형들이 등장하였고, 본 연구에서 다루는 HRM도 등장하게 되었다.

전통적인 문항반응이론 모형들부터 HRM에 이르기까지 문항반응이론 모형들은 대부분 채점자의 특성이 채점에 미치는 영향이 문항과 무관하게 일정한 것으로 모형화하였다. 그러나 채점자가 문항과 무관한 채점 특성을 가지는 채점자-문항 독립성 가정은 현실적으로 지켜지기 어렵고, 분산분석에 바탕을 둔 일반화가능도 연구에서 이를 위배하는 사례들이 흔하게 보고된다. 이러한 상황에서 논술형 문항 채점에 많이 활용되는 신호탐지이론-위계적 채점자 모형(HRM-SDT)에 채점자-문항 독립성 가정에 따른 제약을 제거한 연구가 등장하였으나 모형의 복잡성으로 소수의 논술형 문항 채점에 주로 활용되는 데 그쳤고, 다수의 구성형 문항으로 이루어진 검사에는 적용되지 못하고 있다.

이에 본 연구에서는 채점자의 일관성을 무선효과로 모형화하여 간명성 측면에서 장점을 가지는 HRM에서 채점자-문항 독립성 제약을 완화하여 재모수화하는 모형을 탐색하고, 실제 자료에 적용하는 연구와 모형의 정확성 및 효율성을 검증하는 모의실험 연구를 진행하고자 하였다.

개발, 검증, 적용을 거치는 일반적인 모형 개발 연구와는 달리 본 연구는 기존 모형에 바탕을 두고 있기 때문에 채점자-문항 독립성 제약 완화에 따라 가능한 모든 모형을 먼저 실제 자료에 적용하여 모형의 적용 가능성을 확인하고, 이후에 정확성과 효율성을 검증하는 절차를 거치는 순서로 연구를 진행하였다.

적용 연구에서는 HRM에 포함된 채점자 엄격성 모수와 채점자 분산성 모수의 채점자-문항 독립성 제약 완화에 따라 총 4가지 모형을 분석하였다. 이들 모형을 베이지안 기반의 MCMC 방법으로 국제비교 대규모 성취도 검사인 PIRLS 2006 자료에 적용함으로써 실제 분석 가능성을 탐색하였고, 제안 모형의 정확성과 효율성을 검증하는 모의실험 연구를 진행하였다. 적용 연구 및 모의실험 연구의 결과를 제시하면 다음과 같다.

적용 연구 결과, 채점자의 일관성이 문항에 따라 다름을 가정하고 분산성 모수에서만 채점자-문항 독립성 제약을 완화한 E-HRM(Extended HRM)이 MCMC 연쇄 수렴과 해석가능성에서 가장 나은 결과를 보여 제안 모형으로 타당함을 확인하였다. 그리고 HRM 분석에서는 채점자 특성의 영향이 전체 문항에서는 평균적으로 일관된 것처럼 보이더라도 E-HRM 분석을 통해 채점자 특성의 영향이 특정 문항에서는 일관되지 않을 수 있다는 것을 확인하였고, 상세하고 유의미한 채점자 정보를 추가로 획득할 수 있음을 증명하였다. 또한 이질적인 채점 분포를 가지는 문항들이 혼재된 채점 자료를 분석하였을 때, HRM은 추정이 불안정하거나 이질성을 놓치는 취약점을 보였지만, E-HRM은 안정적인 추정으로 채점 특성의 이질성을 분석할 수 있다는 장점을 보였다. 반면, 이론적으로 1개의 관찰값만 있더라도 추정이 가능하다는 MCMC 방법에서도 채점자-문항당 채점건수가 어느 정도 이상 확보되지 못하면 E-HRM이 적절하게 수렴하지 못하는 것으로 나타났다.

모의실험 연구에서는 '정상범위 채점자 비율', '채점자-문항당 채점건수'의 2개 요인으로 12개의 조건을 구성하여 E-HRM의 정확성과 효율성

을 평가하였다. 모의실험 결과로 첫째, 채점건수가 적을수록 E-HRM의 모수 추정 정확도와 효율성이 낮아지고 전반적 오차가 증가하는 경향이 나타났다. 채점자 엄격성 모수에 대한 추정치 사후분포가 95% 신뢰구간 내에 참값을 포함하는 비율은 12개 실험설계 조건에서 중 하나의 조건을 제외하고는 모두 95% 수준으로 상당히 높은 정확성을 보였다. 그리고 분산성 모수에 대한 추정치의 사후분포가 95% 신뢰구간 내에 참값을 포함하는 비율은 채점자-문항당 채점건수 '200건'에서 정상범위 채점자 비율 '100%', '70%', '40%' 조건들에 대해 각각 .953, .948, .944로 95%에 근사한 수치를 보였다. 하지만 채점건수가 감소할수록 줄어들어 25건에서는 정상범위 채점자 비율 '100%', '70%', '40%' 조건들에 대해 각각 .881, .894, .922로 다소 낮은 것으로 나타났다.

둘째, 정상범위 기준을 벗어나는 채점자 모수에 대한 판별 측면에서 E-HRM은 정상범위 외 채점자 분산성 모수에 대한 높은 판별 정확도를 보였다. 채점건수별로 E-HRM의 민감도와 특이도를 분석한 결과, 채점건수가 '200건'일 때는 민감도 .953, 특이도 .988로 정상범위 채점 특성과 정상범위 외 채점 특성을 매우 높은 확률로 판별해내었다. 그리고 채점건수 '100건'인 경우에도 민감도 .903, 특이도 .985로 90% 이상의 판별 정확성을 보였지만, 채점건수가 감소함에 따라 민감도가 낮아져 '50건'과 '25건'인 경우 민감도는 각각 .864, .794로 다소 낮았다.

셋째, E-HRM은 채점 특성의 영향을 정확히 판별하고 이 영향을 적절히 보정하여 엄격(혹은 관대)하거나 신뢰도가 낮은 채점자 특성이 피험자 모수와 문항 모수의 정확성 및 효율성에 미치는 영향을 통제할 수 있는 것으로 나타났다. 그리고 채점자 모수 추정치의 사후분포가 다봉분포가 뚜렷한 경우가 빈번해 HRM의 수렴 비율과 정확성이 떨어지는 상황에서는 E-HRM은 추정에 큰 문제가 없고, HRM보다 더 나은 모수 복원력을 보였다.

이 연구는 베이지안 추정을 활용한 E-HRM을 하나의 실제 자료에 적용하였고, 채점 자료의 다양한 설계를 모두 포괄하지 못하는 한정적인

조건하에서만 E-HRM의 성능을 평가하였다는 제한점이 있다. 그러나 다양한 채점 설계를 가지는 PIRLS 2006의 24개국 자료를 베이지안 추정을 활용한 E-HRM으로 분석하여 E-HRM의 추정 및 해석 가능성을 확인하고, 모의실험을 통해 특정 조건하에서 E-HRM의 정확성과 효율성을 검증하였다는 점에 의의가 있다. 그리고 E-HRM 유용성을 확인하였다는 점에서 E-HRM이 제공하는 채점자-문항 단위의 상세한 채점자 특성 정보는 채점자의 전문성 향상을 위한 중요한 진단적 정보로 활용되고, 대규모 검사의 채점자 재훈련 및 재채점에 들어가는 노력, 시간, 비용을 크게 절감하는 데에도 매우 유용하게 활용될 수 있으리라고 기대한다.

* 주요어: 위계적 채점자 모형, 베이지안 추정, MCMC, 깁스 표본 추출법, 채점자 효과
* 학 번: 2016-30397

목 차

Ⅰ. 서론 ····· 1
Ⅱ. 이론적 배경 ····· 8
1. 고전적 채점자 모형 ····· 8
2. 위계적 채점자 모형 ····· 15
3. 위계적 채점자 모형의 확장 ····· 22
Ⅲ. 연구방법 ····· 29
1. 적용 연구 대상 ····· 29
2. 적용 연구 분석 절차 ····· 46
3. 모의실험 연구 절차 ····· 50
Ⅳ. 적용 연구 결과: Overview ····· 61
1. 모형 수렴 ····· 61
2. 모형 적합도 ····· 72
Ⅴ. 연구 결과: 국가별 분석 ····· 82
1. 아이슬란드 ····· 82
2. 벨기에(프랑스어권) ····· 97
3. 덴마크 ····· 107
Ⅵ. 모의실험 연구 결과 ····· 119
1. 모의실험 채점자 모수 ····· 119
2. 모형 수렴 ····· 121
3. 모수 복원력 ····· 125
Ⅶ. 결론 ····· 145
1. 요약 ····· 145
2. 논의 ····· 150
참 고 문 헌 ····· 155

표 목 차

<표 Ⅱ-1> HRM 계열 모형별 채점자 모수 수 ····· 26
<표 Ⅱ-2> PIRLS 2006 자료 분석 시 모형별 채점자 모수 수
(러시아, 마케도니아) ····· 27
<표 Ⅲ-1> PIRLS 2006 참여국가별 학교수 및 학생수 ····· 29
<표 Ⅲ-2> PIRLS 2006 문항구획 Antarctica 지시문 관련 문항 일부 ··· 31
<표 Ⅲ-3> PIRLS 2006 정보 획득 및 활용 영역 R011A07C 채점 기준 ··· 32
<표 Ⅲ-4> PIRLS 2006 문학적 경험 R011C10C 채점 기준 ····· 33
<표 Ⅲ-5> PIRLS 2006 문항 구획 및 소책자 구성 ····· 34
<표 Ⅲ-6> 문항 전처리 후 문항별 기술통계 ····· 36
<표 Ⅲ-7> 국가별 채점 규모 통계 ····· 39
<표 Ⅲ-8> 분석 자료 소책자별 문항 구성 ····· 40
<표 Ⅲ-9> 분석 자료 채점자-문항 설계 ····· 42
<표 Ⅲ-10> 분석 자료 채점자-문항별 소규모 채점 비율 ····· 43
<표 Ⅲ-11> PIRLS 2006 분석 자료 채점자-문항당 채점건수 분포 ····· 51
<표 Ⅲ-12> 모의실험 연구 문항 모수 ····· 55
<표 Ⅲ-13> 모의실험 연구 설계 요인 및 수준 ····· 57
<표 Ⅳ-1> 국가 및 모형별 전체 모수 수렴 비율 ····· 61
<표 Ⅳ-2> 문항 모수 수렴률 ····· 63
<표 Ⅳ-3> 채점자 모수 수렴 비율 ····· 65
<표 Ⅳ-4> 채점건수에 따른 채점자 엄격성 모수 수렴 비율 ····· 68
<표 Ⅳ-5> 채점건수에 따른 채점자 분산성 모수 수렴 비율 ····· 69
<표 Ⅳ-6> 모형별 Deviance 통계량 ····· 72
<표 Ⅳ-7> 모형별 pD 통계량 ····· 73
<표 Ⅳ-8> 모형별 DIC 통계량 및 DIC 최소 모형 ····· 74
<표 Ⅳ-9> E-HRM 적합도 악화 채점자 ····· 78
<표 Ⅳ-10> E-HRM 적합도 악화 채점자 채점 특성 ····· 79
<표 Ⅴ-1> 채점자별 채점 배정: 아이슬란드 ····· 83

<표 Ⅴ-2> 문항 모수: 아이슬란드 ····· 84
<표 Ⅴ-3> 채점자 모수: 아이슬란드 ····· 86
<표 Ⅴ-4> 채점자별 채점 배정: 벨기에(프랑스어권) ····· 98
<표 Ⅴ-5> 문항 모수: 벨기에(프랑스어권) ····· 99
<표 Ⅴ-6> 채점자 모수: 벨기에(프랑스어권) ····· 100
<표 Ⅴ-7> 채점자 모수에 따른 이상평정-실제평정의 편의:
벨기에(프랑스어권) 3번 채점자 7번, 20번 문항 ····· 103
<표 Ⅴ-8> 채점자별 채점 배정: 덴마크 ····· 108
<표 Ⅴ-9> 채점자 및 문항별 채점건수: 덴마크 ····· 109
<표 Ⅴ-10> 문항 모수: 덴마크 ····· 110
<표 Ⅴ-11> 채점자 모수: 덴마크 ····· 112
<표 Ⅵ-1> 모의실험 채점자 모수 참값 ····· 120
<표 Ⅵ-2> 모의실험 채점자 모수 수렴 비율 ····· 122
<표 Ⅵ-3> 채점자 모수 추정값 95% 신뢰구간 내 모수 참값 포함 비율 ··· 125
<표 Ⅵ-4> 채점자 모수 추정치 편향 ····· 127
<표 Ⅵ-5> 채점자 모수 추정치 SB ····· 128
<표 Ⅵ-6> 채점자 모수 추정치 VAR ····· 130
<표 Ⅵ-7> 채점자 모수 추정치 MSE ····· 132
<표 Ⅵ-8> 모형별 분산성 모수 이상 기준 판별률 ····· 135
<표 Ⅵ-9> HRM 분산성 모수 이상 기준 판별률 ····· 138
<표 Ⅵ-10> 문항 변별도 모수 추정치 bias, SB, VAR, MSE ····· 140
<표 Ⅵ-11> 문항 곤란도 모수 추정치 bias, SB, VAR, MSE ····· 141
<표 Ⅵ-12> 피험자 능력 모수 추정치 bias, SB, VAR, MSE ····· 143

그 림 목 차

[그림 Ⅱ-1] Patz et al.(2002)의 HRM 모형 ····· 17
[그림 Ⅱ-2] 채점자 엄격성 모수에 따른 채점 확률 ····· 19
[그림 Ⅱ-3] 채점자 분산성 모수에 따른 채점 확률 ····· 20
[그림 Ⅲ-1] PIRLS 2006 문항구획 Antarctica 지시문 일부 ····· 31
[그림 Ⅲ-2] 채점자 사회관계망 분석 그래프 ····· 45
[그림 Ⅳ-1] 채점건수에 따른 채점자 엄격성 모수 수렴 비율 ····· 69
[그림 Ⅳ-2] 채점건수에 따른 채점자 분산성 모수 수렴비율 ····· 71
[그림 Ⅳ-3] HRM 및 E-HRM 채점자별 Deviance ····· 76
[그림 Ⅳ-4] HRM 대비 E-HRM Deviance 비율에 따른
채점자 분산성 모수 최댓값 ····· 77
[그림 Ⅴ-1] 채점자 엄격성 모수: 아이슬란드 ····· 87
[그림 Ⅴ-2] 채점자 분산성 모수: 아이슬란드 ····· 89
[그림 Ⅴ-3] 채점자별 이상 평정 추정치-실제 평정 분포(HRM):
아이슬란드 ····· 91
[그림 Ⅴ-4] 채점자별 이상 평정 추정치-실제 평정 편차 분포(HRM):
아이슬란드 ····· 93
[그림 Ⅴ-5] 문항별 이상 평정 추정치-실제 평정 편차 분포(E-HRM):
아이슬란드 6번 채점자 ····· 96
[그림 Ⅴ-6] MCMC Trace plot: 아이슬란드(E-HRM)
6번 채점자, 7번 문항 분산성 모수 ····· 97
[그림 Ⅴ-7] 채점자 엄격성 모수: 벨기에(프랑스어권) ····· 102
[그림 Ⅴ-8] 채점자 분산성 모수: 벨기에(프랑스어권) ····· 104
[그림 Ⅴ-9] 문항별 이상 평정 추정치-실제 평정 편차 분포(E-HRM):
벨기에(프랑스어권) 4번 채점자 ····· 105
[그림 Ⅴ-10] 문항별 이상 평정 추정치-실제 평정 분포(E-HRM):
벨기에(프랑스어권) 4번 채점자 ····· 106
[그림 Ⅴ-11] 채점자 엄격성 모수: 덴마크 ····· 113

[그림 Ⅴ-12] 채점자 분산성 모수: 덴마크 ····· 114
[그림 Ⅴ-13] 분산성 모수의 추정 대상 채점건수 -
95% 신뢰구간 상한값 분포: 덴마크(E-HRM) ····· 115
[그림 Ⅴ-14] 분산성 모수의 추정 대상 채점건수별 MCSE 평균:
덴마크(E-HRM) ····· 116
[그림 Ⅴ-15] 분산성 모수의 추정 대상 채점건수별 ESS 평균:
덴마크(E-HRM) ····· 116
[그림 Ⅴ-16] 분산성 모수의 추정 대상 채점건수별 MCSE 평균:
전체 국가(E-HRM) ····· 117
[그림 Ⅴ-17] 분산성 모수의 추정 대상 채점건수별 ESS 평균:
전체 국가(E-HRM) ····· 118
[그림 Ⅵ-1] HRM 채점자 모수의 MCMC 표본 분포
(정상범위 채점자 70%, 채점건수 200건 조건, 반복1, 채점자 4번) ··· 124

Ⅰ. 서론

21세기에 들어서면서 평가에 대한 세계적인 동향 중의 하나는 선택형(selected- response item format) 검사를 위주로 하여 학생들을 선발·분류·배치하기 위한 양적 평가로부터, 수행평가(performance assessment)를 위주로 하여 개별 학생에게 지도하고 조언하고 충고하기 위한 질적 평가로의 전환이 이루어지고 있다는 점이다(백순근, 2000). 수행평가에 대한 관심은 비단 우리나라만의 현상은 아니다. 서구의 여러 국가들에서도 역시 전통적인 평가 도구인 지필식 선택형 검사를 지양하고 새로운 형태의 평가 방식을 시도하고 있다. 기존의 평가 도구로는 최근 강조되고 있는 창의적 문제해결력, 협동학습능력, 의사소통능력과 같은 고차원적이고 복합적인 역량을 적절하게 측정하기 어려운 상황에서 기존 평가 도구의 한계를 극복할 수 있는 대안으로 수행평가가 부상하게 된 것이다(박도순, 2007; Kane & Mitchell, 1996; Linn, 1994).

우리나라에서는 1990년 초에 수행평가와 관련한 개념이 소개되었고, 1996년경부터 초등학교의 열린교육 확산 추세에 따라 수행평가가 학생평가에 있어서 하나의 대안으로 관심을 끌기 시작하였다(백순근, 1999). 그리고 2000년부터 시행된 제7차 교육과정과 함께 초・중등교육에 수행평가가 전면적으로 도입되어 학생평가의 변화를 시도하였다. 한편 우리나라는 수행평가가 학교 단위에서 주로 활용된 것과 달리, 우리보다 먼저 1980년대에 수행평가를 도입한 미국에서는 더 큰 단위인 주 단위, 또는 국가 단위의 평가에 활용하였다(박종임, 노은희, 김유향, 전경희, 2017). 미국의 경우 성취기준 기반 교육개혁(standard-based reform)의 맥락에서 California Learning and Assessment System(CLAS), Maryland State Performance Assessment Program(MSPAP), Wyoming Body of Evidence(BOE) 등과 같은 주 단위 평가에 수행평가를 도입하였고(Wei, Pecheone, & Wilczak, 2014), 이를 통해 학교 교육의 질 향상과 책무성 보장을 꾀하였다(박종임 외, 2017).

수행평가는 초·중등교육의 학업성취도 평가뿐만 아니라 사법시험, 대학입학 논술고사 등과 같은 고부담 시험(high-stake test)에도 활용되고 있다(법무부, 2005; 지은림, 1996; 허명회, 한상태, 1996). 그리고 수행평가의 여러 가지 형태 중 하나인 구성형 문항(constructed response item)이 국가수준 학업성취도 평가, 한국어능력시험(TOPIK), TOEFL (Test Of English as a Foreign Language), TOEIC(Test Of English for International Communication) Speaking and Writing Tests, IELTS(International English Language Testing System), GRE(Graduate Record Examination), 대학생 핵심역량 진단(K-CESA) 등과 같은 대규모 표준화 검사에서도 활용되고 있다. 또한 수행평가 확산 추세는 대규모의 국제비교 연구에도 영향을 미쳐, 2001년부터 5년 주기로 시행되는 국제읽기역량평가(Progress in International Reading Literacy Study; 이하 PIRLS), 1995년 이후 4년 주기로 실시되고 있는 수학·과학 성취도 국제비교 연구(the Third International Mathematics and Science Study; 이하 TIMSS) 및 반복연구(TIMSS Repeat; TIMSS-R)나 OECD 주관의 학업성취도 국제비교 연구(Programme for International Student Assessment; 이하 PISA)의 검사 도구도 구성형 문항을 다수 포함하고 있다(박정, 박경미, 2004).

이와 같이 널리 활용되고 있는 구성형 문항으로 구성된 검사 결과가 측정학적인 측면에서 신뢰도와 타당도를 가지기 위해서 선행되어야 할 몇 가지 전제 조건들이 여러 연구자들로부터 제기되고 있다. 특히 측정학적인 측면에서 구성형 문항과 관련된 여러 가지 문제들이 해결되어야 한다는 주장이 제기되었다(남명호, 1995; 남현우, 1998; 민경석, 양길석, 2013; 허경철, 김왕근, 조덕주, 양미경, 차우규, 박선미, 2001; Engelhard, 1996; Engelhard & Wind, 2017; Hattie, Jaeger, & Bond, 1999; Lunz & Stahl, 1990). 구성형 문항이 고차원적이고 복합적인 역량을 제대로 측정할 수 있는지, 기존의 선다형 문항 검사에 비해 다소 적은 문항을 활용할 수밖에 없는 구성형 문항으로 이루어진 검사를 통해 피험자 능력을

정확히 추정할 수 있는지, 구성형 문항 응답에 대한 채점자의 평정으로 획득한 다분 반응 자료를 통해서 피험자의 능력을 제대로 추정할 수 있는지 등이 측정학적인 측면에서의 구성형 문항에 대한 물음이다.

이상의 구성형 문항에 대한 측정학적인 관점에서의 문제 제기 중에서 이 연구에서 집중하여 다루고자하는 문제는 구성형 문항 응답에 대한 채점자의 평정으로 획득한 다분 반응 자료를 통해서 피험자의 능력을 제대로 추정할 수 있는지와 관련된 신뢰성의 문제이다. 구성형 문항의 신뢰성에 영향을 미치는 대표적인 요인으로 채점자의 전문성(expertise), 엄격성(severity), 일관성(consistency) 등을 들 수 있다(지은림, 2008; Engelhard, 1994; Engelhard et al., 2017; Leckie & Baird, 2011; Myford & Wolfe, 2009). 채점자 전문성은 채점자가 구성형 문항 과제들의 주제 내용, 문항 형태 등에 대해 얼마나 전문적인 지식을 가지고 있는지와 채점 기준에 대해 얼마나 정확하게 이해할 수 있는지를 나타내는 것으로 채점에 대한 경험과도 연계된다. 그리고 채점자 엄격성은 동일한 대상에 대해 높은 점수를 주는 경향이 있는 채점자가 있는 반면에, 낮은 점수를 주는 경향이 있는 채점자가 있다는 것을 말한다. 즉, 낮은 점수를 부여하는 채점자의 경우에 엄격성이 높다고 할 수 있고, 전통적인 문항반응이론 연구에서는 구성형 문항에서 채점자간 차이를 유발하는 가장 본질적인 이유를 채점자간에 엄격성의 차이로 보고 있다(지은림, 1996; Congdon & McQueen, 2000; Linacre, 1994; Lunz & Stahl, 1990). 마지막으로 채점자 일관성은 채점자의 평정이 안정적이고 일관되며 오류가 없다는 것을 의미하고, 때로는 평가자 신뢰도나 평가자 간의 동의(agreement)로 다루어질 수 있다(Fitzpatrick, Ercikan, Yen, & Ferrara, 1998; Yen, 2016).

구성형 문항 검사의 신뢰성을 확보하기 위해서는 앞서 살펴본 채점자의 전문성, 엄격성, 일관성 등이 확보되어야 하겠지만, 현실에서는 채점자 특성이 피험자의 능력과 무관하게 채점 결과에 영향을 주는 것을 완벽히 통제하는 데에 한계가 있고, 채점자에 따라 구성형 문항의 신뢰성

을 확보하기 어려운 상황이 발생할 가능성을 배제할 수 없다. 특히 PIRLS, PISA, TIMSS와 같은 국제비교 연구나 우리나라의 국가수준 학업성취도 평가에서 이루어지는 대규모 학업성취도 검사에서는 이미 복잡한 피험자-문항 설계에 더하여 구성형 문항의 비중이 확대되어 채점에 대한 신뢰성 확보가 더욱 어려운 경우가 발생할 수 있다. 단적인 예로 TIMSS 2011까지만 하더라도 지필 검사지를 직접 이용하여 수기 채점 방식으로 채점을 진행하면서 각각의 채점자가 한 검사지에 포함된 수학 및 과학 문항을 함께 채점하는 설계였고(김수진 외, 2015), 때문에 과학 교과 전문가가 수학 교과까지 채점하거나 반대로 수학 교과 전문가가 과학 교과까지 채점하여 구성형 문항 채점의 신뢰성을 의심받을 수 있는 상황이었다.

채점 신뢰성 문제를 해결하는 일차적인 방안은 채점자가 수행형 문항의 답안을 평정할 때 활용하는 채점 기준(rubric)을 명확하게 개발하여 채점자 특성의 영향이 발생할 가능성을 줄이는 것이다(백순근 외, 2010; Eckes, 2017). 이는 Hattie et al.(1999), Engelhard et al.(2017) 등이 제시한 교육 검사에서의 신뢰성 및 타당성과 관련한 5가지 주요 쟁점들인 측정 모형(measurement models), 평가 활용(use of assessments), 평가 개발(assessment development), 평가 관리(administration of assessments), 메타 평가(evaluation of assessments) 중 평가 개발 측면에서의 노력이라 할 수 있다. 이어서 채점 기준을 적용하는 과정에서 나타나는 채점자 특성의 영향을 줄이기 위해서는 채점 절차를 진행하기 전에 채점자 훈련을 통하여 채점자간 특성 차이를 최소화하고 채점자내 일관성을 최대화하는 평가 관리 측면에서 노력이 필요하다(Brennan, 2001; Eckes, 2017).

하지만 평가 개발, 평가 관리 측면에서의 노력들에도 불구하고 채점자 특성의 영향을 완전히 통제하기란 사실상 불가능하다(이현숙, 2008; Lunz & Stahl, 1990). 따라서 채점자 특성에 따른 차이를 배제하고 일반적으로 검사의 목적인 피험자 능력의 정확한 추정을 위해서는 통계적 분석 방법을 통해서 채점자 특성의 영향을 제거하여 보정하는 측정 모형

측면에서의 접근이 필요하다. 그리고 채점자 특성에 따른 영향을 제거하기 위해서는 먼저 구성형 문항 채점에서 나타나는 채점자 특성에 따른 영향을 정확하게 파악하는 것이 무엇보다 필수적이다.

이러한 이유로 구성형 문항의 채점 자료에 대한 분석 방법이 점차 더 관심을 받게 되었고, 고전검사이론에 바탕을 두는 일반화가능도 이론(Generalizability Theory)과 문항반응이론에 바탕을 두는 다국면 라쉬 모형(Many-Facet Rasch Model)을 활용한 연구들이 활발히 이루어졌다(Engelhard et al., 2017). 그리고 최근에는 여러 가지 채점자 효과의 분석이 가능한 위계적 채점자 모형(Hierarchical Rater Model; 이하 HRM)을 적용한 다양한 연구들이 진행되었다(Casabianca & Wolfe, 2017; Casabianca, Junker, Nieto & Bond, 2017; DeCarlo, 2002; DeCarlo, 2005; DeCarlo, 2008; DeCarlo, 2010; Kim, 2009; Kim 2013; Nieto & Casabianca, 2019; Patz & Junker, 1999; Patz, Junker, Johnson & Mariano, 2002).

HRM에 대한 연구들에서는 전통적인 문항반응이론 기반의 채점자 효과 모형에서부터 주로 다루어 오던 채점자 엄격성 효과에서 더 나아가 채점자의 일관성과 중심경향성 효과까지 여러 방식으로 채점자 효과를 분리하였지만, 전통적 채점자 효과 모형과 마찬가지로 채점자의 영향이 문항과 독립적이라는 가정을 전제로 분석되어 왔다. 그러나 구성형 문항 중에서도 답안의 길이가 긴 논술형(essay) 문항 채점과 같이 소수의 문항만을 활용하는 검사가 아니라 국제비교연구들과 같이 수십 개의 짧은 구성형 문항으로 이루어진 검사의 채점 자료에서 채점자의 특성에 따른 영향이 문항과 독립적일 것이라고 기대하기 어렵다. 또한 채점자와 문항이 독립적이지 않을 수 있다는 근거는 고전적인 일반화가능도 연구에서도 발견된다. 채점 자료에 대한 다변량 일반화가능도 연구에서 (피험자,p)×(문항,i)×(채점자,r) 설계로 문항과 채점자의 상호작용을 분석하는 경우가 빈번하고(Briggs & Wilson, 2007), 검사의 여러 가지 특성에 따라 문항과 채점자의 상호작용이 거의 없는 경우도 있지만 큰 경

우에는 분산의 20%를 넘어가는 경우도 발견된다(Shavelson, Webb, & Rowley, 1989; 김창환, 이규민, 2008; 황지유, 원미진, 2019)는 점에서 채점자와 문항의 독립성 가정은 실제 자료에서 충족되기 쉽지 않은 가정이라고 판단된다.

DeCarlo, Kim, & Johnson(2011), Patterson(2013), Patterson, Wind, & Engelhard(2017)의 연구에서는 채점자 모수를 문항에 따라 각각 다르게 설정하는 모수화 방식을 신호탐지이론 기반 위계적 채점자 모형(Hierarchical Rater Model based on Signal Detection Theory; 이하 HRM-SDT)에 적용하여 채점자와 문항의 독립성 가정을 제거한 모형의 활용 가능성을 보여주었다. 하지만 채점 범주 간 경계 모수들을 활용하는 HRM-SDT의 모형의 특성상 채점자 모수의 개수는 채점자수와 범주수가 곱해지는 방식으로 늘어난다. 게다가 채점자와 문항의 독립성 가정을 풀면서 문항 수마저 채점자수, 범주수와 함께 곱으로 모형의 모수가 증가하기 때문에 추정해야할 모수가 지나치게 많아지는 문제가 존재한다. 이러한 모형 추정의 현실적 어려움 때문에 채점자 모수와 문항의 독립성을 가정한 HRM-SDT 활용 연구에서는 5개 범주 점수로 채점되는 2~4개 논술형 문항을 분석하는데 그쳤고(DeCarlo et al., 2011; Patterson, 2013; Patterson et al., 2017), 상대적으로 답안의 길이가 짧은 다수의 구성형 문항으로 이루어진 PIRLS, TIMSS, PISA 등의 대규모 검사에는 적용하기 쉽지 않다.

다수의 구성형 문항으로 이루어진 대규모 검사에 HRM 계열의 적용할 때에 나타나는 '채점자-문항 독립성으로 인한 정보 부족 문제'와 '현실적 추정 가능성의 문제' 사이의 타협 지점을 찾기 위한 시도로 본 연구에서는 무선효과 추정의 장점을 가진 Patz et al.(2002)의 HRM을 채점자-문항 독립성 제약을 완화하는 방식으로 재모수화한 모형을 제안한다. Patz et al.(2002)의 HRM은 채점자-범주경계별 채점자 모수를 모형에 포함하는 DeCarlo(2002)의 HRM-SDT에 비하여 상대적으로 간명한 모형이다. Patz et al.(2002)의 HRM은 DeCarlo(2002)의 HRM-SDT과 달

리 채점자에 문항-범주경계별 채점 특성 정보를 줄 수는 없지만, 채점자-문항 간의 독립성 가정을 완화하더라도 모형 복잡성 증가가 적다는 점에서 추정에 유리하기 때문에 다수의 구성형 문항을 포함하는 대규모 검사에서도 상세한 채점자-문항별 채점 특성 정보들을 제공할 수 있는 가능성이 높다. 따라서 본 연구에서는 채점자-문항 독립성 제약이 완화된 HRM을 다수의 구성형 문항으로 이루어진 실제 자료인 PIRLS 2006에 적용하여 채점자와 문항의 상호작용으로 발생하는 채점자 효과를 찾아내고, 다양한 설계 상황에서 모형이 잘 작동할 수 있는지에 대한 연구를 진행하고자 한다. 이러한 문제의식을 담아 연구 문제들을 기술하면 다음과 같다.

첫째, 채점자-문항 독립성 제약이 완화된 위계적 채점자 모형은 채점자 모수의 문항별 차이를 판별할 수 있는가?

둘째, 채점자-문항 독립성 제약이 완화된 위계적 채점자 모형의 모수 추정은 얼마나 정확하고 효율적인가?

Ⅱ. 이론적 배경

1. 고전적 채점자 모형

채점 자료에 대한 분석은 고전검사이론에서부터 다양한 접근이 이루어졌다. 고전검사이론에 바탕을 둔 일치도 통계, Cohen의 Kappa나 일반화가능도 이론에서의 채점자 영향 분석 접근은 채점자 간에 평정 결과가 얼마나 일치하느냐 즉, 채점자 간 신뢰도에 초점을 두고 있었다. 따라서 이 접근들에서는 채점자 특성으로 인한 영향이 피험자 능력 추정의 정확성에 어떠한 영향을 미치는지 파악하고, 채점자 영향이 배제된 피험자의 정확한 능력치를 추정하는 수단이 될 수는 없다는 한계점을 가진다. 이에 따라 채점자 영향을 분석하거나 피험자 능력 추정치에 대하여 채점자 영향으로 인한 차이를 사후에 보정하는 방법으로 문항반응이론(item response theory)에 바탕을 둔 방법들이 제안되었다(Engelhard et al., 2017; Linacre, 1989; 이현숙, 2008)

문항반응이론 모형은 추정되는 문항모수의 수에 따라 1모수, 2모수, 3모수 모형(또는 Birnbaum 모형)으로 구분되며 이러한 모형들은 초기에 반응이 정답 또는 오답으로 표현되는 이분 문항반응(dichotomous item response)을 중심으로 연구되었다(Baker & Kim, 2004). 이후 피험자들의 반응을 이분 반응으로 한정하지 않고 보다 세분화된 반응 즉, 다분 문항반응(polytomous item response)에 대한 필요성이 제기되었고 여러 연구자들에 의해 다분 문항반응이론(polytomous item response theory) 모형이 개발되어 왔다(Samejima, 1969; Andrich, 1978; Masters, 1982; Bock, 1972; Muraki, 1992).

1모수 다분 문항반응이론 모형으로는 Masters(1982)가 제안한 부분점수모형(Partial Credit Molel; PCM)과 Andrich(1978)의 평정척도모형(Rating Scale Model; RSM)이 대표적이다. 부분점수모형은 비교의 대상

이 되는 준거집단(reference group)이 없어도 점수를 해석할 수 있고, 동시에 특별한 의미를 갖는 '0'점을 확보할 수 있으며, 채점 시 모든 문항의 각 점수 급간별 거리가 동일하다는 가정 없이도 시용할 수 있다는 장점을 가진다(백순근, 1996). 그리고 평정척도모형은 문항범주 간 거리가 모든 문항에서 같다고 가정함으로써 부분점수모형에 비해 단순화된 모형으로 리커트(likert) 척도로 채점된 유형의 문항에 적용할 수 있고, 이후에 등장한 다국면 라쉬모형 역시 평정척도모형의 국면(facet)을 확장한 모형이다(Eckes, 2009).

2모수 다분 문항반응이론에서 가장 널리 활용되는 모형으로는 Samejima(1969)의 등급반응모형(Graded Response Model; 이하 GRM)과 Muraki(1992)의 일반화부분점수모형(Generalized Partial Credit Model; 이하 GPCM)이 있다. GRM은 서열화된 문항 반응에 적용할 수 있도록 Samejima가 제안한 다분 문항반응이론 모형으로 서열 변수인 다분 문항 반응 자료를 활용하여 문항과 피험자 모수를 추정할 수 있다. 그리고 GPCM은 Muraki가 Masters(1982)의 부분점수모형을 문항별로 서로 다른 변별도를 가질 수 있도록 확장시킨 모형으로 부분점수모형과 달리 범주곤란도 지수가 서열화되어 있다는 가정을 할 필요가 없어 문항을 제작할 때 범주난이도 지수에 대한 제한점을 갖지 않도록 함으로서 현실에서 많이 사용될 수 있다(박정, 2001).

가. 다분 문항반응이론 채점자 영향 모형

Samejima(1969)가 제안한 GRM은 피험자의 반응을 여러 가지 유형으로 분류하고, 각 유형에 대해 차등적인 점수를 부여하여 이를 토대로 피험자의 능력이나 속성을 추정하도록 제안된 수리적 모형으로, 예를 들면 대학생들의 학점과 같이 유목화된 채점이나, 태도에 대한 설문과 같은 리커트 척도, 부분 점수와 같은 채점 방식에 적용될 수 있다. 이 모형은 개별 문항에 대한 반응을 K개의 유형으로 나누고 개별 피험자의 능력이

나 속성에 따라서 선택하는 확률을 수리적으로 추정하는 모형이다. 이때, 변별도 α와 단계 곤란도 모수 β_{ik}를 가지는 문항 i에 대하여 능력(혹은 잠재특성) 수준이 θ인 피험자가 $k(k=1,2,\cdots,K)$번째 유형의 점수를 선택하거나 분류될 확률을 $P_{ik}(\theta)$라고 하면 GRM은 다음 식과 같이 표현된다.

$$P_{ik}{}^*(\theta)=\begin{cases}1 & (k=0)\\ \dfrac{1}{1+\exp\{-\alpha_i(\theta-\beta_{ik})\}} & (k=1,2,\cdots,K-1)\\ 0 & (k=K)\end{cases}$$

$$P_{ik}(\theta)=P_{i(k-1)}{}^*(\theta)-P_{ik}{}^*(\theta)$$

θ : 피험자의 능력수준
α_i : 문항 i의 변별도
β_{ik} : 문항 i의 단계 곤란도 모수
K : 문항의 범주수
k : 특정 범주 점수($1\sim K$)

등급화된 응답 모형을 적용하여 분석할 수 있는 문항은 하나의 문항에 대하여 다양한 수준의 점수가 부여되기 때문에 이분 문항의 문항특성곡선과는 달리 등급별로 문항의 특성을 동시에 보여준다. 따라서 다분반응 문항의 문항특성곡선은 문항반응유형특성곡선(item response category characteristic curve; IRCCC)라고 한다(Baker, 2004).

Muraki(1992)는 GRM에서 연결함수로 사용하는 누적범주로짓(cumulative category logit) 대신 이웃범주로짓(adjacent category logit)을 연결함수로 활용하여 다분 문항반응 자료를 다루기 위한 모형인 GPCM을 제안하였다(강태훈, 김명연, 2012). GPCM 하에서 다분 반응문항에 대하여 특정한 점수 $k(k=1,2,\cdots,K)$를 획득하기 위한 확률은 다음 식과 같이 표현될 수 있다.

$$P_{ik}(\theta)=\frac{\exp\sum_{h=0}^{k}\left[\alpha_i\{\theta-(\beta_i-s_{ih})\}\right]}{\sum_{h=0}^{K}\exp\sum_{g=0}^{k}\left[\alpha_i\{\theta-(\beta_i-s_{ig})\}\right]}$$

$$\sum_{g=0}^{0}\{\theta-(\beta_i-s_{ig})\}=0$$

θ : 피험자의 능력수준
α_i : 문항 i의 변별도
β_i : 문항 i의 곤란도 모수
s_{ik} : 문항 i의 k번째 단계 모수
K : 문항의 범주수
k : 특정 범주 점수($1\sim K$)

위 식에서 α_i는 문항 i의 변별도 혹은 기울기 모수이며, β_i-s_{ih}는 문항 i에서 $h(h=1,2,\cdots,K)$라는 점수를 받을 때의 어려운 정도를 의미하는 '문항 범주 곤란도' 혹은 '문항 단계 곤란도'이다. 그런데 다분 문항반응이론 모형에서 기울기 모수를 이분문항반응 모형에서와 같은 의미의 문항의 변별도라고 해석하는 것은 적절하지 않다. 다분 문항반응이론 모형에서의 기울기 모수는 범주 곤란도 모수들 간의 거리에도 영향을 받기 때문에 이분 문항반응이론 모형의 경우처럼 하나의 곤란도 모수를 갖는 경우와는 특성과 해석에 차이가 있다. 때문에 다분 문항반응이론 모형에서 범주 곤란도 모수는 능력 척도 상에서의 범주 간 경계치로 해석하고, 실제적인 변별 정도는 각 문항 범주 곤란도 모수들 간 간격에 따라서 달라진다.

이상의 다분 문항반응이론 모형을 바탕으로 Muraki(1993)는 채점자 특성의 영향을 추가한 여러 가지 다분 문항반응이론 모형을 제안하였다. Muraki(1993)가 제안한 다분 반응이론 모형들은 기존 다분 문항반응이론 모형을 다집단 상황으로 확장한 것으로 채점자 영향 모형(rater's effect model), 차별기능문항 모형(DIF model), 경향 모형(trend model)

등이 있다(Muraki & Bock, 2003).

Muraki(1993)의 모형들은 다집단 분석의 맥락에서 GPCM에 채점자에 따라 집단을 구분하고, 차별기능 효과 모수 d_{ji}를 추가하여 아래 식과 같이 집단 j의 문항 i에 대한 수리적 모형을 제시했다.

$$P_{jik}(\theta) = \frac{\exp \sum_{h=0}^{k} \left[\alpha_i \{ \theta - (\beta_i - s_{ih} + d_{ji}) \} \right]}{\sum_{h=0}^{K} \exp \sum_{g=0}^{k} \left[\alpha_i \{ \theta - (\beta_i - s_{ig} + d_{ji}) \} \right]}$$

θ : 피험자의 능력수준
α_i : 문항 i의 변별도
β_i : 문항 i의 곤란도 모수
s_{ik} : 문항 i의 k번째 단계 모수
K : 문항의 범주수
k : 특정 범주 점수($1 \sim K$)
d_{ji} : 문항 i에 대한 집단 j의 차별기능 효과 모수

채점자 영향 모형에서는 차별기능문항과 달리 채점자 영향이 문항과 독립적이기 때문에 문항에 따라 변화하는 d_{ji} 대신 문항에 무관하게 고정적인 d_j를 모형에 투입하여 모든 문항에 동일한 채점자 영향이 미치는 모형을 설정할 수 있다. 이에 따라 채점자 j의 문항 i에 대한 채점자 영향 모형은 다음과 같이 수리적으로 표현된다.

$$P_{jik}(\theta) = \frac{\exp \sum_{h=0}^{k} \left[\alpha_i \{ \theta - (\beta_i - s_{ih}) - d_j \} \right]}{\sum_{h=0}^{K} \exp \sum_{g=0}^{k} \left[\alpha_i \{ \theta - (\beta_i - s_{ih}) - d_j \} \right]}$$

$$\sum_{g=0}^{0} \{ \theta - (\beta_i - s_{ig}) - d_j \} = 0$$

나. 다국면 라쉬모형

채점자 영향을 분석하거나 피험자 능력 추정치에 대하여 채점자 엄격성 차이를 사후에 보정하는 대표적인 방법인 다국면 라쉬 모형(Many-Facet Rasch Model; Linacre, 1989)은 다양한 수행평가 및 논술형 검사 채점 상황에서 많이 활용되고 있다(이현숙, 2008, Engelhard et al., 2017). 다국면 라쉬 모형을 활용한 평정 척도 모형은 피험자의 능력과 무관하게 문항의 곤란도에 대한 모수가 추정되고 문항의 곤란도와 무관하게 피험자의 능력 모수가 추정되는 기존의 라쉬 모형에 채점자 엄격성과 같은 별도의 국면을 추가함으로써 채점자들의 엄격성과 무관하게 피험자의 능력과 문항의 곤란도가 추정되는 모형이다.

다국면 라쉬 모형을 활용하면 채점자의 엄격성이 직접적으로 측정되고, 별도의 척도 점수의 조정 절차 없이 피험자의 능력이 동시에 추정되어 척도화된다는 점에서 고전적인 표준화 방식과는 차별성을 가진다(Engelhard, 1996; Engelhard et al., 2017). 이러한 장점 때문에 다국면 라쉬 모형은 다양한 맥락에서 널리 활용되었고, 채점자간 엄격성 차이를 사후에 통계적으로 보정하는 방법으로도 여러 연구들에서 활용되어 왔다(김성숙, 2001; 이현숙, 2008; 지은림, 2008; 채선희, 1996; Engelhard, 1996; Linacre, 1994; Lunz, Wright, & Linacre, 1990).

다국면 라쉬 모형을 좀 더 구체적으로 살펴보면 이 모형은 기존 문항반응이론에 바탕을 두고 개발된 라쉬 모형 중 하나이다(Linacre, 1989). 기본적인 라쉬 모형에서는 문항과 피험자를 평가한 뒤 이를 각각 난이도와 능력의 차이에 따라 하나의 등간척도로 나열해준다. 기본적인 일차원 이분 반응 자료에 대한 라쉬 모형은 다음 수식과 같이 표현된다.

$$\log\left(\frac{P_{in}}{1-P_{in}}\right)=B_n-D_i$$

P_{in} : 피험자 n이 문항 i를 맞힐 확률

B_n : 피험자 n의 능력 수준

D_i : 문항 i의 곤란도

라쉬 모형에서 피험자의 능력 추정치는 사용된 특정 문항 표본에 따라 그 값이 달라지지 않으며, 문항의 곤란도 추정치 또한 특정 피험자 표본으로부터 영향을 받지 않는다. 나아가 문항 곤란도와 피험자 능력만을 분석할 수 있었던 기존 라쉬 모형과 달리, 다국면 라쉬 모형은 채점자 엄격성을 비롯한 다양한 국면들까지 동시에 그리고 상호 독립적으로 분석하여 하나의 척도 상에 제시해준다(Linacre, 1989; McNamara, 1996; Weigle, 1994). 기존 이분 반응 자료에 대한 라쉬 모형을 다분 반응 자료로 확장하고 채점자 특성 영향을 추가한 다국면 라쉬 모형을 수식으로 표현하면 다음과 같다(Linacre, 2008).

$$\log\left(\frac{P_{ijkn}}{P_{ij(k-1)n}}\right) = B_n - D_i - C_j - F_k$$

P_{ijkn} : 피험자 n이 문항 i에 대하여 채점자 j로부터 점수 k를 받을 확률

$P_{ij(k-1)n}$: 피험자 n이 문항 i에 대하여 채점자 j로부터 점수 $(k-1)$을 받을 확률

B_n : 피험자 n의 능력 수준

D_i : 문항 i의 곤란도

C_j : 채점자 j의 엄격성 수준

F_k : 점수 k-1에 비해 점수 k를 받기 어려운 정도(단계 모수)

위의 식에서 나타나는 바와 같이 다국면 라쉬 모형은 피험자의 능력 수준, 문항의 곤란도, 채점자 엄격성 정도를 동일한 척도에서 수치화할

수 있는 모형이다. 이 때 피험자 능력 수준의 평균, 문항 곤란도의 평균, 채점자 엄격성 수준의 평균을 척도상의 0점에 고정시키면 문항 곤란도와 채점자 엄격성의 영향력이 통제된 상태에서 피험자 능력에 대한 로짓 점수를 얻을 수 있다. 다국면 라쉬 모형은 이와 같은 특성을 가지고 있기 때문에 피험자의 답안을 관대하거나 혹은 엄격한 채점자에 의해 평정되더라도 채점자 엄격성에 의한 영향력이 보정된 피험자 능력 모수를 추정할 수 있다는 장점을 가진다.

한편, 한 피험자의 답안을 여러 채점자가 점수를 부여하는 경우, 이 점수들은 한 답안에 대한 채점 결과이므로 서로 높은 상관을 가질 수밖에 없다. 하지만 다국면 라쉬 모형과 앞 절에서 살펴본 채점자 영향 모형에서는 이와 같이 채점이 답안에 내재되는 구조를 반영하지 못하고, 한 답안에 대한 채점들을 서로 독립적인 것으로 다루기 때문에 피험자 능력 추정에서 심각한 편향이 발생하게 된다(Donoghue & Hombo, 2000; Patz, 1996; Patz, Junker, Johnson & Mariano, 2002; Verhelst & Verstralen, 2001; Wilson & Hoskens, 2001). 또한 다국면 라쉬 모형은 채점자 효과에 대해서 채점자 엄격성만을 알려줄 수 있고 채점자가 서로 다른 답안들을 변별해내는 정도에 대한 정보를 제공할 수 없다는 한계를 가진다. 이러한 채점자 모형들의 한계를 극복하기 위한 대안 제시된 모형들 중 하나가 위계적 채점자 모형이다.

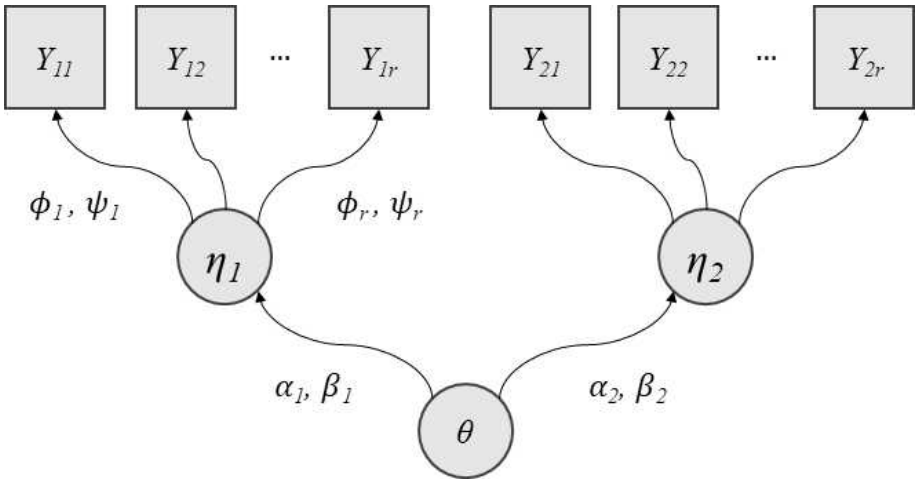
2. 위계적 채점자 모형

Patz(1996), Patz & Junker(1999), Patz, Junker, Johnson, & Mariano(2002)에 의해 제안된 위계적 채점자 모형(Hierarchical Rater Model; 이하 HRM)에서는 피험자에 대한 여러 채점자의 다분 채점 결과를 위계적 자료 구조로 파악한다. HRM에서 채점자가 특정 피험자의 특정 문항응답에 부여한 점수는 문항반응이론 모형과 달리 피험자의 능력을 나타내는 직접 지표가 되지 않고, 채점 기준(rubric)으로 정의되는 진

점수(true score) 혹은 이상 평정(ideal rating)에 대한 지표이다(Patz, Junker, Johnson, & Mariano, 2002).

HRM의 1수준 모형은 채점자의 관찰된 실제 평정 결과(Y_{ir})로 구성된다. 그런데 각 피험자의 응답에 대한 채점자의 이상 평정(η_i)은 잠재변수이기 때문에 관찰 가능하거나 알 수 있는 것이 아니다. 따라서 HRM의 맥락에서 채점자의 역할은 주어진 피험자의 응답이 어떤 수준에 속하는지를 결정하는 것으로, 이상 평정에 대한 신호 탐지(signal detection)라 할 수 있다. 따라서 채점자의 채점 자료를 포함하는 HRM의 1수준 모형은 신호 탐지 모형(signal detection model; SDM)으로 볼 수도 있다(DeCarlo et al., 2011). 이러한 HRM의 접근에서는 채점자의 판단(실제 평정)이 이상 평정에 대한 오류 가능성을 가진 판단으로 보고, 채점자가 이상 평정을 설제 평정으로 정확하게 탐지하는 정도를 채점자 엄격성(severity) 또는 관대성(leniency)과 채점자 일관성(consistency) 또는 비일관성(inconsistency)으로 측정한다(Barr & Raju, 2003; DeCarlo et al., 2011).

2수준 모형에서는 기존의 다분 문항반응이론 모형을 활용하는데, 일반적인 1수준 다분 문항반응이론 모형에서는 관찰 점수가 피험자의 능력 수준에 대한 지표로 활용되었다면, HRM의 2수준 모형에서는 잠재 변수인 이상 평정이 피험자의 능력 수준에 대한 지표로 활용된다. 따라서 피험자의 응답은 채점자를 통해 채점 기준에 정의된 평정 범주에 속하게 되고, 그 범주가 피험자의 능력 수준을 나타내는 지표가 된다. 이와 같은 지표 및 추정치들 사이의 관계를 모형으로 나타내면 다음 [그림 Ⅱ-1]과 같고, 능력수준이 θ인 피험자의 i번 문항에 대한 답안에 r번째 채점자가 평정한 점수 Y_{ir}은 실제로는 알 수 없는 잠재변수인 이상 평정 η_i에 대한 겉으로 드러난 지표이고, 이상 평정 η_i는 피험자의 능력수준인 θ에 대한 지표가 된다.



[그림 Ⅱ-1] Patz et al.(2002)의 HRM 모형

위 그림에서 제시된 상황에서 채점자가 3명($R=3$)이라면 각 피험자가 2개의 문항에 대하여 응답하고, 각 응답에 대하여 3명의 채점자가 채점을 하여 총 6개의 채점 결과를 가지는 HRM을 나타낸다. 채점 결과는 서열 반응으로 관찰되고, Y_{ir}는 i번째 문항에 r번째 채점자의 대한 채점 결과를 나타낸다. 채점자들은 Y_{ir}을 통해서 두 개의 문항에 대한 각각의 응답에 대한 이상 평정 η_1, η_2을 탐지한다고 볼 수 있고, η_i에서 Y_{ir}로 향하는 화살표가 곡선인 것은 Y_{ir}와 η_i간의 관계가 비선형(nonlinear)적인 관계임을 나타낸다. 이상의 1수준 모형의 관계를 Patz et al.(2002)에서는 다음 수식과 같이 모형을 수리적으로 표현하였다.

$$P(Y_{ir}=k|\eta_i)=\frac{\exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}}{\sum_{m=1}^{K}\exp\left\{-\frac{1}{2\psi_r^2}\left[m-(\eta_i+\phi_r)\right]^2\right\}}$$

Y_{ir} : 문항 i에 대한 채점자 r의 실제 평정 결과(1~K)
η_i : 문항 i에 대한 이상 평정(잠재 변수, 1~K)
i : 문항
r : 채점자
K : 문항의 채점 범주수
k : 특정 범주 점수(1~K)
ϕ_r : 채점자 r의 엄격성 모수
ψ_r : 채점자 r의 분산성 모수

1수준 모형에서 확률 함수의 분모인 $\sum_{m=1}^{K} \exp\left\{-\frac{1}{2\psi_r^2}\left[m-(\eta_i+\phi_r)\right]^2\right\}$는 k와 무관하게 고정적인 값이기 때문에 상수로 취급하고 아래와 같이 비례 관계로 표현하기도 한다.

$$P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}$$

위 식에서 e의 지수부는 정규분포 밀도 함수의 지수부와 동일함을 알 수 있고, 관찰된 실제 평정 Y_{ir}이 이상 평정 η_i로부터 ϕ_r만큼 이동된 평균 $(\eta_i+\phi_r)$를 중심으로 분산 ψ_r^2을 갖는 정규 분포를 이산적으로 따르는 것으로 이해할 수 있다. 따라서 Patz et al.(2002)이 제안한 HRM의 1수준 모형은 문항 i에 대한 채점자 r의 실제 평정 Y_{ir}가 이상 평정 η_i로부터 얼마나(ϕ_r) 이동(편향)하고, 이 지점으로부터 얼마나(ψ_r) 흩어져 분포(분산)하는가를 추정하는 모형이다. 두 채점자 모수 ϕ_r, ψ_r에 대해서는 연구자마다 서로 다른 다양한 표현을 사용하고, 한 연구자도 연구마다 조금씩 다른 표현을 사용하고 있기에 본 연구 내에서는 이후 HRM을 제안한 Patz et al.(2002)가 최초로 사용한 표현을 따라 '채점자 엄격성(rater severity)'과 '채점자 분산성(rater variability)'이라고 표현하겠다.

HRM의 1수준 모형에 포함된 채점자 모수가 이상 채점(η)이 실제 채

점(Y)에 어떻게 영향을 미치는지 확인하기 위하여 채점자 모수에 따른 이상 채점과 실제 채점 확률의 관계를 [그림 Ⅱ-2], [그림 Ⅱ-3]과 같이 제시하였다.

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	99.61	.39	.00	.00
1	.38	99.23	.38	.00
2	.00	.38	99.23	.38
3	.00	.00	.39	99.61

(A1) 4범주, ϕ=0, ψ=0.3

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	99.61	.39	.00
1	.38	99.23	.38
2	.00	.39	99.61

(B1) 3범주, ϕ=0, ψ=0.3

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	90.22	9.78	.00	.00
1	.01	90.21	9.78	.00
2	.00	.01	90.21	9.78
3	.00	.00	.01	99.99

(A2) 4범주, ϕ=0.3, ψ=0.3

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	90.22	9.78	.00
1	.01	90.21	9.78
2	.00	.01	99.99

(B2) 3범주, ϕ=0.3, ψ=0.3

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	50.00	50.00	.00	.00
1	.00	50.00	50.00	.00
2	.00	.00	50.00	50.00
3	.00	.00	.00	100.00

(A3) 4범주, ϕ=0.5, ψ=0.3

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	50.00	50.00	.00
1	.00	50.00	50.00
2	.00	.00	100.00

(B3) 3범주, ϕ=0.5, ψ=0.3

[그림 Ⅱ-2] 채점자 엄격성 모수에 따른 채점 확률

엄격성 모수와 분산성 모수는 함께 실제 채점 확률에 영향을 미치기 때문에 분산성 모수가 .3으로 동일할 때, 문항의 채점 범주가 4개인 경우(A1~3)와 채점 범주가 3개인 경우(B1~3)에 대하여 엄격성 모수에 따른 실제 채점 확률 변화를 비교하였다. 엄격성 모수 ϕ가 클수록(채점자가 관대할수록) 이상 채점 η보다 더 높은 실제 채점 Y를 획득할 확률이 커지고 ϕ가 .5에 이르면 정확하게 이상 채점과 동일한 실제 채점 확률과

인접 범주로 채점할 확률이 같게 된다. 이러한 채점자 특성과 채점 분포와의 관계를 역으로 해석하자면 채점자가 이상 채점 η보다 높은 점수로 평정하는 경향이 있다면 채점자 엄격성 모수는 0보다 크고, 이상 채점 η보다 낮은 점수로 평정하는 경향이 있다면 채점자 엄격성 모수는 0보다 작은 값으로 추정된다.

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	99.61	.39	.00	.00
1	.38	99.23	.38	.00
2	.00	.38	99.23	.38
3	.00	.00	.39	99.61

(A1) 4범주, ϕ=0, ψ=0.3

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	99.61	.39	.00
1	.38	99.23	.38
2	.00	.39	99.61

(B1) 3범주, ϕ=0, ψ=0.3

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	88.05	11.92	.03	.00
1	10.65	78.68	10.65	.03
2	.03	10.65	78.68	10.65
3	.00	.03	11.92	88.05

(A2) 4범주, ϕ=0, ψ=0.5

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	88.05	11.92	.03
1	10.65	78.70	10.65
2	.03	11.92	88.05

(B2) 3범주, ϕ=0, ψ=0.5

η	P(Y|η)			
	Y=0	Y=1	Y=2	Y=3
0	72.60	26.17	1.23	.01
1	20.74	57.54	20.74	.97
2	.97	20.74	57.54	20.74
3	.01	1.23	26.17	72.60

(A3) 4범주, ϕ=0, ψ=0.7

η	P(Y|η)		
	Y=0	Y=1	Y=2
0	72.60	26.17	1.23
1	20.95	58.11	20.95
2	1.23	26.17	72.60

(B3) 3범주, ϕ=0, ψ=0.7

[그림 Ⅱ-3] 채점자 분산성 모수에 따른 채점 확률

채점자 분산성 모수에 따른 이상 채점과 실제 채점간의 관계에 대한 [그림 Ⅱ-3]을 살펴보면 문항의 채점 범주가 4개인 경우(A1~3)와 채점 범주가 3개인 경우(B1~3) 모두에서 분산성 모수인 ψ값이 커질수록 이상 채점이 η일 때, 정확한 η로 실제 채점할 확률은 낮아지고 인접한 점

수로 실제 채점할 확률이 점차 증가하여 채점의 일관성이 떨어지는 채점 결과를 확인할 수 있다. 채점 분포와 채점자 특성과의 관계로 본다면 채점자가 주로 이상 채점 η에 엄격성만큼 조정된 점수와 거의 일치하는 평정을 했다면 채점자 분산성 모수는 0에 가깝게 추정된다. 반대로 채점자가 이상 채점 η에 엄격성만큼 조정된 점수로부터 차이가 큰 평정을 많이 할수록 채점자 분산성 모수는 큰 값을 갖게 된다.

HRM의 2수준 모형은 일반적으로 기존의 다분 문항반응이론 모형인 PCM이나 GPCM이 활용되는데, 관찰된 채점자의 평정 Y_{ir}를 활용하여 피험자의 반응 확률 $P(Y_i = k|\theta)$를 모형화하지 않고, 1수준 모형을 통해 추정되는 잠재변수인 채점자의 이상 평정 η_i을 획득할 확률 $P(\eta_i = k|\theta)$로 다분 문항반응이론 모형을 추정한다는 것이 차이점이다. HRM의 2수준 모형에서는 [그림 Ⅱ-1]에 제시된 바와 같이 이상 평정 η_i가 피험자 능력 모수 θ에 대한 서열 지표로 역할하고, 이상 평정 η_i의 확률과 피험자 능력 모수 θ 사이의 관계 역시 연결함수를 활용하는 문항반응이론의 특성에 따라서 비선형적인 관계이기 때문에 화살표가 곡선으로 표현되었다. 2수준 모형의 모수들은 GPCM을 적용할 경우, 변별도 모수 α_i와 곤란도 모수 β_{im}(본 연구의 실제 분석에서는 문항 곤란도 모수와 단계 모수들을 각각 추정하지 않고 문항 곤란도 모수 β_i에 단계 모수 s_{im}을 뺀 β_{im}으로 추정하였기에 앞서 이론적 배경의 GPCM에서 다룬 표기 대신 분석에 사용한 모수를 따라 표기)를 사용하여 다음과 같은 다분 문항반응이론 모형과 동일한 형태의 수식으로 표현된다.

$$P(\eta_i = k|\theta) = \frac{\exp \sum_{m=1}^{k} \alpha_i(\theta - \beta_{im})}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \alpha_i(\theta - \beta_{im})}$$

한편, DeCarlo(2002)는 Patz et al.(2002)의 HRM에서 1수준 모형을 달리하여 다음과 같은 HRM-SDT를 제안하였다.

$$P(Y_{ir}=k|\eta_i)=P(Y_{ir}\geq k-1|\eta_i)-P(Y_{ir}\geq k|\eta_i)$$

$$\begin{cases} P(Y_{ir}\geq k|\eta_i)=\dfrac{1}{1+\exp(c_{rk}-d_r\eta_i)} \ (k=1,\ldots K-1) \\ P(Y_{ir}\geq 0|\eta_i)=1 \\ P(Y_{ir}\geq K|\eta_i)=0 \end{cases}$$

Y_{ir} : 문항 i에 대한 채점자 r의 실제 평정 결과(1~K)
η_i : 문항 i에 대한 이상 평정(잠재 변수, 1~K)
i : 문항
r : 채점자
K : 문항의 채점 범주수
k : 특정 범주 점수(1~$K-1$)
c_{rk} : 채점자 r의 k번째 범주 경계 모수
d_r : 채점자 r의 변별도 모수

Patz et al.(2002)의 HRM에서는 실제 평정 Y_i가 엄격성 모수 ϕ_r의 고정효과로 평균적으로 이상 채점 η_i으로부터 벗어나고, 그 지점을 기준으로 분산성 모수 ψ_r의 무선효과로 흩어져 분포할 확률을 모형화했다면 DeCarlo(2002)의 HRM-SDT는 고정효과인 변별도 모수 d_r와 (범주수)-1개의 범주 간 경계 모수 c_{rk}를 통해 채점 범주마다 이상 채점 η_i으로부터 실제 평정 Y_i가 벗어날 확률을 모형화한 것으로 이해할 수 있다.

3. 위계적 채점자 모형의 확장

DeCarlo et al.(2011)는 채점자 모수가 문항에 따라 다른 값을 가질 수 있도록 DeCarlo(2002)의 기존 HRM-SDT의 1수준 모형을 다음과 같이 확장하였다.

$$P(Y_{ir}=k|\eta_i)=P(Y_{ir}\geq k-1|\eta_i)-P(Y_{ir}\geq k|\eta_i)$$

$$\begin{cases} P(Y_{ir}\geq k|\eta_i)=\dfrac{1}{1+\exp(c_{irk}-d_{ir}\eta_i)}\ (k=1,\ldots K-1) \\ P(Y_{ir}\geq 0|\eta_i)=1 \\ P(Y_{ir}\geq K|\eta_i)=0 \end{cases}$$

Y_{ir} : 문항 i에 대한 채점자 r의 실제 평정 결과(1～K)
η_i : 문항 i에 대한 이상 평정(잠재 변수, 1～K)
i : 문항
r : 채점자
K : 문항의 채점 범주수
k : 특정 범주 점수(1～K)
c_{irk} : 문항i에 대한 채점자 r의 k번째 범주 경계 모수
d_{ir} : 문항i에 대한 채점자 r의 변별도 모수

기존 HRM-SDT와의 차이는 채점자별로 추정되던 범주 경계 모수 c_{rk}와 변별도 모수 d_k를 채점자-문항별로 추정하도록 각각 c_{irk}와 d_{ik}로 변경하였다는 점이다. Patz et al.(2002)의 HRM 모형을 재모수화하려는 본 연구에서도 DeCarlo et al.(2011)의 아이디어와 같은 방식으로 채점자 모수를 문항별로 추정할 수 있도록 재모수화를 하고자 하였다. 이러한 아이디어는 완전히 새로운 것은 아니고, 이미 Robitzsch & Steinfeld (2018) 연구에서 소개한 R 소프트웨어의 immer package에서 다음 수식과 같이 HRM의 1수준(채점자 수준) 모형의 모수의 제약을 풀어 ϕ_{ir}과 ψ_{ir}를 추정할 수 있도록 분석 가능성을 열어두었다.

$$P(Y_{ir}=k|\eta_i)\propto\exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k-(\eta_i+\phi_{ir})\right]^2\right\}$$

그러나 Metropolis-Hastings 알고리즘을 토대로 Robitzsch가 2018년

에 개발한 HRM 분석 코드는 ϕ_r 또는 ψ_r의 제약을 풀어서 추정할 경우, package에 포함된 채점 자료로도 정상적인 분석 결과를 얻을 수 없고, Robitzsch가 HRM에 대한 후속 연구를 진행하지 않고 있어 채점자-문항 독립성 제약을 푼 모형이 제대로 모수의 추정이 가능한지, 그리고 추정의 정확성이 어느 정도 수준인지 확인할 수 없었다. 따라서 본 연구에서는 채점자-문항 독립성 제약을 완화하여 확장한 모형을 '확장된 위계적 채점자 모형(Extended Hierarchical Rater Model; E-HRM)'으로 설정하고, 모형이 실제 문항과 피험자, 그리고 채점자의 특성을 정확하게 분석할 수 있는지에 대해 적용 연구와 모의실험 연구를 진행하고자 한다.

HRM의 1수준(채점자 수준) 모형에서 채점자-문항 독립성 제약을 완화하는 방안은 총 세 가지가 가능하다. 엄격성 모수의 확장, 분산성 모수의 확장, 두 모수 모두를 확장하는 방안이 그것이다. 이에 따라 연구 모형은 네 가지를 설정하였는데, 먼저 채점자-문항 독립성 제약을 완화한 모형들과 비교를 위해 Patz et al.(2002)의 HRM(M1)을 기저모형으로 사용하였고, HRM과 2수준(피험자 수준) 모형은 GPCM으로 동일하지만 채점자 분산성 모수 ψ_r의 제약만을 없애고 문항별로 채점자의 분산성 모수 ψ_{ir}을 추정하는 모형을 M2, 채점자 엄격성 모수 ϕ_r의 제약만을 완화하고 문항별 채점자 엄격성 모수 ϕ_{ir}을 추정하는 M3, 마지막으로 채점자 엄격성 모수와 분산성 모수 모두의 제약을 제거하여 ϕ_{ir}, ψ_{ir}을 추정하는 모형을 M4로 설정하였다. 이 모형들의 1수준 모형을 기술하면 아래 네 가지 모형이 된다.

$$\text{M1} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}$$

$$\text{M2} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}$$

$$\text{M3 : } P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_{ir})\right]^2\right\}$$

$$\text{M4 : } P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k-(\eta_i+\phi_{ir})\right]^2\right\}$$

이상의 모형들 중 본 연구에서 제안하고자 하는 모형은 채점자 분산성 모수 ψ_r의 제약만을 완화하여 문항별 채점자 분산성 모수 ψ_{ir}을 추정하는 M2이고, 이를 E-HRM으로 명명하였다. 우선 채점자 엄격성 모수와 분산성 모수 모두의 제약을 제거하여 ϕ_{ir}, ψ_{ir}을 추정하는 M4의 경우, 본 연구에서 관심을 가지는 것과 같이 다수의 구성형 문항을 포함하는 검사 상황에서는 하나의 채점자 모수 제약만을 완화하는 M2나 M3에 비하여 추정할 채점자 모수가 지나치게 많아져 추정의 어려움이 예상된다. 때문에 채점자 엄격성 모수와 분산성 모수 중 하나의 모수에 대해서만 제약을 완화하는 것이 추정 가능성이 높은 모형이라고 판단된다.

한편, 채점자 엄격성 모수와 분산성 모수를 앞서 2절에서 살펴본 바와 같이 상호의존적인 관계를 가지면서 실제 채점이 이상 채점으로부터 벗어날 확률을 결정하게 된다. 이와 관련하여 선행연구에서 밝혀진 바에 따르면 HRM 분석에서는 엄격성 모수 추정이 분산성 모수에 영향을 받는 정도가 크고, 특히 분산성 모수가 작을 경우에 엄격성 모수의 추정이 불안정해지는 것으로 알려져 있다(Patz et al., 2002; Casabianca et al., 2017). 또한 엄격성 모수와 분산성 모수가 실제 채점의 확률에 미치는 영향을 보여주는 [그림 Ⅱ-2], [그림 Ⅱ-3]을 살펴보면 실제 대규모 표준화 검사에서 관찰하기 어려운 .5에 가까운 큰 편향이 있는 경우가 아니라면, 엄격성 모수가 HRM의 1수준 모형에서 관찰된 실제 채점의 확률 변화에 미치는 영향은 분산성 모수에 비하여 상대적으로 작을 것이라고 판단된다.

모수 추정 가능성의 측면 외에도 E-HRM을 연구의 제안 모형으로 삼은 근거로는 Lunz & Stahl(1990)와 DeCarlo et al.(2011)의 연구 결과가

있었다. Lunz & Stahl(1990)에서는 동일한 피험자에 대한 세 가지 방법의 시험 답안을 채점한 자료를 FACET 모형으로 분석한 결과, 채점자 간에는 엄격성의 차이가 있었지만 대부분의 채점자가 세 가지 시험 방법에 대한 일관된 엄격성을 보였다. 또한 DeCarlo et al.(2011)의 연구에서는 일부 5점 척도의 2개 문항에 대한 13명의 채점자의 문항 간에 엄격성을 비교하였는데, 일부 범주 경계 모수에서 나타나는 엄격성은 문항 간에 차이를 보였으나 문항 전체에서 평균적인 채점자의 엄격성은 그다지 큰 차이를 보이지 않았다. 이러한 연구 결과들은 채점자의 엄격성이 시험 방법 혹은 문항과 독립적으로 기능할 가능성을 보여주고, E-HRM에서 엄격성 모수의 채점자-문항 독립성 제약을 유지하는 것이 채점자 엄격성의 문항 독립적인 특성을 반영할 수 있다고 판단하였다.

이상의 HRM 및 확장 모형들과 HRM-SDT를 포함하여 HRM 계열 모형별로 추정하는 채점자 모수의 수는 <표 Ⅱ-1>과 같다.

<표 Ⅱ-1> HRM 계열 모형별 채점자 모수 수

모형		번호	채점자-문항 독립성제약	채점자 모수 수
HRM	Patz et al. (2002)	M1	제약	$2\times R$
E-HRM	ψ_{ir} 확장	M2	부분완화	$R\times(I+1)$
대안모형	ϕ_{ir} 확장	M3	부분완화	$R\times(I+1)$
	ϕ_{ir}, ψ_{ir} 확장	M4	없음	$2\times R\times I$
HRM-SDT	DeCarlo (2002)		제약	$R\times K$
	DeCarlo et al. (2012)		없음	$R\times I\times K$

(I : 문항수, R : 채점자수, K : 범주수)

<표 Ⅱ-1>의 모형별 채점자 모수 산출 식에 따라서 본 연구에서 활용하고자 하는 PIRLS 2006에서 채점자가 가장 적은 러시아(2명)와 채점자가 가장 많은 마케도니아(22명)에 적용해보면, <표 Ⅱ-2>와 같다.

<표 Ⅱ-2> PIRLS 2006 자료 분석 시 모형별 채점자 모수 수 (러시아, 마케도니아)

모형		번호	채점자 모수 수	
			러시아	마케도니아
HRM	Patz et al. (2002)	M1	4	44
E-HRM	ψ_{ir} 확장	M2	60	660
대안모형	ϕ_{ir} 확장	M3	60	660
	ϕ_{ir}, ψ_{ir} 확장	M4	116	1,276
HRM-SDT	DeCarlo (2002)		8	88
	DeCarlo et al. (2012)		184	2,024

<표 Ⅱ-2>에서 제시된 채점자 모수 수는 채점자-문항 설계가 완전구획설계로 구성되어 있어 모든 채점자가 모든 문항을 채점하는 상황에서 이론적으로 계산된 모수의 수이다. E-HRM과 대안모형의 경우, 채점자가 특정 문항에 대하여 채점을 하지 않았다면 해당 채점자-문항에 대한 모수가 추정에서 제외되어 실제 추정되는 모수는 감소한다.

모형 간의 모수 수 차이를 살펴보면 채점자-문항 독립성을 가정하고, 채점자별 무선효과로 채점자 분산성을 추정하는 HRM이 추정하는 채점자 모수 수가 가장 적은 것으로 나타났다. 반면에 가장 복잡한 모형은 DeCarlo et al.(2012)에서 제안한 채점자-문항 독립성 제약을 제거한 HRM-SDT로 이 모형에서는 채점자 모수가 채점자 수, 문항 수, 범주 수에 따라 승산적으로 증가하기 때문이다. 따라서 채점자-문항 독립성 제약을 제거한 HRM-SDT는 1~3개의 논술형 문항으로 구성된 검사 상황이 아닌 PIRLS와 같이 다수의 구성형 문항을 포함하는 대규모 표준화 검사에 활용하기에는 모수 추정의 부담이 지나치게 큰 모형이라는 것을 확인할 수 있다.

그리고 HRM-SDT(DeCarlo et al., 2012)와 같이 채점자-문항 독립성 제약을 제거하여 채점자 엄격성 모수와 분산성 모수를 채점자-문항 단위에서 모두 추정하는 대안모형 M4의 경우에도 HRM-SDT(DeCarlo et

al., 2012)보다는 적은 수의 채점자 모수를 갖지만 E-HRM이나 대안모형 M3에 비해서는 2배에 가까운 모수수를 포함하기 때문에 모수 추정이 쉽지 않을 것임을 예상할 수 있다.

마지막으로 본 연구의 제안 모형인 E-HRM과 대안모형 M3는 HRM나 HRM-SDT(DeCarlo, 2002)보다는 많은 채점자 모수를 가지지만 HRM-SDT(DeCarlo et al., 2012)이나 대안모형 M4보다는 적은 채점자 모수를 가지게 된다. 이러한 모수 수 비교에서 드러나듯 E-HRM의 목적은 HRM에 비해 모형의 간명성은 다소 희생하되, 최대한 상세하고 다양한 채점자 정보를 실제적으로 추정이 가능한 범위 내에서 획득하는 것이다.

Ⅲ. 연구방법

1. 적용 연구 대상

가. PIRLS 2006 채점 자료

PIRLS는 읽기 역량에 대한 국제비교연구로 TIMSS와 함께 국제 교육성취도 평가 협회(이하 IEA)의 핵심 연구 중 하나이다. Boston College에 위치한 TIMSS 및 PIRLS 국제 연구 센터의 주관으로 2001년부터 5년마다 실시되어 온 PIRLS는 4학년 학생들의 읽기 성취도 추세를 평가한다. 평가 대상이 4학년인 이유는 이 시기를 어린이의 읽기 발달의 핵심 전환 단계로 보고 있기 때문이고(Mullis, Kennedy, Martin, & Sainsbury, 2004), 이 단계에서 읽기 성취도를 평가하여 교육자와 정책 입안자에게 정보를 제공하고 개선점을 찾는 것이 PIRLS의 목적이다.

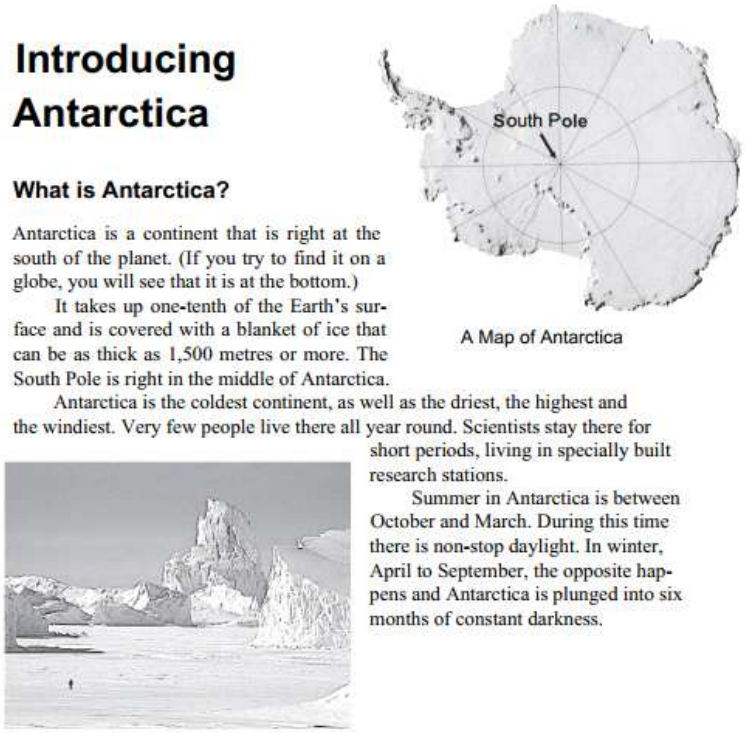
PIRLS 2006에는 40개국에서 7,762개 학교 단위로 219,740명의 학생들이 참여하였고, 이중 몇 개 국가는 지역별로 따로 참여하거나 5학년을 별도의 집단으로 추가하기도 하였다(Mullis et al., 2004). 전체 참여국 목록과 참여 학생 규모는 다음 <표 Ⅲ-1>과 같다.

<표 Ⅲ-1> PIRLS 2006 참여국가별 학교수 및 학생수

국가	학교수	학생수	국가	학교수	학생수
오스트리아	158	5,067	카타르	146	4,701
불가리아	137	4,479	루마니아	178	5,101
대만	150	4,552	러시아	150	4,002
덴마크	143	3,863	싱가포르	150	4,036
프랑스	150	4,243	슬로바키아	159	3,249
조지아	148	4,150	슬로베니아	139	4,156
독일	201	4,436	남아공	243	6,256
홍콩	180	3,988	스페인	135	3,837

국가	학교수	학생수	국가	학교수	학생수
헝가리	185	3,748	스웨덴	148	4,854
아이슬란드	150	4,589	트리니다드 토바고	119	6,680
인도네시아	145	4,001	마케도니아	146	4,273
이란	148	4,036	미국	232	4,720
이스라엘	169	4,404	영국	130	3,775
이탈리아	149	4,402	스코틀랜드	178	6,390
쿠웨이트	405	7,899	벨기에(네덜란드어권)	167	5,380
라트비아	144	4,712	벨기에(프랑스어권)	145	5,337
리투아니아	149	4,068	캐나다(온타리오)	429	16,073
룩셈부르크	128	3,673	캐나다(퀘벡)	152	4,094
몰도바	168	4,774	캐나다(앨버타)	147	4,394
모로코	236	5,411	캐나다(브리티시컬럼비아)	147	3,951
네덜란드	149	3,908	캐나다(노바스코샤)	183	5,190
뉴질랜드	150	3,581	아이슬란드(5학년)	35	1,379
노르웨이	149	3,958	노르웨이(5학년)	66	1,808
폴란드	147	4,162	전체	7,762	219,740

평가 문항 구성에 있어서는 총 126개 문항이 활용되었는데, 이 중에서 채점자를 필요로 하는 구성형 문항은 62개 문항이었고, 0~1점 배점으로 이분 반응을 가지는 28개 문항을 제외하면 본 연구에서 관심은 가지는 다분 반응 구성형 문항은 34개이다. 문항들은 각각이 독립적인 문항으로 제시되지 않고, 하나의 주제와 관련된 몇 개의 글로 구성된 제시 글을 먼저 제공하고 이와 관련된 내용을 묻는 여러 개의 선다형 문항, 이분 반응 구성형 문항, 다분 반응 구성형 문항들이 제시된다. 실제 검사에 포함된 문항 예시로 PIRLS 2006의 공개 문항 중 “Antarctica” 주제에 대한 제시 글의 일부와 다분 반응 구성형 문항들을 살펴보면 다음 [그림 Ⅱ-1], <표 Ⅲ-2>와 같다.



Introducing Antarctica

What is Antarctica?

Antarctica is a continent that is right at the south of the planet. (If you try to find it on a globe, you will see that it is at the bottom.)

It takes up one-tenth of the Earth's surface and is covered with a blanket of ice that can be as thick as 1,500 metres or more. The South Pole is right in the middle of Antarctica.

Antarctica is the coldest continent, as well as the driest, the highest and the windiest. Very few people live there all year round. Scientists stay there for short periods, living in specially built research stations.

Summer in Antarctica is between October and March. During this time there is non-stop daylight. In winter, April to September, the opposite happens and Antarctica is plunged into six months of constant darkness.

A Map of Antarctica

[그림 Ⅲ-1] PIRLS 2006 문항구획 Antarctica 지시문 일부

<표 Ⅲ-2> PIRLS 2006 문항구획 Antarctica 지시문 관련 문항 일부

문항코드	문항내용
R011A04C	Think about what the article says about Antarctica. Give two reasons why most people who visit Antarctica choose not to go there between April and September.
R011A07C	Give three ways penguins are able to keep warm in Antarctica.
R011A08C	What are two things you learn about food in Antarctica from Sara's letter?
R011A09C	Think about whether you would like to visit Antarctica. Use what you have read in both Introducing Antarctica and A Letter from Antarctica to explain why you would or would not like to visit.

<표 Ⅲ-2>에 제시된 문항들은 Antarctica에 포함된 11개 문항 중에서 다분 반응 구성형 문항들로 3개 문항(R011A04C, R011A08C, R011A09C)은 0~2점의 3개 범주로 채점되고, R011A07C 문항은 0~3점의 4개 범주

로 채점된다. 문항별 채점을 위해서는 구체적인 채점 기준(rubric)은 <표 Ⅲ-3>과 같이 제시되고, 채점자들은 이러한 기준을 활용하여 피험자들의 답안을 채점하게 된다.

<표 Ⅲ-3> PIRLS 2006 정보 획득 및 활용 영역 R011A07C 채점 기준

점수	기준 설명
3점	These responses demonstrate extensive comprehension by identifying most of the ideas in the article from which penguins' ability to stay warm can be inferred. The response describes at least three of the ways penguins are able to stay warm listed below.
2점	These responses demonstrate satisfactory comprehension by identifying some of the ideas in the article from which penguins' ability to stay warm can be inferred. The response describes two of the ways penguins are able to stay warm listed below.
1점	These responses demonstrate limited comprehension by identifying one idea in the article from which penguins' ability to stay warm can be inferred. The response describes only one of the ways penguins are able to stay warm listed below.

<Ideas from Article Explaining how Penguins Stay Warm>
① They have many feathers which overlap each other.
② They have woolly down feathers.
③ They have feathers (only counts as a separate idea if neither of the first two ideas about feathers is included in the response).
④ They have a thick layer of fat.
⑤ They huddle together in groups.

문항 R011A07C을 포함한 "Antarctica" 문항구획은 PIRLS 2006의 평가 개념틀에서 규정하는 두 개의 읽기 목적인 '문학적 경험(literary experience)', '정보 획득 및 활용(acquire and use information)' 중에서 '정보 획득 및 활용'에 해당하는 문항구획이다. 따라서 이 문항구획은 채점 기준에 있어서 피험자가 답해야할 정보가 명확하고, 몇 개의 정보를 정확하게 파악하여 답했는지에 따라 채점하기 때문에 상대적으로 명확한 기준을 가진다. 한편, '문학적 경험'에 대한 문항들의 경우에는 문학적인

글이 제시되고, '피험자가 문학적인 글에 참여하여 사건, 설정, 행동, 결과, 인물, 분위기, 감정 및 아이디어에 참여하고 상상하는 것을 목적'으로 하고 있다(Mullis et al., 2004). 따라서 이 영역의 문항에 대한 답안은 제시 글에 대한 피험자의 주관적 수용과 해석을 거친 결과물이기 때문에 '정보 획득 및 활용'에 비해 답안이 주관적이다. 그리고 이러한 답안에 대한 채점 기준 역시 <표 Ⅲ-4>에서 제시된 바와 같이 '정보 획득 및 활용'에 비해 기준에 대한 설명이 길고, 상대적으로 주관적인 판단의 여지가 많은 특징을 가진다.

<표 Ⅲ-4> PIRLS 2006 문학적 경험 R011C10C 채점 기준

점수	기준 설명
3점	These responses demonstrate extensive comprehension by integrating ideas from across the text to fully support an interpretation of why the clay's feelings changed during the story. The response provides an appropriate description of the clay's feelings at the beginning and at the end. It includes information from the story to explain why they changed. In the explanation the response demonstrates understanding of one of the following aspects of the clay's proud feelings about itself at the end of the story: fulfillment, usefulness, or beauty/aesthetics.
2점	These responses demonstrate satisfactory comprehension by integrating ideas from across the text to support an interpretation of why the clay's feelings changed during the story. The response provides an appropriate description of the clay's feelings at the beginning and at the end. It includes information from the story to explain why they changed. However, the explanation for why they changed does not demonstrate understanding of one of the following aspects of the clay's proud feelings about itself at the end of the story: fulfillment, usefulness, or beauty/aesthetics. Or, the response provides an appropriate explanation of his feelings at the beginning or the end (but not both) and in the explanation of that feeling demonstrates understanding of one of the following aspects of the clay's proud feelings about itself at the end: fulfillment, usefulness, or beauty/aesthetics
1점	These responses demonstrate limited comprehension of how the clay's feelings changed during the story. The response

점수	기준 설명
	provides an appropriate description of the clay's feelings at the beginning or at the end, or both, but does not include appropriate information from the story to explain why they changed. Or, the response provides an explanation of the change that demonstrates feelings of fulfillment, usefulness, or aesthetics, but does not describe his feelings at the beginning or the end.

구체적인 채점 절차에 있어서 구성형 문항들은 국가별로 채점자들에 의해 평정되었는데 국가 내 신뢰도 측정을 위하여 국가별로 최소 200명 이상을 채점할 것을 권장하였고, 채점 과정에서 채점 일치도가 85% 이하로 떨어지면 채점자를 재훈련하고 이미 채점된 답안도 모두 다시 채점하도록 하였다(Martin, Mullis, & Kennedy, 2007). 그리고 국가 내 신뢰도 채점 외에도 국가 간 비교와 연도별 추이 비교를 위한 국가 간 신뢰도 채점, 추이 신뢰도 채점도 이루어졌다.

PIRLS 2006의 설계와 관련하여 문항-피험자 설계는 문항을 10개의 구획(block)으로 나누고, 두 개의 구획이 하나의 소책자(booklet)를 구성하여 피험자가 하나의 소책자에 응시하도록 하여 문항-피험자 간 가교 설계를 구성하였다. 구체적인 문항 구획과 소책자 구성은 <표 Ⅲ-5>와 같다.

<표 Ⅲ-5> PIRLS 2006 문항 구획 및 소책자 구성

소책자 (Booklet)	구획 (Block)									
	L1	L2	L3	L4	I1	I2	I3	I4	L5	I5
1	○	○								
2		○	○							
3			○	○						
4				○	○					
5					○	○				
6						○	○			

소책자 (Booklet)	구획 (Block)									
	L1	L2	L3	L4	I1	I2	I3	I4	L5	I5
7							○	○		
8	○							○		
9	○				○					
10		○				○				
11			○				○			
12				○				○		
Reader									○	○

전체적인 소책자 설계는 1~12번 소책자에 L1~4, I1~4 구획을 포함하면서 두 번째 다른 소책자와 공통 구획을 가지는 방식으로 가교를 구성하는 불완전 구획 나선형 설계(incomplete block spiraling design) 형태이다. 다만 특이한 점은 10개의 구획들 중 2개의 구획은 “Reader” 소책자에만 포함시켰고(Martin et al., 2007), 이로 인하여 이 소책자로 시험을 치른 학생들에 소책자에 포함된 문항들은 나머지 피험자나 문항들과 가교가 없는 구조이다. 이와 같이 가교가 없을 경우, 모형의 추정에 있어서 수리적으로 해의 미결정성 문제를 가지기 때문에 일반적인 문항반응이론 모형을 통한 모수 추정이 불가능하고, 베이지안 추정을 활용하는 경우에도 해의 미결정성 문제로 사후분포가 결정될 수 없고 우연적으로 모수가 수렴되는 값으로 추정될 수 있지만 이를 받아들일 수 없다.

PIRLS 2006에서는 피험자-문항 설계 외에 다른 채점자수, 채점자-문항 설계, 채점자-피험자 설계 등 채점 설계와 관련한 명확한 강제 사항이 없어 국가별로 다양한 형태로 채점 설계가 이루어졌다. 또한 무선(random)적인 배정 방식도 아니기에 이러한 비체계적인 설계는 측정학적으로는 아쉽게 평가할 부분이지만, 한편으로는 다양한 채점 설계를 가지는 대규모 실제 자료를 분석할 수 있다는 점에서 실제 자료에 HRM과 그 확장 모형의 적용을 시도하는 본 연구에는 매우 유용한 상황이다.

나. 채점 자료 전처리 절차

PIRLS 2006은 전체 47개 국가 및 지역(이후 편의상 국가라고 표현) 단위로 본 채점과 신뢰도 채점 자료가 제공되어 총 94개 자료 파일이 제공된다. 이 자료를 모두 통합한 결과, PIRLS 2006 참여자 219,740명 중 성취도 자료에 포함된 피험자는 본 채점 대상자 218,324명이다. 그리고 본 채점 대상자 중 일부에 대한 신뢰도 채점 대상 47,002명의 자료를 합쳐 인원으로 총 265,326명의 자료에 대한 자료를 구성하였다.

가장 먼저 전체 62개 구성형 문항 중 HRM에서 다루는 범주수 3개 이상의 구성형 문항을 선별한 결과, 범주수 3개인 28개 문항과 범주수 4개인 6개 문항을 포함하는 총 34개 문항이 추출되었다. 그런데 이 문항들 중 'Reader' 소책자에 포함된 5개 문항(범주수 3개 4문항, 범주수 4개 1문항)은 나머지 12개 소책자에 포함된 문항들과 공통 피험자가 존재하지 않아 가교 설정이 불가능하므로 이를 제외한 29개 문항을 최종 문항으로 선택하고, 'Reader' 소책자에 응시한 피험자 54,962명을 분석 대상에서 제외하였다.

그리고 R011F12C 문항은 PIRLS 2006 기술보고서인 Martin, Mullis, & Kennedy(2007)에서 권고를 따라 총 793건(0.3%)의 3점 응답을 2점으로 변경하였다. 또한 'not reached', 'omitted'와 같은 결측값에 대해서는 일반적으로 문항반응이론 모형 분석에서 오답으로 처리하여 0점으로 변환하기도 하지만 HRM에서는 채점자가 답안에 대하여 0점을 부여한 것으로 처리되어 채점자 모수에 영향을 주기 때문에 문항 자체가 제시되지 않은 것으로 처리되도록 결측값 자체를 제거하였다. 이상의 문항에 대한 전처리 결과, 남은 문항과 그에 대한 기술 통계량은 다음 <표 Ⅲ-6>과 같다.

<표 Ⅲ-6> 문항 전처리 후 문항별 기술통계

문항	채점건수	최솟값	최댓값	평균	표준편차
R011A04C	49,356	0	2	1.090	.631
R011A08C	49,381	0	2	1.430	.696
R011A09C	46,826	0	2	1.180	.746
R011C08C	48,730	0	2	.970	.883
R011C11C	45,733	0	2	1.120	.703
R011F07C	48,127	0	2	.880	.908
R011F09C	49,339	0	2	1.500	.747
R011F12C	44,030	0	2	.790	.861
R011L08C	45,636	0	2	.800	.707
R011L10C	43,061	0	2	.760	.792
R011L12C	41,210	0	2	.890	.680
R021E05C	49,744	0	2	1.330	.848
R021E07C	49,214	0	2	1.160	.807
R021E09C	48,719	0	2	.750	.742
R021E12C	44,623	0	2	1.040	.787
R021K01C	50,719	0	2	1.400	.743
R021K07C	45,393	0	2	1.110	.815
R021K10C	38,319	0	2	.710	.844
R021N03C	47,183	0	2	.560	.709
R021N11C	44,424	0	2	1.100	.878
R021N12C	47,185	0	2	.970	.833
R021Y09C	48,475	0	2	1.460	.761
R021Y12C	49,148	0	2	1.050	.955
R021Y14C	47,945	0	2	.910	.901
R011A07C	50,043	0	3	2.040	1.099
R011C10C	45,832	0	3	1.600	1.179
R011L04C	48,597	0	3	1.250	.913
R021K12C	42,560	0	3	1.800	1.075
R021Y13C	42,585	0	3	1.330	1.055

문항별로 해당 문항에 대한 채점건수는 평균 46,625건이고, 최대 채점건수를 가지는 문항은 R011A07C로 50,719건의 채점이 이루어졌고, 최소 채점건수를 가지는 문항인 R021K10C에는 38,319건의 채점이 진행되었다.

문항에 대한 전처리를 거친 210,364건의 채점 자료에 대하여 채점자와 관련된 자료 전처리를 진행하였다. 가장 먼저 채점자ID가 결측되어 채점자 모형에서 가장 필수적인 채점자 식별이 불가능한 10,933건이 제외되었는데, 프랑스의 경우 모든 채점자ID가 결측되었기 때문에 국가 전체가 분석에서 제외되었다. 다음으로 서로 다른 채점자가 신뢰도 채점을 하도록 한 PIRLS의 권고를 따르지 않고 본 채점과 신뢰도 채점을 동일한 사람이 진행한 1,158건을 삭제하였다. 그리고 HRM의 경우, 한 피험자 혹은 한 답안에 한명의 채점자만 배정된 경우에도 분석이 가능하지만, 문항-피험자에 대한 행렬표집설계로 인하여 가교 설계가 복잡한 상황을 조금이나마 간소하게 만들고, 공통 채점자를 통한 채점자 간 직접 가교를 강화하여 채점자 효과 추정이 더 정확해지도록 단일 채점자가 채점한 사례 131,331건을 제거하여 66,942건이 남았다. 이 과정에서 신뢰도 채점자의 ID가 결측되어 신뢰도 채점자들이 모두 제거된 독일은 분석 대상 국가에서 제외되었다.

피험자 능력 모수 추정의 정확성 확보를 위하여 피험자 단위에서 답안수 즉, 피험자가 몇 문항에 응답하였는지를 확인하여 답안수가 3건 이하인 4,464건을 제외하였다. 그리고 채점자 모수 추정의 정확성 확보를 위해서 채점 대상자 단위가 아닌 답안 단위의 채점 건수가 30건 미만인 채점자를 삭제하였다. 그리고 이 채점자 삭제로 앞서 적용했던 단일 채점자가 다시 발생하였기에 30건 미만 채점의 채점자 삭제와 단일 채점자를 삭제를 반복하였는데, 총 3회를 반복하여 6,156건이 제외되어 56,322건이 남았고, 국가 단위에서는 쿠웨이트, 몰도바, 캐나다(알버타주), 캐나다(브리티시 컬럼비아주)가 대상 국가에서 제외되었다.

최소 답안 및 채점 건수를 확보한 자료에서는 공통 피험자를 통한 모

든 채점자간 가교가 가능한지를 사회관계망 분석을 통해 점검하였다. 확인 결과, 채점자가 2명씩인 리투아니아와 러시아는 채점자 간의 직접 가교가 가능했고, 대부분의 국가들은 1~3명의 다른 채점자를 거쳐 모든 채점자간의 가교가 가능하였고, 마케도니아는 7명의 채점자를 거쳐야 모든 채점자와 가교가 가능한 경우도 발생하였다. 한편, 오스트리아, 불가리아, 대만, 인도네시아, 이란, 이스라엘, 이탈리아, 폴란드, 싱가폴, 남아공, 스페인, 트리니다드 토바고 등 12개국은 공통 피험자를 통한 전체 채점자들의 가교 연결이 불가능하여 분석에서 제외하였다.

다음 단계로는 문항 범주에서 특정 범주가 누락되었는지를 확인하였는데, 모로코, 카타르, 남아공, 캐나다(온타리오주), 캐나다(퀘백주) 등 5개국이 특정 범주 혹은 문항이 누락되어 문항 수준에서 GPCM을 적용하는 연구 모형에 맞지 않아 분석 대상 국가에서 제외되어 총 24개국 18,034명의 피험자를 최종 분석 대상으로 선정하였고, 각 국가별로 채점 통계치를 제시하면 아래 <표 Ⅲ-7>과 같다.

<표 Ⅲ-7> 국가별 채점 규모 통계

국가명	피험자수	채점건수	채점자수
덴마크	594	8,079	16
조지아	691	8,946	12
홍콩	620	8,592	12
헝가리	789	10,971	6
아이슬란드	674	9,032	11
라트비아	831	11,541	10
리투아니아	796	11,186	2
룩셈부르크	803	10,938	8
네덜란드	709	9,981	8
뉴질랜드	1,043	14,268	11
노르웨이	636	8,429	14
루마니아	750	10,256	8
러시아	858	12,034	2

국가명	피험자수	채점건수	채점자수
슬로바키아	721	9,903	11
슬로베니아	817	11,015	9
스웨덴	852	11,693	14
마케도니아	714	9,337	22
미국	931	13,087	7
영국	757	10,487	7
스코틀랜드	638	8,778	6
벨기에(네덜란드어권)	825	11,487	6
벨기에(프랑스어권)	761	10,025	8
아이슬란드(5학년)	723	9,819	11
노르웨이(5학년)	501	6,851	13

다. 분석 자료 특성

자료 전처리 절차를 거쳐 최종적인 분석 대상인 자료의 특성을 피험자, 문항, 채점자 설계를 중심으로 살펴보겠다. 먼저 피험자-문항 설계는 소책자들마다 피험자를 고른 비율로 분배하여 한 피험자가 한 소책자만 풀도록 한다는 점에서 소책자별 문항 구성을 살펴보면 피험자-문항 설계를 파악할 수 있다. 12개 소책자와 29개 문항이 포함된 최종 분석 자료의 12개 소책자별 29개 문항 구성은 다음 <표 Ⅲ-8>과 같다.

<표 Ⅲ-8> 분석 자료 소책자별 문항 구성

문항	소책자											
	1	2	3	4	5	6	7	8	9	10	11	12
R011F07C	○	○								○		
R011F09C	○	○								○		
R011F12C	○	○								○		
R011C08C	○							○	○			
R011C11C	○							○	○			
R011C10C	○							○	○			

문항	소책자											
	1	2	3	4	5	6	7	8	9	10	11	12
R021Y09C		○	○								○	
R021Y12C		○	○								○	
R021Y14C		○	○								○	
R021Y13C		○	○								○	
R021E05C			○	○								○
R021E07C			○	○								○
R021E09C			○	○								○
R021E12C			○	○								○
R011A04C				○	○				○			
R011A08C				○	○				○			
R011A09C				○	○				○			
R011A07C				○	○				○			
R011L08C					○	○				○		
R011L10C					○	○				○		
R011L12C					○	○				○		
R011L04C					○	○				○		
R021N03C						○	○				○	
R021N11C						○	○				○	
R021N12C						○	○				○	
R021K01C							○	○				○
R021K07C							○	○				○
R021K10C							○	○				○
R021K12C							○	○				○
문항수	6	7	8	8	8	7	7	7	7	7	7	8

소책자별 문항 구성은 6개 문항으로 구성된 1번 소책자를 제외하면 7개 혹은 8개 문항으로 구성되어 있고, PIRLS 2006의 소책자-구획 설계에 따라 불완전 구획 나선형 설계를 갖추고 모든 문항들이 소책자를 가교로 서로 연결될 수 있어 동일 척도 상의 분석이 가능함을 확인할 수 있다.

분석 자료의 채점자와 문항 간의 가교 설계는 정확하게 규정되어 있던 피험자-문항 설계와 달리 국가마다 차이가 있었고, 채점자-문항 설계

조건의 셀 단위로 정리하면 다음 <표 Ⅲ-9>와 같다.

<표 Ⅲ-9> 분석 자료 채점자-문항 설계

국가명	채점자-문항 설계 특성			
	전체셀수	배정셀수	배정비율	완전구획설계
덴마크	464	361	77.8%	
조지아	348	348	100.0%	○
홍콩	348	348	100.0%	○
헝가리	174	174	100.0%	○
아이슬란드	319	316	99.1%	
라트비아	290	218	75.2%	
리투아니아	58	58	100.0%	○
룩셈부르크	232	208	89.7%	
네덜란드	232	181	78.0%	
뉴질랜드	319	318	99.7%	
노르웨이	406	406	100.0%	○
루마니아	232	139	59.9%	
러시아	58	58	100.0%	○
슬로바키아	319	307	96.2%	
슬로베니아	261	218	83.5%	
스웨덴	406	406	100.0%	○
마케도니아	638	638	100.0%	○
미국	203	196	96.6%	
영국	203	163	80.3%	
스코틀랜드	174	147	84.5%	
벨기에(네덜란드어권)	174	174	100.0%	○
벨기에(프랑스어권)	232	232	100.0%	○
아이슬란드(5학년)	319	296	92.8%	
노르웨이(5학년)	377	377	100.0%	○

채점자-문항 설계에서 모든 채점자가 모든 문항을 채점하는 경우는 총 11개국(조지아, 홍콩, 헝가리, 리투아니아, 노르웨이, 러시아, 스웨덴,

마케도니아, 벨기에-네덜란드어권, 벨기에-프랑스어권, 노르웨이-5학년) 이고, 이 국가들은 채점자-문항 설계에 있어서 완전구획설계가 이루어졌다고 할 수 있다. 나머지 국가들은 채점자-문항 설계 조건의 모든 셀들을 채우지 못하였지만, 공통 문항 혹은 공통 채점자를 가교로 삼아 모든 채점자들과 문항들이 서로 연결되는 것을 확인하였다.

연구 모형 중 E-HRM을 자료 분석에 적용하는 경우, 채점자-문항별 분석이 이루어지기 때문에 채점자-문항별로 채점건수를 확인할 필요성이 있고, 상당수의 채점자-문항 조건에서 적은 채점건수가 발견되어 5건을 기준으로 <표 Ⅲ-10>과 같이 정리하였다.

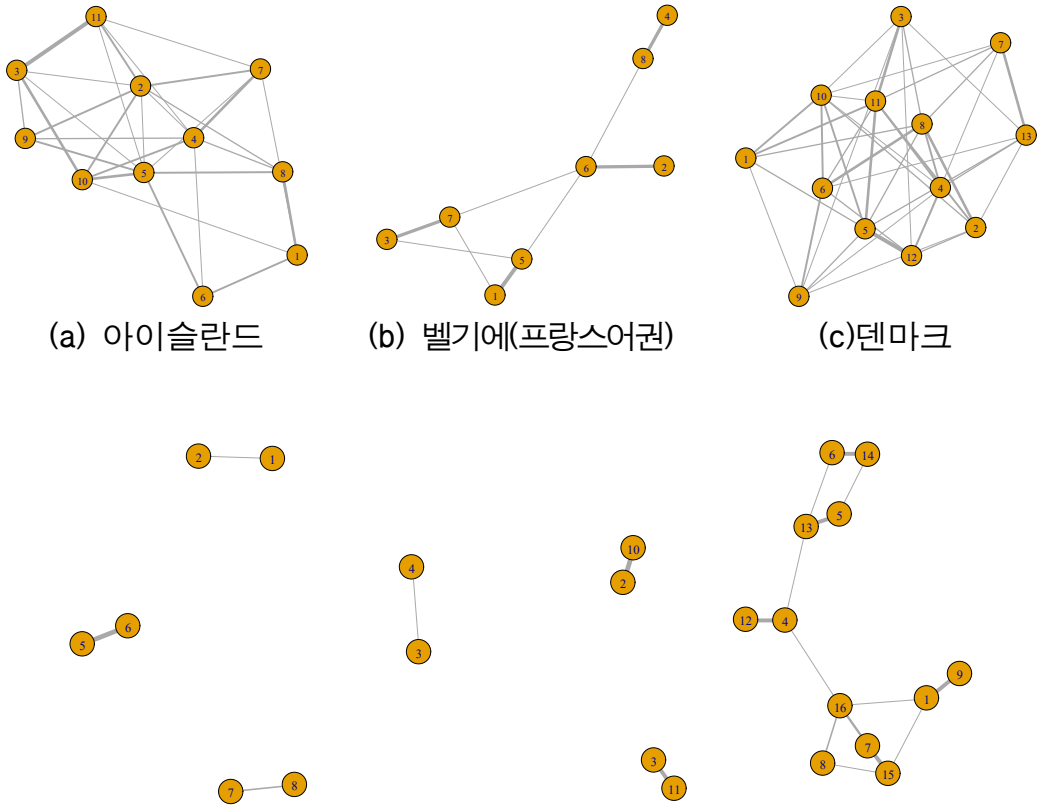
<표 Ⅲ-10> 분석 자료 채점자-문항별 소규모 채점 비율

국가명	채점자-문항별 소규모 채점 비율					
	1건	2건	3건	4건	5건	5건 이하
덴마크	11.9%	4.2%	4.4%	0.8%	1.1%	22.4%
조지아	-	-	0.3%	1.1%	1.1%	2.6%
홍콩	-	0.9%	-	1.1%	5.5%	7.5%
헝가리	-	-	-	-	-	-
아이슬란드	-	-	-	-	-	-
라트비아	0.5%	1.4%	-	-	-	1.8%
리투아니아	-	-	-	-	-	-
룩셈부르크	0.5%	-	-	-	-	0.5%
네덜란드	-	-	0.6%	2.8%	-	3.3%
뉴질랜드	0.3%	1.9%	3.1%	4.4%	1.9%	11.6%
노르웨이	0.5%	0.5%	1.2%	0.2%	0.5%	3.0%
루마니아	12.2%	0.7%	1.4%	0.7%	-	15.1%
러시아	-	-	-	-	-	-
슬로바키아	-	-	-	1.3%	-	1.3%
슬로베니아	1.8%	-	-	1.8%	-	3.7%
스웨덴	-	-	-	-	-	-
마케도니아	-	-	0.2%	-	0.2%	0.3%
미국	-	-	-	-	-	-

국가명	채점자-문항별 소규모 채점 비율					
	1건	2건	3건	4건	5건	5건 이하
영국	-	-	-	0.6%	3.7%	4.3%
스코틀랜드	-	-	-	-	3.4%	3.4%
벨기에(네덜란드어권)	-	-	-	-	-	-
벨기에(프랑스어권)	-	-	-	-	-	-
아이슬란드(5학년)	2.7%	-	-	0.7%	0.7%	4.1%
노르웨이(5학년)	-	-	-	-	-	-

채점자-문항별로 헝가리, 아이슬란드, 리투아니아, 러시아, 스웨덴, 미국, 벨기에(네덜란드어권), 벨기에(프랑스어권), 노르웨이(5학년)의 경우에는 5건 이하의 채점건수를 보이는 경우는 없었다. 반면에 덴마크의 경우 채점자-문항 셀에 1건의 채점만 포함한 경우가 전체 채점자-문항 조건 셀들 중 11.9%에 이르렀고, 5건 이하인 경우를 모두 합하면 무려 채점자-문항 조건 셀의 22.4%로 매우 높은 비율을 보였다. 덴마크 외에도 루마니아, 뉴질랜드가 채점자-문항별 채점건수가 소규모인 경우가 많았던 것으로 파악되어 이 국가들의 경우 모수 추정에 활용되는 채점자-문항별 채점건수 부족으로 인하여 해당 채점자 모수의 추정에 있어 어려움이 있을 것으로 예상된다.

마지막으로 피험자-채점자 설계에서는 피험자는 공통 채점자 가교로, 채점자는 공통 피험자 가교로 서로 연결될 수 있다. 공통 채점자를 가교로 피험자들이 연결되는 것은 사회관계망 분석(social network analysis)을 통해 [그림 Ⅲ-2]와 같이 도식화할 수 있다. [그림 Ⅲ-2]에서 채점자 사이를 연결하는 선(edge)의 굵기는 공통 채점한 피험자 수에 비례하여 표현하였고, 선의 굵기가 굵을수록 가교 설계 측면에서 채점자 간의 가교 강도를 의미한다.



(a) 아이슬란드 (b) 벨기에(프랑스어권) (c)덴마크

(d) 오스트리아 (e) 인도네시아

[그림 Ⅲ-2] 채점자 사회관계망 분석 그래프

[그림 Ⅲ-2]의 (a)~(c)는 모든 채점자가 공통 채점을 통해서 서로 연결될 수 있음을 잘 보여주는 국가들로 이후 5장의 국가별 분석에서 자세히 분석할 예정이다. 이 중 (a), (c)는 상대적으로 가교를 고르게 가지는 채점자 설계를 갖추고 있고, (b)의 경우 4쌍의 채점자들이 각각의 쌍 내에서는 서로 강한 연결을 가지는 것을 보여주지만 다른 쌍과는 적은 연결을 가지는 설계 특성을 보여주고 있다. 반면, (d)와 (e)는 모든 채점자가 가교 피험자를 통해서 연결되지 못하여 분석에서 제외된 오스트리아와 인도네시아 사례이다. 오스트리아는 두 명씩 네 집단을 이루어 집단 내에서만 연결된 경우이고, 인도네시아는 다수의 채점자들이 서로 연결되어 집단을 이루고 있고 소수 채점자들은 그 집단에 연결되지 못하고

있음을 확인할 수 있다.

2. 분석 절차

가. 적용 연구 분석

앞 절에서 전처리 과정을 거친 PIRLS 2006 자료에 대한 분석은 전체 234명의 채점자가 채점한 246,735건의 채점 자료를 국가별로 나누어 분석하는 방식으로 진행하였다. 분석에 활용한 주요 연구 모형은 3장 2절과 3절에서 각각 제시한 Patz et al.(2002)의 HRM과 HRM에서 채점자 분산성 모수의 채점자-문항 독립성 제약을 완화하여 ψ_{ir}을 추정하는 E-HRM이다. 그리고 채점자 엄격성 모수의 채점자-문항 독립성 제약을 완화하여 ϕ_{ir}을 추정하는 M3과 모든 채점자 모수의 채점자-문항 독립성 제약을 완전히 제거하고 ϕ_{ir}, ψ_{ir}을 모두 추정하는 M4를 대안모형으로 앞서 두 모형과 비교하고자 한다. 이에 따라 연구 모형들을 다시 제시하면 다음 4개의 수식과 같다.

$$\text{M1(HRM)} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}$$

$$\text{M2(E-HRM)} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k-(\eta_i+\phi_r)\right]^2\right\}$$

$$\text{M3(대안모형1)} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_r^2}\left[k-(\eta_i+\phi_{ir})\right]^2\right\}$$

$$\text{M4(대안모형2)} : P(Y_{ir}=k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k-(\eta_i+\phi_{ir})\right]^2\right\}$$

이상의 위계적 채점자 모형에 대한 실제 추정과정에서는 일반적으로

베이지안 통계학(Bayesian Statistics)에 기초한 마르코프 연쇄-몬테카를로(Markov Chain Monte Carlo; 이하 MCMC) 추정 방법이 주로 활용되는데(Patz et al., 2002), 본 연구에서는 깁스 표본 추출(Gibbs sampling) 방법으로 MCMC 추정을 하는 JAGS(Just another Gibbs sampler) 소프트웨어를 활용하여 분석을 진행하였다. MCMC 추정은 무작위로 표본을 추출하여 모수의 값을 확률적으로 추정하는 방법이고, 깁스 표본 추출 방법은 두 개 이상의 확률변수의 결합확률분포(Joint probability distribution)로부터 표본을 무작위로 추출하여 모수의 분포를 추정하는 방법이다. 깁스 표본 추출 방법이 결합확률분포를 사용한다는 점은 모수 간의 의존성을 가지는 확률분포를 다룰 수 있다는 것이므로, 하위 수준의 모수가 상위 수준 모수에 의존하는 같은 HRM과 위계적 모형 추정에 유용한 분석 방법이다(Kruschke, 2014).

MCMC 추정에는 채점 자료와 함께 모형에 포함된 모수에 대한 사전분포와 초기치 설정이 필요하다. 먼저 MCMC 연쇄가 모수 분포를 탐색하기 시작하는 지점이자 MCMC 연쇄 수렴의 빠르기에 영향을 미칠 수 있는 초기치는 특정한 값으로 설정하지 않고, JAGS 소프트웨어가 MCMC 연쇄별로 다양한 초기치를 무선적으로 설정하도록 하였다. 그리고 사전분포 설정이 필요한 모수는 피험자 능력모수 θ, 문항 변별도 모수 α, 문항 곤란도 모수(단계모수 포함) β, 채점자 엄격성 모수 ϕ, 채점자 분산도 모수 ψ이다. 이 중에서 2수준(피험자 수준) 모형인 GPCM에 포함되는 능력모수와 문항 모수들은 아래 수식들과 같이 사전분포의 분산이 크지 않은 정보적 사전분포(informative prior)를 활용하였다. 이는 MCMC 추정을 활용하는 GPCM에서 모형 식별(identification)을 위한 조치이기도 하다(Li & Baser, 2012).

$$\theta \sim N(0,1)$$
$$\log(\alpha) \sim N(0,0.12)$$
$$\beta \sim N(0,1)$$

반면에 본 연구에서 관심 모수들인 채점자 엄격성 모수와 분산성 모수는 이론적 분포가 존재하지 않고, 경험적으로 제시된 분포도 존재하지 않기 때문에 무정보적 사전분포(non-informative prior)를 활용하여 MCMC 연쇄가 넓은 범위에서 채점자 모수들의 분포를 탐색하도록 하였다. 다만, 양수와 음수 값을 모두 가질 수 있는 엄격성 모수와 달리 분산성 모수는 항상 양수가 되어야 하므로 정규분포 대신 로그정규분포 형태의 무정보적 사전분포를 활용하였고, 각각의 모수에 대한 구체적인 사전분포 설정은 아래와 같다.

$$\phi \sim N(0,10)$$
$$\log(\psi) \sim N(0,10)$$

MCMC 추출의 반복(iteration) 횟수는 연쇄의 안정화를 위한 burn-in 10,000번을 포함하여 20,000번으로 설정하였고, 솎아내기(thinning) 없이 연쇄 당 10,000개의 표본을 얻어서 6개의 연쇄들로부터 60,000개의 MCMC 표본을 추출하였다.

표집된 MCMC 표본으로부터 모형에 포함된 모수의 사후분포를 얻을 수 있고, 사후분포를 모수 해석에 활용하기는 어렵기 때문에 사후분포로부터 모수에 대한 점 추정치(point estimates)를 산출하였다. 사후분포의 점 추정치로 사용되는 대푯값으로는 일반적으로 사후 평균(posterior mean), 사후 중위값(posterior median), 근사 사후 최빈값(approximate posterior mode) 등이 활용된다. 한편 Patz et al.(2002)의 연구에서는 베이지안 추정을 활용하는 IRT 모형의 사후분포는 단봉-대칭(symmetric unimodal) 분포와 큰 차이를 보이는 경우가 종종 발생하기 때문에 사후 평균이나 사후 표준편차(posterior deviation)를 활용하지 않고, 사후 중위값과 신뢰구간(credible interval)을 대푯값으로 활용하였다. 이를 참고하여 본 연구에서도 문항 모수와 채점자 모수의 사후분포에 대한 점 추

정치로 사후 중위값을 활용하였고, 오차와 관련된 분석에는 주로 사후분포의 신뢰구간이나 유효 표본 크기(effective sample size), 몬테카를로 표준오차(Monte Carlo standard error)를 활용하였다.

반면, 능력 모수 추정에서는 θ가 $N(0,1)$인 정규 분포를 가정하고 있으므로 EAP(expected a priori) 추정치를 능력 모수의 대푯값으로 사용하였고, 능력모수의 오차로는 능력모수의 사후 표준편차(posterior standard deviations)를 활용하였다.

이와 같이 MCMC 방법을 통해 추정된 모수는 MCMC 연쇄가 적절하게 수렴되었는지를 평가하는 것이 필요하다. 다만, 피험자 모수는 지나치게 많고 연구의 주요 관심인 모형 비교에는 핵심적인 모수가 아니기에, 문항과 채점자 모수에 대해서만 평가하였다. 수렴 평가 방법은 Gelman & Rubin(1996)이 제안한 $\hat{R}$ 통계량을 활용하였고, Gelman & Rubin(1996)이 제시한 “<1.1” 기준을 적용하였다.

모형 비교에 있어서는 각 모형의 Deviance 통계량과 Deviance Information Criteria(이하 DIC) 통계량을 산출하여 모형 간에 비교하는 방법을 활용하였다. DIC는 deviance에 모형의 유효 모수 수(effective number of parameter)를 의미하는 pD를 더한 지수로 산출되는데, 이는 모형의 복잡도에 따른 패널티를 반영한 지수라 볼 수 있다. 이때 유효 모수 수는 AIC(Akaike Information Criterion)나 BIC(Bayes Information Criterion)에서 활용하는 모형 표현식에서 나타나는 단순한 모수 수를 의미하는 것이 아니고, 모수들 간의 상관이나 의존성을 감안하여 모형의 실질적인 모수가 몇 개인지를 의미하는 지수이다.

일반적으로 pD를 산출하는 방식은 세 가지 방법이 활용된다. 먼저, Spiegelhalter, Best, & Van Der Linde(2002)에서 제시한 방식은 $pD = \overline{(D(\theta))} - D(\overline{\theta})$이고, 두 번째 방법은 Gelman, Carlin, Stern, & Rubin(2004), Sturtz, Ligges, & Gelman(2005)에서 제시한 $pD = pV = \frac{1}{2}\overline{var(D(\theta))}$이다. 그리고 마지막 방법은 본 연구에서 MCMC

추정에 활용한 JAGS 프로그램의 개발자인 Plummer(2008)가 제안한 방법으로 연쇄 간 쿨백-라이블러 발산(Kullback-Leibler divergences)의 기댓값을 활용한 pD 산출 방법이다. 그런데 JAGS 프로그램에 포함되어 있어 가장 쉽게 활용할 수 있는 Plummer(2008)의 방식은 무정보적(noninformative) 사전분포를 사용할 경우, pD 값이 0이나 무한대 값으로 계산되는 문제점이 발생할 수 있어 채점자 모수의 분산을 10으로 설정한 본 연구 모형에서는 활용할 수 없다. 또한 Spiegelhalter et al.(2002)의 방법은 JAGS에서는 별도의 계산 과정과 모니터링이 필요한 부담이 있어 최종적으로 본 연구에서는 DIC 산출에 있어서 Gelman et al.(2004)와 Sturtz et al.(2005)가 제시한 pD를 활용하였다.

분석에 활용한 소프트웨어 정보를 상세히 밝히자면 자료 전처리에는 EXCEL 2019, SPSS 25.0 버전, R 3.6.3 버전 등을 활용하였고, MCMC 추정에는 JAGS 4.3.0 버전을 사용하였다. 그리고 분석과 자료 입출력의 편의를 위해 R 3.6.3 64-bit 소프트웨어 내에서 JAGS를 실행할 수 있도록 wrapper function을 제공하고, 여러 개의 MCMC 연쇄를 동시에 분석할 수 있도록 지원하는 runjags package를 활용하여 MCMC 분석을 진행하였다.

3. 모의실험 연구 절차

가. 모의실험 연구 설계

모의실험 연구 설계에서는 가능한 적용 연구에서 활용하는 PIRLS 2006의 자료 특성을 반영하면서 본 연구의 제안 모형인 E-HRM이 정확하게 작동할 수 있는 최소한의 조건을 탐색하고, 채점자-문항 독립성 완화에 따라 E-HRM과 기존의 HRM과의 차이점을 드러내는데 주안점을 두었다. 이 목적 하에 검사 규모를 결정하는 피험자수, 문항수, 전체 채점자수, 답안 당 채점자수 등의 요인과 이 요인들과 관련된 모수들의 특

성을 포함하는 모의실험 설계 조건을 구성하였다.

먼저, E-HRM이 작동할 수 있는 최소한의 조건과 관련해서는 가장 중요한 요인 중 하나는 모수 추정을 위한 사례수(피험자수 및 채점 답안수)의 결정이다. 수렴이나 성능 평가 대상이 아닌 피험자 능력모수를 제외하고 E-HRM의 모수들 중 추정 과정에서 가장 추정이 어려울 것으로 예상되는 모수는 채점자 분산성 모수이다. 문항 모수인 변별도 모수와 곤란도 모수의 경우 문항 단위에서 추정되고, 채점자 엄격성 모수의 경우 채점자 단위에서 추정되기 때문에 추정에 활용되는 피험자수가 상대적으로 충분하다. 반면에 E-HRM의 분산성 모수의 경우에는 채점자-문항 단위에서 추정된다는 점에서 채점자의 문항별 분산성 모수 추정에 활용되는 사례수가 나머지 모수들에 비하여 매우 적어질 수밖에 없고, 실제 적용 연구 대상 자료인 PIRLS 2006 자료에서도 <표 Ⅲ-11>에서 확인할 수 있듯이 채점자-문항당 채점건수가 50명 이하인 경우가 73.5%로 매우 높았다.

<표 Ⅲ-11> PIRLS 2006 분석 자료 채점자-문항당 채점건수 분포

채점건수	빈도	비율(%)	채점건수	빈도	비율(%)
[1, 10]	649	10.3	[121, 130]	20	.3
[11, 20]	1,331	21.2	[131, 140]	11	.2
[21, 30]	1,333	21.2	[141, 150]	20	.3
[31, 40]	823	13.1	[151, 160]	2	.0
[41, 50]	484	7.7	[161, 170]	3	.0
[51, 60]	473	7.5	[171, 180]	18	.3
[61, 70]	424	6.7	[181, 190]	18	.3
[71, 80]	245	3.9	[191, 200]	26	.4
[81, 90]	161	2.6	[201, 210]	40	.6
[91, 100]	71	1.1	[211, 220]	4	.1
[101, 110]	58	.9	[221, 230]	16	.3
[111, 120]	57	.9			

채점자-문항당 채점건수를 결정하기 위해서 고려되어야할 요인들은 피험자수, 문항수, 전체 채점자수, 답안 당 채점자수, 채점자배치 방식 등이 있다. 첫 번째로 피험자수와 관련해서 기존의 문항반응이론 모형 모의실험 연구들에서는 피험자 수를 100, 250, 500, 1,000 등과 같이 설정한 경우가 많았다(강태훈, 김명연, 2012; 이현숙, 2008; Casabianca & Wolfe, 2017; Jiao, Wang, & Kamata, 2005; DeCarlo, 2008; DeCarlo, 2010; Kim 2013; Kim, Morris, Nering, & Oshima, 2009; Raju, Fortmann-Johnson, Wang, Bradlow, & Wainer, 2002). 앞서 적용 연구 대상에서 살펴본 PIRLS 2006의 분석 단위인 국가별 피험자수의 최솟값이 약 500명 규모였다는 점을 감안하여 본 연구에서는 피험자수를 500명으로 설정하였다.

두 번째로 문항수의 경우, 적용 연구에서 활용하는 PIRLS 2006의 피험자들이 응답한 다분 반응 구성형 문항수를 고려하였다. 앞서 적용 연구 대상에서 PIRLS 2006 자료의 전처리 결과, 피험자들이 하나씩 할당받게 되는 소책자들은 1번 소책자를 제외하면 7개 혹은 8개 문항으로 구성되어 있다. 그리고 본 연구와 유사하게 대규모 표준화 검사 상황을 가정한 HRM 모의실험 연구인 Casabianca & Wolfe(2017)에서는 2~8개의 문항수를 조건으로 활용하고 있는데, 문항별로 채점자 분산성 모수의 다양성을 충분히 설정할 수 있도록 적용 연구 분석 자료나 Casabianca & Wolfe(2017)에 비하여 조금 더 많은 문항수를 10개로 결정하였다. 그리고 모의실험 연구 조건이 지나치게 복잡해지는 것을 방지하기 위해 국제비교 대규모 검사에서 활용하는 행렬표집설계는 배제하고, 모든 피험자(500명)가 모든 문항(10개)에 응시하는 설계를 적용하여 각 모의실험 자료별로 5,000개씩의 답안을 포함하도록 하였다.

다음으로 전체 채점자수, 답안 당 채점자수, 채점자-문항당 채점 건수를 포함하는 채점 설계 요인은 E-HRM을 적용할 수 있는 설계 조건의 탐색과 가장 밀접한 요인이다. 이는 피험자수와 문항수를 고정적인 요인으로 전제하였을 때, 채점자 모수의 추정에 가장 큰 영향을 미치는 자료의 구조적 특성인 채점자-문항당 채점 건수가 전체 채점자수, 답안 당

채점자수를 각각 몇 명으로 설정할지에 따라서 결정되기 때문이다. 먼저, 답안 당 채점자수는 현실적인 비용과 시간 등의 문제로 인하여 실제 대규모 검사 채점에서 특별한 목적이 없다면 1~2명의 채점자가 하나의 답안을 채점하는 것이 일반적이라는 점을 고려하고, 채점자-피험자 및 채점자-문항 간 가교가 안정적으로 이루어지도록 피험자 당 채점자를 2명으로 설정하였다. 그리고 배정된 2명의 채점자가 해당 피험자가 답한 10개의 답안을 모두 채점하도록 채점 배정을 설계하여 답안 당 채점자수 역시 2명으로 설정되었다.

답안 당 채점자수가 2명으로 고정되면, 채점 설계 요인 중 채점자-문항당 채점 건수 요인과 전체 채점자수 요인의 결정이 남게 된다. 그런데 나머지 채점 설계 요인들이 고정되었을 때, 전체 채점자수 요인과 채점자-문항당 채점건수 요인은 독립적으로 결정될 수 없고, 전체 채점자수에 따라 채점자-문항당 채점건수가 결정된다. 전체 답안의 채점을 고르게 배분한다는 전제하에 전체 채점자수를 조정함으로써 채점자-문항당 채점건수 요인의 하위 조건들을 설정할 수 있다. 전체 R명의 채점자 중 2명씩 조합하고, 이 조합을 피험자 500명에 대해 채점자별 채점건수를 동일하게 조정하여 배분하면 한 명의 채점자는 $(1000/R)$명의 피험자를 채점하게 된다. 그리고 모든 피험자는 모든 문항에 응시하였기 때문에 채점자-문항당 채점건수도 역시 $(1000/R)$건이 된다.

한편, 이 연구에서도 전체 채점자수는 분명히 E-HRM의 추정에 영향을 줄 수 있는 설계 요인이지만, 실제 대규모 성취도 검사에서 전체 채점자수는 일반적으로 검사 규모에 따라 부차적으로 결정되는 요인이고, 전체 채점자수에 따른 분석 결과의 차이가 줄 수 있는 실제적 함의도 거의 없다는 점에서 이 연구에서 관심을 갖는 요인이 아니다. 또한 채점자-문항당 채점건수를 포함한 다른 채점 설계 요인들이 고정일 때에 전체 채점자수를 늘릴 방법은 피험자수를 늘리는 방법밖에 없기 때문에 현실적으로 전체 채점자수를 채점자-문항당 채점건수와 독립적인 설계 요인으로 포함하는 것은 불가능하다고 판단하였다. 따라서 본 연구의 모의실

험 설계에서는 채점자-문항당 채점건수 요인이 주요 설계 요인이 되고, 전체 채점자수는 채점자-문항당 채점건수에 종속적인 요인으로 다루어졌다.

채점자-문항당 채점건수 설계 요인은 고전적 문항반응이론 모형에 대한 연구들(Bamber & van Santen, 1985; Hanson & Beguin, 2002)과 베이지안 추정을 활용한 동등화 목적의 연구인 남현우(2015)에서 문항 단위 최소 사례수로 100명이 제시된 것을 참고하여 100건을 우선 결정하였고, 그보다 큰 200건과 그보다 작은 50건, 25건을 추가하여 총 4가지 수준으로 채점자-문항당 채점건수를 결정하였다. 이에 따라 채점자-문항당 채점건수에 반비례하는 전체 채점자수는 5명, 10명, 20명, 40명이 된다.

그리고 채점자 배정은 무선 배정을 계획하였는데, 무한하지 않은 사례에 대해 무선 배정을 하게 되면 채점자-채점자 가교나 채점건수에서 불균형이 발생한다. 때문에 실제 채점자 배정은 무선적으로 문항에 채점자를 배정하는 방식이 아니라, 전체 채점자수에 따라 가능한 2명의 채점자 조합을 모두 생성하고 무선적으로 순서를 바꾼 뒤, 이 조합을 피험자별로 배정하여 공통 채점을 통한 채점자-채점자 가교가 일정하도록 하였다. 다만 가능한 모든 채점자 조합의 수가 피험자수의 배수가 아니기 때문에 채점자별 채점 대상 피험자수에 1명 정도의 차이가 발생하였다.

자료 구조적 설계 요인 다음으로는 모형에 포함된 피험자 모수, 문항 모수, 채점자 모수를 결정하여야 한다. 먼저 피험자 모수는 매 모의실험 자료 생성 과정마다 표준정규분포 $N(0,1)$에 따라 무선적으로 추출하는 가장 일반적인 모의실험 방식을 따랐다. 그리고 문항 모수의 경우, 적용 연구에서 활용하는 PIRLS 2006의 기술보고서에 제시된 문항반응이론 분석 결과로부터 문항 모수를 택하였는데, 3개 범주를 갖는 24개 문항의 모수 분석 결과들에서 문항 곤란도 모수의 단계 모수가 역전되지 않은 경우들 중 단계 모수들 간의 간격이 큰 10개의 문항을 선택하였다. 그리고 문항 곤란도 모수와 단계 모수가 합쳐진 형태의 문항 곤란도 모수로 변환하고, 전체 문항 곤란도 모수의 평균이 0이 되도록 조정하여 모의실

험 연구의 10개 문항 모수에 고정 값으로 활용하였다. 모의실험 연구 자료 생성에 활용된 문항 모수는 다음 <표 Ⅲ-12>에 제시하였다.

<표 Ⅲ-12> 모의실험 연구 문항 모수

문항	문항 변별도	문항 곤란도	
	α	β_1	β_2
1	.677	-.560	-.110
2	.755	.449	.999
3	.844	-.321	.349
4	.606	-1.501	-.541
5	.607	-.060	1.002
6	.724	-.794	.270
7	.822	-.222	1.082
8	.827	-.325	1.079
9	.812	-.843	.631
10	.787	-1.329	.741

마지막으로는 가장 중요하고 본 연구의 관심인 채점자 모수 결정이 남았는데, 채점자-문항 독립성 가정을 완화하였기 때문에 E-HRM이 가질 수 있는 기존 HRM과의 차이점이 드러나도록 설계를 구성하고자 하였다. HRM에 비해 E-HRM이 가지는 가장 뚜렷한 차별점이자 강점은 문항별로 채점자 분산성을 각각 추정하여 HRM에서 판별하지 못했던 문항에 따른 차별적 채점자 영향을 판별해내는 것이다. 따라서 일부 문항의 정상범위 기준인 '$\psi_{ir} < .5$'를 벗어나는 채점자 분산성 모수를 갖는 조건이 필요하고, 채점자 엄격성 역시 정상범위 기준인 '$|\psi| < .5$'를 벗어나는 조건 역시 고려되어야 하며, 이러한 조건들 하에서 E-HRM과 HRM의 분석 결과를 비교해 볼 필요가 있겠다. 이에 정상범위 채점자 비율 요인을 모의실험 설계 요인에 포함하고, 정상범위 채점자 모수만을 가지

는 채점자와 정상범위를 벗어나는 채점자 모수를 가지는 채점자의 비율을 달리하는 하위 수준을 계획하였다. 정상범위 채점자 비율의 하위 수준은 모든 채점자가 엄격성 및 분산성 모수를 가지는 '100%' 하위 수준과 정상범위 채점자 모수만을 가지는 채점자의 비율이 점차 늘어나고 정상범위를 벗어나는 채점자 모수를 가지는 채점자 비율이 점점 늘어나는 '70%', '40%' 조건으로 이루어진 총 3개의 하위 수준을 설정하였다.

한편, 채점자 모수 설정에 있어서는 채점자 모수에 대해서는 알려진 이론적 분포가 존재하지 않고 무선적으로 추출할 경우에는 현실적이고 합리적인 범위를 벗어나는 모수가 나올 수 있어 무선 추출 방식을 활용하는 것은 부적절하다고 판단하였다. 그리고 기존 HRM 연구들에서 보고된 모수 값들이 있긴 하지만 대부분 문항수가 적어 다수의 모수 값을 E-HRM 연구에 활용하기에는 부족하였기 때문에 적용 연구에서 도출된 E-HRM의 채점자 모수들 중 일부를 추출하여 활용하고자 하였다. 다만, 매 모의자료마다 채점자들의 엄격성과 분산성이 다양한 범위를 가질 수 있도록 적용 연구에서 얻은 채점자 모수들을 채점자 단위에서 '정상범위', '엄격', '저신뢰', '엄격-저신뢰'의 4가지 유형의 집단으로 구분하였다.

구체적인 채점자 구성에서는 정상범위 채점자 비율이 '100%'인 수준에서는 '정상범위' 집단의 채점자만을 무선 추출하였고, 정상범위 채점자 비율 '70%' 수준에서는 '정상범위' 채점자 유형에서 먼저 70%를 무선 추출하고, '엄격', '저신뢰', '엄격-저신뢰' 유형에서 각각 10%씩 무선 추출하여 총 30%의 채점자가 정상범위 밖의 채점자 모수를 갖는 채점자로 구성되도록 설계하였다. 단, 설계요인 중 전체 채점자수 요인이 5명이고, 정상범위 채점자 비율이 '70%'인 조건에서는 5명의 70%가 정수가 아니기 때문에 '엄격'과 '저신뢰' 유형에서 1명씩 정상범위를 벗어나는 채점자를 추출하여 나머지 전체 채점자수 설계 요인의 조건들에 비하여 높은 40%의 채점자가 정상범위 밖의 채점자 모수를 갖게 된다. 그리고 정상범위 채점자 비율 '40%' 수준에서는 '정상범위' 채점자 유형에서 전체 채점자의 40%를 무선 추출하고, '엄격', '저신뢰', '엄격-저신뢰' 유형에서

각각 20%씩 무선 추출하여 총 60%의 채점자가 정상범위 기준을 충족하지 못하는 채점자 모수를 갖는 채점자로 구성되도록 설계하였다. 구체적인 채점자 모수는 모의실험 연구 결과의 도입에서 간략히 제시할 예정이다.

이상의 모의실험 설계 관련 사항을 정리하면, 고정 설계 요인으로는 피험자수(500명), 문항수(10개), 문항 범주수(3개), 피험자당 채점자수(2명), 문항 모수, 피험자 분포, 채점자 풀 등이 있고, 변동 설계 요인으로는 채점자-문항당 채점건수, 정상범위 채점자 비율이 있다. 모의실험 설계의 전체 요인 및 수준은 다음 <표 Ⅲ-13>과 같고, 자료 생성에 있어서 조정하는 것은 '전체 채점자수'이지만, 연구의 관심이 E-HRM의 정상적인 작동 조건 탐색에 있기 때문에 설계 요인은 '채점자-문항당 채점건수'로 제시하였다.

<표 Ⅲ-13> 모의실험 연구 설계 요인 및 수준

구분	설계 요인	수준수	수준 설계
고정 요인	피험자수	1	500명
	문항수	1	10개
	문항 범주수	1	3개
	피험자당 채점자수	1	2명
	문항 모수	1	<표 Ⅲ-12> 참조
	피험자 분포	1	$N(0,1)$
	채점자 풀	1	<표 Ⅵ-13> 참조
변동 요인	채점자-문항당 채점건수 (전체 채점자수)	4	200건 (5명)
			100건 (10명)
			50건 (20명)
			25건 (40명)
	정상범위 채점자 비율	3	100%
			70%
			40%

나. 모의자료 생성 및 분석 방법

모의실험 연구 설계의 조건에 따라서 모의자료를 반복 생성하는 데에는 R 3.6.3 64-bit 프로그램을 활용하였다. 이 연구에서 분석에 활용하는 최종 모의자료는 모의 채점 자료이다. 모의 채점 자료를 생성하기 위하여 1단계로 앞서 모의실험 설계에서 제시한 사항들 따라서 피험자 모수는 표준정규분포에서 무선 추출하고, 채점자 모수는 채점자 모수 풀로부터 무선 추출한다. 2단계에서는 p명의 피험자가 i개 문항에 대해 획득할 $(p \times i)$개의 이상 채점($0 \sim k$점, $k = K-1$, K : 범주수)을 얻기 위해 1단계에서 추출된 피험자 및 문항 모수를 아래의 부분점수모형에 투입하여 확률적 다분 문항 반응을 생성한다.

$$P(\eta_i = k|\theta) = \frac{\exp \sum_{m=1}^{k} \alpha_i(\theta - \beta_{im})}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \alpha_i(\theta - \beta_{im})}$$

3단계에서는 앞 단계에서 생성한 $(p \times i)$개의 다분 문항 반응과 함께 r명의 채점자 모수를 다음의 E-HRM의 1수준(채점자 수준) 모형에 투입하여 $(p \times i \times r)$개의 확률적인 채점 자료를 생성한다.

$$P(Y_{ir} = k|\eta_i) \propto \exp\left\{-\frac{1}{2\psi_{ir}^2}\left[k - (\eta_i + \phi_r)\right]^2\right\}$$

마지막으로 4단계에서는 전체 채점자들 중 2명씩 조합하여 각 조합을 피험자별로 무선적으로 할당하는 채점자 배정을 생성하고, 채점 배정에 따라 채점이 존재하지 않는 채점자-피험자의 채점 자료를 제거하여 최종적인 모의실험 자료를 생성한다. 이상의 4단계에 걸친 모의실험 자료

생성은 각각의 모의실험 조건별로 총 50회씩 반복한다.

한편, 본 연구에서 모의실험 자료 생성에 활용한 R 3.6.3 소프트웨어를 포함한 모든 소프트웨어에서 제공하는 무작위 값 생성 알고리즘은 통계학적인 의미에서의 난수가 아니라 아주 큰 난수표로부터 순서대로 값을 가져온 의사난수(pseudo-random)이다. 그리고 난수표를 결정하는 것을 시드(seed)라고 하는데, 시드를 동일하게 설정하는 난수 조절을 통해서 동일한 무선 추출 자료를 얻는 것이 가능하다. 난수 생성기의 원리를 활용하여 모의실험 설계 조건과 반복 회차가 동일하고 분석 모형만 다른 경우에는 같은 시드를 갖도록 설정하여 동일한 모의실험 자료가 생성되도록 하였다. 이와 같은 난수 조절을 통해 E-HRM과 HRM 간 비교를 수행하는 데에 있어서 무선적인 영향을 최소화하고자 하였다. 이상의 절차로 반복 생성된 모의 자료에 E-HRM과 HRM을 적용하여 MCMC 분석을 하는 과정에서는 앞서 적용 연구와 동일하게 JAGS 4.3.0을 사용하였고 설정한 사전분포 역시 동일하였다. 하지만 50회의 반복 분석임을 고려하여 연쇄의 수는 적용 연구의 6개보다 적은 3개만 활용하였고, 추출된 MCMC 표본은 4개 중 1개만 저장하는 thinning을 적용하였다.

다. 모의자료 분석 결과 평가 준거

모의실험 연구에서 반복 생성된 모의 자료에 대한 모수 추정 방법은 적용 연구와 마찬가지로 깁스 표본 추출법을 활용한 MCMC 추정 방법을 사용하였기 때문에 모수의 정확성과 효율성을 평가하기에 앞서 추정치에 대한 연쇄가 적절하게 수렴되었는지를 평가하는 과정이 선행되어야 한다. 적용 연구의 기준과 동일하게 피험자 모수는 수렴 평가 대상에서 제외하였고, 문항 모수(문항 변별도 모수, 곤란도 모수)와 채점자 모수(채점자 엄격성 모수, 채점자 분산성 모수)에 대한 수렴 평가는 Gelman & Rubin(1996)이 제안한 $\hat{R}$ 통계량을 활용하고, 기준 역시 Gelman & Rubin(1996)이 제시한 “<1.1” 기준을 적용하였다.

모의실험을 통해 추정된 모수의 정확성과 효율성은 편향(bias), 평균제곱오차(Mean Square Error; 이하 MSE), 제곱편향(squared bias; 이하 SB), 분산(variance; 이하 VAR)으로 평가하였다. 먼저, 정확성은 모수 추정값의 평균이 모수의 참값과 얼마나 일치하는지를 의미하고, 편향을 통해 확인할 수 있다(Hoogland & Boomsma, 1998). 그리고 효율성은 모수 추정값이 얼마나 일관되게 추정되는지를 나타내고 추정값의 표준편차 또는 분산을 통해서 평가할 수 있다(Carsey & Harden, 2014). 마지막으로 MSE는 SB와 VAR의 합으로 나타나고, 모수 추정값의 정확성과 효율성을 함께 평가할 수 있는 값이다(Carsey & Harden, 2014; Lee & Ban, 2010; 강태훈, 심혜진, 2015).

Ⅳ. 적용 연구 결과: Overview

이 장에서는 HRM과 E-HRM을 활용한 PIRLS 2006 자료에 대한 적용 연구 분석에서 나타난 경향성에 대해서 제시하고자 한다. 앞서 연구 방법에서 설명하였듯이 적용 연구 분석은 연구의 주요 관심 모형인 HRM과 E-HRM과 더불어 두 개의 대안모형까지 총 4개의 모형으로 24개 국가를 분석하여 총 96번의 분석을 진행하였다. 구체적인 국가별 분석 결과에 대한 해석은 이후 Ⅴ장에서 제시하였고, 이 장에서는 96번의 분석 결과를 전반적으로 살펴보면서 모형들의 적용 가능성과 해석 가능성을 검토하고, 모형 간 혹은 국가 간에 특징적인 차이가 드러나는지 탐색하였다.

1. 모형 수렴

가. 모형별 수렴

모형별로 각각의 모수에 대하여 10,000개의 burnin 이후에 표집된 10,000개의 MCMC 표본의 수렴 여부를 진단한 결과, 각 국가 및 모형별 전체 모수 수와 수렴 비율은 아래 <표 Ⅳ-1>과 같다.

<표 Ⅳ-1> 국가 및 모형별 전체 모수 수렴 비율

국가	HRM		E-HRM		대안모형1		대안모형2	
	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율
덴마크	124	1.000	469	.977	469	.972	814	.885
조지아	116	1.000	452	.951	452	.931	788	.943
홍콩	116	1.000	452	1.000	452	.969	788	.931
헝가리	104	.990	272	1.000	272	.915	440	.852
아이슬란드	114	1.000	419	1.000	419	.933	724	.907

국가	HRM		E-HRM		대안모형1		대안모형2	
	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율
라트비아	112	1.000	320	.988	320	.969	528	.890
리투아니아	96	1.000	152	1.000	152	.914	208	.837
룩셈부르크	108	1.000	308	.994	308	.831	508	.839
네덜란드	108	1.000	281	1.000	281	.868	454	.813
뉴질랜드	114	1.000	421	.988	421	.941	728	.970
노르웨이	120	1.000	512	.957	512	1.000	904	.940
루마니아	108	1.000	239	.987	239	.996	370	.968
러시아	96	1.000	152	1.000	152	.849	208	.793
슬로바키아	114	1.000	410	1.000	410	.959	706	.946
슬로베니아	110	1.000	319	.994	319	.928	528	.867
스웨덴	120	1.000	512	.992	512	1.000	904	.937
마케도니아	136	1.000	752	.980	752	.927	1,368	.966
미국	106	1.000	295	1.000	295	1.000	484	.917
영국	106	1.000	262	1.000	262	.969	418	.971
스코틀랜드	104	1.000	245	.992	245	1.000	386	.907
벨기에 (네덜란드어권)	104	1.000	272	.996	272	.967	440	.925
벨기에 (프랑스어권)	108	.972	332	1.000	332	1.000	556	.995
아이슬란드 (5학년)	114	1.000	399	.990	399	.920	684	.836
노르웨이 (5학년)	118	1.000	482	.925	482	.936	846	.882

모형별로 모든 모수가 수렴($\hat{R}<1.1$)한 국가는 기저 모형인 HRM에서 22개국으로 나타났고, 채점자 분산성 모수에 대하여 채점자-문항 독립성 제약을 완화하여 추정할 채점자 분산성 모수의 개수가 증가한 E-HRM에서는 10개국만이 모든 모수가 수렴한 것으로 나타났다. 그리고 채점자 엄격성 모수의 채점자-문항 독립성 제약을 완화한 대안모형1의 경우, E-HRM과 비교하였을 때, 전반적으로 수렴 비율이 떨어졌고 5개 국가 자료에서만 수렴이 가능한 것으로 나타났다. 마지막으로 채점자 엄격성

과 분산성 모수 모두 채점자-문항 독립성 제약을 제거한 대안모형2의 경우, 앞서 Ⅱ장의 모형 검토에서 예측한 바와 같이 모든 국가 자료에서 수렴하지 않는 것으로 나타나 실질적으로 정상적인 추정이 불가능한 것으로 판단된다.

나. 모수 유형별 수렴

모수 유형별로 자세한 수렴 여부를 살펴보기 위하여 문항 모수와 채점자 모수 각각의 수렴 여부를 확인한 결과 중 문항 모수별 MCMC 연쇄의 수렴에 대한 분석 결과는 아래 <표 Ⅳ-2>와 같다. 정보적 사전분포를 활용한 문항모수인 α모수 29개와 β모수 63개(점수 범주가 3개인 24개 문항은 각각 2개, 범주가 4개인 5개 문항은 각각 3개)의 경우, HRM와 E-HRM에서는 모든 국가에서 충분히 수렴하는 것으로 나타났다. 반면에 채점자-문항 독립성 제약이 채점자 엄격성 모수에 대해 완화된 대안모형1과 두 가지 채점자 모수 모두에 대해 완화된 대안모형2에서는 문항 모수에 대한 MCMC 연쇄가 잘 수렴되지 않는 것으로 나타났다.

<표 Ⅳ-2> 문항 모수 수렴률

국가	모수	모수수	수렴 비율			
			HRM	E-HRM	대안1	대안2
덴마크	α	29	1.000	1.000	.966	.862
	β	63	1.000	1.000	.984	.825
조지아	α	29	1.000	1.000	.931	1.000
	β	63	1.000	1.000	.921	1.000
홍콩	α	29	1.000	1.000	.966	1.000
	β	63	1.000	1.000	.984	.937
헝가리	α	29	1.000	1.000	.966	.966
	β	63	1.000	1.000	.921	.841
아이슬란드	α	29	1.000	1.000	.966	.966
	β	63	1.000	1.000	.968	.905

국가	모수	모수수	수렴 비율			
			HRM	E-HRM	대안1	대안2
라트비아	α	29	1.000	1.000	.966	.931
	β	63	1.000	1.000	.968	.905
리투아니아	α	29	1.000	1.000	.966	.897
	β	63	1.000	1.000	.905	.794
룩셈부르크	α	29	1.000	1.000	.862	.828
	β	63	1.000	1.000	.825	.810
네덜란드	α	29	1.000	1.000	.897	.828
	β	63	1.000	1.000	.905	.841
뉴질랜드	α	29	1.000	1.000	.966	.966
	β	63	1.000	1.000	.937	.968
노르웨이	α	29	1.000	1.000	1.000	1.000
	β	63	1.000	1.000	1.000	.984
루마니아	α	29	1.000	1.000	1.000	1.000
	β	63	1.000	1.000	1.000	.952
러시아	α	29	1.000	1.000	.966	.862
	β	63	1.000	1.000	.873	.762
슬로바키아	α	29	1.000	1.000	.966	.931
	β	63	1.000	1.000	.984	.952
슬로베니아	α	29	1.000	1.000	.966	.931
	β	63	1.000	1.000	.952	.889
스웨덴	α	29	1.000	1.000	1.000	.931
	β	63	1.000	1.000	1.000	.937
마케도니아	α	29	1.000	1.000	.966	1.000
	β	63	1.000	1.000	.968	.984
미국	α	29	1.000	1.000	1.000	.897
	β	63	1.000	1.000	1.000	.921
영국	α	29	1.000	1.000	.966	1.000
	β	63	1.000	1.000	.968	.952
스코틀랜드	α	29	1.000	1.000	1.000	1.000
	β	63	1.000	1.000	1.000	.905
벨기에(네덜란드어권)	α	29	1.000	1.000	.966	.931
	β	63	1.000	1.000	.968	.921
벨기에(프랑스어권)	α	29	1.000	1.000	1.000	1.000
	β	63	1.000	1.000	1.000	.984
아이슬란드(5학년)	α	29	1.000	1.000	1.000	.828
	β	63	1.000	1.000	.921	.810
노르웨이(5학년)	α	29	1.000	1.000	1.000	1.000
	β	63	1.000	1.000	.921	.889

대안모형1은 E-HRM과 추정하는 모수 수가 동일한데도 불구하고, 6개 국가에 대한 분석에서만 문항 모수가 수렴되고, 나머지 국가들에 대한 분석에서는 수렴되지 않았다는 점에서 대안모형1은 단순히 HRM보다 모수수가 많기 때문에 문항 모수가 MCMC 연쇄의 비수렴이 발생할 것이라기보다 채점자 엄격성 모수를 문항별로 추정함으로 인하여 문항 곤란도와 문항 변별도 모수 추정의 안정성에 영향을 미친 것으로 판단된다. 그리고 엄격성 모수에 더해 분산성 모수까지 채점자-문항 독립성 제약이 완화된 대안모형2는 1개 국가에서만 모든 문항 모수가 수렴되었고, 나머지 23개 국가 자료에 대해서는 92개의 문항 모수 중 평균 7.9개, 최대 24개까지도 수렴되지 않는 추정의 불안정성을 보였다.

무정보적 사전분포를 활용하여 추정한 채점자 모수에서는 정보적 사전분포를 사용한 문항 모수에 비하여 수렴되지 않는 사례가 더 많이 발생하였고, 채점자 엄격성 모수 ϕ와 분산성 모수 ψ의 수렴 결과를 제시하면 아래 <표 Ⅳ-3>과 같다.

<표 Ⅳ-3> 채점자 모수 수렴 비율

국가	모수	HRM		E-HRM		대안1		대안2	
		모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율
덴마크	ϕ	16	1.000	16	1.000	361	.970	361	.884
	ψ	16	1.000	361	.970	16	1.000	361	.898
조지아	ϕ	12	1.000	12	1.000	348	.931	348	.940
	ψ	12	1.000	348	.937	12	1.000	348	.931
홍콩	ϕ	12	1.000	12	1.000	348	.968	348	.914
	ψ	12	1.000	348	1.000	12	.917	348	.943
헝가리	ϕ	6	1.000	6	1.000	174	.914	174	.828
	ψ	6	.833	174	1.000	6	.667	174	.862
아이슬란드	ϕ	11	1.000	11	1.000	316	.921	316	.889
	ψ	11	1.000	316	1.000	11	1.000	316	.921
라트비아	ϕ	10	1.000	10	1.000	218	.968	218	.867
	ψ	10	1.000	218	.982	10	1.000	218	.904

국가	모수	HRM		E-HRM		대안1		대안2	
		모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율	모수수	수렴 비율
리투아니아	ϕ	2	1.000	2	1.000	58	.897	58	.793
	ψ	2	1.000	58	1.000	2	1.000	58	.897
룩셈부르크	ϕ	8	1.000	8	1.000	208	.832	208	.798
	ψ	8	1.000	208	.990	8	.750	208	.889
네덜란드	ϕ	8	1.000	8	1.000	181	.862	181	.796
	ψ	8	1.000	181	1.000	8	.625	181	.818
뉴질랜드	ϕ	11	1.000	11	1.000	318	.940	318	.969
	ψ	11	1.000	318	.984	11	.909	318	.972
노르웨이	ϕ	14	1.000	14	1.000	406	1.000	406	.943
	ψ	14	1.000	406	.946	14	1.000	406	.926
루마니아	ϕ	8	1.000	8	1.000	139	.993	139	.971
	ψ	8	1.000	139	.978	8	1.000	139	.964
러시아	ϕ	2	1.000	2	1.000	58	.776	58	.759
	ψ	2	1.000	58	1.000	2	.500	58	.828
슬로바키아	ϕ	11	1.000	11	1.000	307	.954	307	.932
	ψ	11	1.000	307	1.000	11	.909	307	.961
슬로베니아	ϕ	9	1.000	9	1.000	218	.917	218	.849
	ψ	9	1.000	218	.991	9	.889	218	.872
스웨덴	ϕ	14	1.000	14	1.000	406	1.000	406	.924
	ψ	14	1.000	406	.990	14	1.000	406	.951
마케도니아	ϕ	22	1.000	22	1.000	638	.926	638	.958
	ψ	22	1.000	638	.976	22	.773	638	.972
미국	ϕ	7	1.000	7	1.000	196	1.000	196	.888
	ψ	7	1.000	196	1.000	7	1.000	196	.949
영국	ϕ	7	1.000	7	1.000	163	.969	163	.963
	ψ	7	1.000	163	1.000	7	1.000	163	.982
스코틀랜드	ϕ	6	1.000	6	1.000	147	1.000	147	.891
	ψ	6	1.000	147	.986	6	1.000	147	.905
벨기에 (네덜란드어권)	ϕ	6	1.000	6	1.000	174	.966	174	.897
	ψ	6	1.000	174	.994	6	1.000	174	.954
벨기에 (프랑스어권)	ϕ	8	.875	8	1.000	232	1.000	232	.991
	ψ	8	.750	232	1.000	8	1.000	232	1.000
아이슬란드 (5학년)	ϕ	11	1.000	11	1.000	296	.909	296	.801
	ψ	11	1.000	296	.986	11	1.000	296	.878
노르웨이 (5학년)	ϕ	13	1.000	13	1.000	377	.931	377	.881
	ψ	13	1.000	377	.905	13	1.000	377	.873

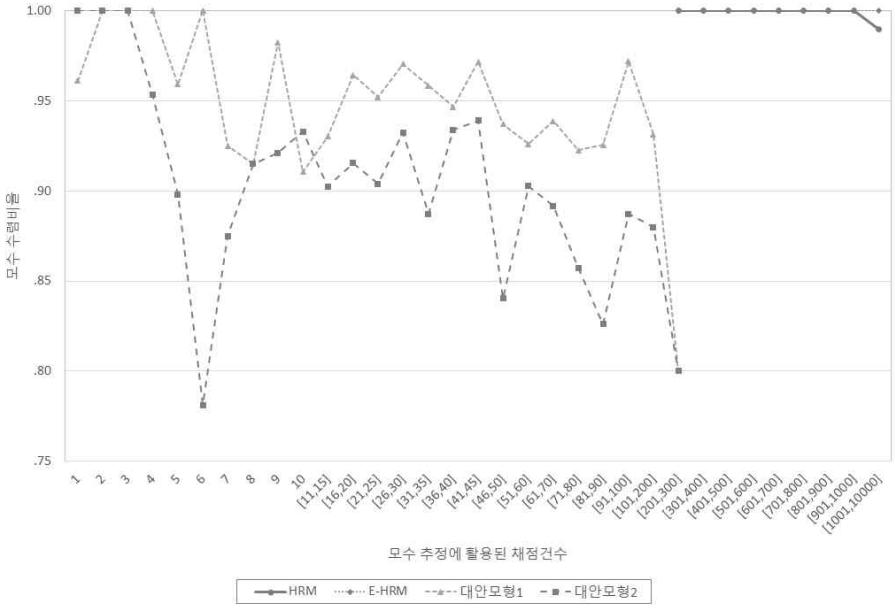
채점자 모수 중 엄격성 모수 ϕ의 경우, HRM에서는 벨기에(네덜란드어권)의 분석 결과에서만 1개 모수가 수렴하지 않는 것으로 확인되었고, E-HRM에서는 모든 엄격성 모수 ϕ_r이 적절하게 수렴된 것으로 나타났다. 그러나 문항별 채점자 모수 ϕ_{ir}를 추정한 대안모형1에서는 5개 국가에서만 엄격성 모수가 모두 수렴하는 것으로 분석되었고, 채점자 엄격성 모수와 분산성 모수를 모두 문항별로 추정한 대안모형2에서는 모든 국가에서 수렴되지 않은 엄격성 모수가 나타나 채점자 엄격성 모수를 문항 단위에서 분석할 경우 MCMC 연쇄의 수렴이 어렵다는 것을 확인할 수 있었다.

본 연구에서 가장 관심을 가지는 모수인 분산성 모수 ψ의 경우, 엄격성 모수의 수렴 비율이 심각하게 나빴던 대안모형2를 제외한 나머지 모형들에서 엄격성 모수 ϕ에 비해 수렴 비율이 떨어지는 것으로 나타났다. 모형별로는 HRM에서 2개국 3개 모수가 수렴하지 않는 것으로 나타나 가장 양호하였다. 그리고 E-HRM에서는 12개국 133개 모수, 대안모형1에서는 9개국 17개 모수, 대안모형2에서는 벨기에(프랑스어권)을 제외한 23개국 464개 모수가 MCMC 연쇄의 수렴에 도달하지 못하였다.

주로 채점자-문항 독립성 제약을 완화한 모형들에서 채점자 분산성 모수가 수렴되지 않은 것으로 나타난 원인을 추측해보자면, 모수의 분포가 단봉 분포(unimodal distribution)가 아니어서 정규분포를 가정하는 $\hat{R}$ 통계량이 충족되기 어렵거나 모수 추정을 위한 채점건수가 부족한 것이 원인이 될 수 있다. 이와 관련해서는 엄밀하게는 모의실험 연구로 확인이 필요한 사항이고, 적용 연구로는 채점건수에 따른 모수의 MCMC 연쇄 수렴 여부를 검토해볼 수 있다. 모수 추정에 활용된 채점건수에 따른 채점자 엄격성 모수의 MCMC 연쇄 수렴 비율은 <표 Ⅳ-4>, [그림 Ⅳ-1]과 같다.

<표 Ⅳ-4> 채점건수에 따른 채점자 엄격성 모수 수렴 비율

채점건수 범위	HRM	E-HRM	대안모형1	대안모형2
1			.961	1.000
2			1.000	1.000
3			1.000	1.000
4			1.000	.953
5			.959	.898
6			1.000	.781
7			.925	.875
8			.915	.915
9			.982	.921
10			.910	.933
[11,15]			.930	.902
[16,20]			.964	.915
[21,25]			.952	.904
[26,30]			.970	.932
[31,35]			.958	.887
[36,40]			.947	.934
[41,45]			.972	.939
[46,50]			.937	.840
[51,60]			.926	.903
[61,70]			.939	.892
[71,80]			.922	.857
[81,90]			.925	.826
[91,100]			.972	.887
[101,200]			.931	.880
[201,300]	1.000	1.000	.800	.800
[301,400]	1.000	1.000		
[401,500]	1.000	1.000		
[501,600]	1.000	1.000		
[601,700]	1.000	1.000		
[701,800]	1.000	1.000		
[801,900]	1.000	1.000		
[901,1000]	1.000	1.000		
[1001,10000]	.989	1.000		



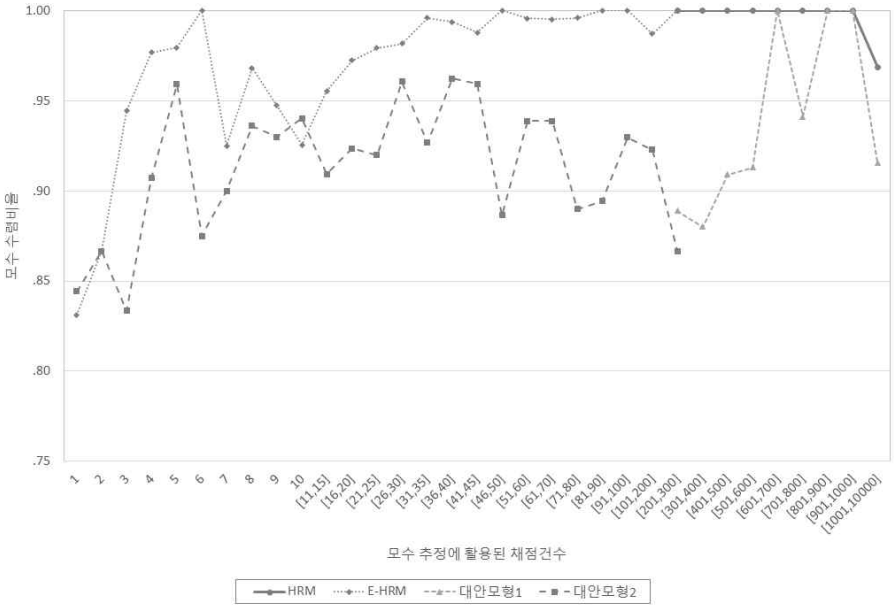
[그림 Ⅳ-1] 채점건수에 따른 채점자 엄격성 모수 수렴 비율

<표 Ⅳ-4>, [그림 Ⅳ-1]을 살펴보면 HRM는 모수 추정에 활용된 채점건수가 1,311건인 1개를 제외하고 엄격성 모수가 모두 수렴되었고, E-HRM은 모든 엄격성 모수가 수렴되는 것으로 나타나 엄격성 모수를 채점자 단위에서만 추정한 모형들에서는 채점건수의 영향은 나타나지 않았다. 반면에 엄격성 모수를 채점자-문항 단위에서 추정한 대안모형들에서는 진폭이 있지만 모수 추정에 활용한 채점건수가 많을수록 MCMC 연쇄의 수렴 비율이 낮다는 점에서 모수 추정에 활용된 채점건수가 수렴 비율에 직접적으로 영향을 주었다기보다 분산성 모수나 다른 요인의 영향으로 수렴이 어려웠던 것으로 추측된다.

다음으로 모수 추정에 활용된 채점건수에 따른 채점자 분산성 모수의 MCMC 연쇄 수렴 비율을 제시하면 다음 <표 Ⅳ-5>, [그림 Ⅳ-1]과 같다.

<표 Ⅳ-5> 채점건수에 따른 채점자 분산성 모수 수렴 비율

채점건수 범위	HRM	E-HRM	대안모형1	대안모형2
1		.831		.844
2		.867		.867
3		.944		.833
4		.977		.907
5		.980		.959
6		1.000		.875
7		.925		.900
8		.968		.936
9		.947		.930
10		.925		.940
[11,15]		.955		.909
[16,20]		.972		.923
[21,25]		.979		.920
[26,30]		.982		.961
[31,35]		.996		.927
[36,40]		.994		.962
[41,45]		.988		.959
[46,50]		1.000		.887
[51,60]		.996		.939
[61,70]		.995		.939
[71,80]		.996		.890
[81,90]		1.000		.894
[91,100]		1.000		.930
[101,200]		.987		.923
[201,300]	1.000	1.000	.889	.867
[301,400]	1.000		.880	
[401,500]	1.000		.909	
[501,600]	1.000		.913	
[601,700]	1.000		1.000	
[701,800]	1.000		.941	
[801,900]	1.000		1.000	
[901,1000]	1.000		1.000	
[1001,10000]	.968		.916	



[그림 Ⅳ-2] 채점건수에 따른 채점자 분산성 모수 수렴비율

채점자별로 ψ_r을 추정한 HRM에서는 채점건수가 1천 건이 넘는 모수 중 2건(모두 프랑스어권 벨기에)만이 수렴되지 않는 것으로 나타났다. 반면에 역시 채점자별로 ψ_r을 추정한 대안모형1에서는 모수 추정에 활용된 채점건수가 200건을 초과하는데도 불구하고 수렴 비율이 낮게 나타나 수렴 비율이 낮았던 엄격성 모수와의 관련성이 있을 것으로 추측해볼 수 있겠다. 그리고 채점자-문항별로 ψ_{ir}을 추정한 E-HRM과 대안모형2는 채점건수가 10건 미만일 때는 불안정한 수렴 비율을 보이다가, E-HRM은 채점건수 증가에 따라서 점차 수렴 비율이 1에 근사하게 되었고, 대안모형2는 채점건수가 증가하여도 85~95% 수준에 머물러 MCMC 연쇄가 적절히 수렴하지 않았다.

이상의 모수 수렴을 검토하였을 때, 엄격성 모수에 대한 채점자-문항 독립성 제약이 완화된 두 가지 대안모형은 엄격성 모수뿐만 아니라 분산성 모수의 MCMC 연쇄 수렴에 문제가 있고, HRM과 E-HRM에서는 안정적이었던 문항모수마저 수렴에 문제가 있어 실제 MCMC 분석에서의 추정 문제로 인하여 사용이 어렵다고 판단된다. 따라서 이후 분석에서는

대안모형1, 대안모형2에 대한 결과는 더 이상 제시하지 않고, HRM과 E-HRM을 중심으로 논의를 진행하겠다. 그리고 채점자 모수 특히 채점자 분산성 모수의 MCMC 연쇄 수렴에 큰 영향이 있는 것으로 나타난 모수 추정에 활용된 채점건수는 모의실험 연구에서 주요한 설계 요인으로 반영할 예정이다.

2. 모형 적합도

국가별로 연구 모형 HRM과 E-HRM의 적합도를 비교하기 위하여 국가 및 모형별로 Deviance 통계량, pD 통계량을 산출하였고, 두 지수의 합산으로 DIC를 획득하였다. 각각의 결과는 <표 Ⅳ-6>, <표 Ⅳ-7>, <표 Ⅳ-8>에 제시하였다.

<표 Ⅳ-6> 모형별 Deviance 통계량

국가명	Deviance		Deviance 검정
	HRM	E-HRM	
덴마크	1001.4	875.4	1.000
조지아	8157.0	6756.8	<.001
홍콩	2613.2	2289.5	.676
헝가리	1126.9	968.8	.698
아이슬란드	2504.5	2008.5	<.001
라트비아	5961.7	5091.5	<.001
리투아니아	2827.8	2544.5	<.001
룩셈부르크	4140.1	3548.7	<.001
네덜란드	584.7	501.6	1.000
뉴질랜드	5696.8	5033.9	<.001
노르웨이	7007.8	6003.6	<.001
루마니아	928.5	854.0	1.000
러시아	1108.2	1010.2	<.001
슬로바키아	2769.2	2409.1	.006
슬로베니아	1848.8	1598.0	.025
스웨덴	5757.1	4788.7	<.001
마케도니아	2267.5	2047.4	1.000

국가명	Deviance		Deviance 검정
	HRM	E-HRM	
미국	5717.3	5316.7	<.001
영국	1429.7	1288.3	.793
스코틀랜드	1604.1	1471.2	.674
벨기에(네덜란드어권)	6590.4	5629.7	<.001
벨기에(프랑스어권)	1988.8	1742.8	.150
아이슬란드(5학년)	2515.2	2141.0	<.001
노르웨이(5학년)	6157.2	5196.9	<.001

<표 Ⅳ-7> 모형별 pD 통계량

국가명	pD	
	HRM	E-HRM
덴마크	120.3	292.4
조지아	3535.4	4635.5
홍콩	549.8	665.1
헝가리	155.3	176.3
아이슬란드	393.5	506.1
라트비아	2016.0	2364.8
리투아니아	447.0	473.5
룩셈부르크	1290.0	1198.2
네덜란드	49.7	87.2
뉴질랜드	1390.6	1696.8
노르웨이	2868.3	3544.4
루마니아	69.2	172.7
러시아	102.3	118.8
슬로바키아	544.2	624.5
슬로베니아	340.4	320.0
스웨덴	1791.8	2064.5
마케도니아	606.0	1194.8
미국	1292.5	1690.9
영국	137.7	259.0
스코틀랜드	286.0	289.0
벨기에(네덜란드어권)	1909.4	3445.4
벨기에(프랑스어권)	479.7	295.9
아이슬란드(5학년)	509.7	592.1
노르웨이(5학년)	2912.4	3630.3

<표 Ⅳ-8> 모형별 DIC 통계량 및 DIC 최소 모형

국가명	Deviance Information Criteria		DIC 최소 모형
	HRM	E-HRM	
덴마크	1121.7	1167.8	HRM
조지아	11692.4	11392.3	E-HRM
홍콩	3162.9	2954.6	E-HRM
헝가리	1282.2	1145.1	E-HRM
아이슬란드	2898.0	2514.6	E-HRM
라트비아	7977.8	7456.3	E-HRM
리투아니아	3274.8	3018.1	E-HRM
룩셈부르크	5430.0	4746.9	E-HRM
네덜란드	634.4	588.8	E-HRM
뉴질랜드	7087.4	6730.7	E-HRM
노르웨이	9876.0	9548.1	E-HRM
루마니아	997.7	1026.7	HRM
러시아	1210.5	1129.0	E-HRM
슬로바키아	3313.4	3033.6	E-HRM
슬로베니아	2189.2	1917.9	E-HRM
스웨덴	7548.9	6853.2	E-HRM
마케도니아	2873.6	3242.2	HRM
미국	7009.8	7007.6	E-HRM
영국	1567.5	1547.4	E-HRM
스코틀랜드	1890.1	1760.2	E-HRM
벨기에(네덜란드어권)	8499.8	9075.1	HRM
벨기에(프랑스어권)	2468.5	2038.7	E-HRM
아이슬란드(5학년)	3024.8	2733.1	E-HRM
노르웨이(5학년)	9069.7	8827.2	E-HRM

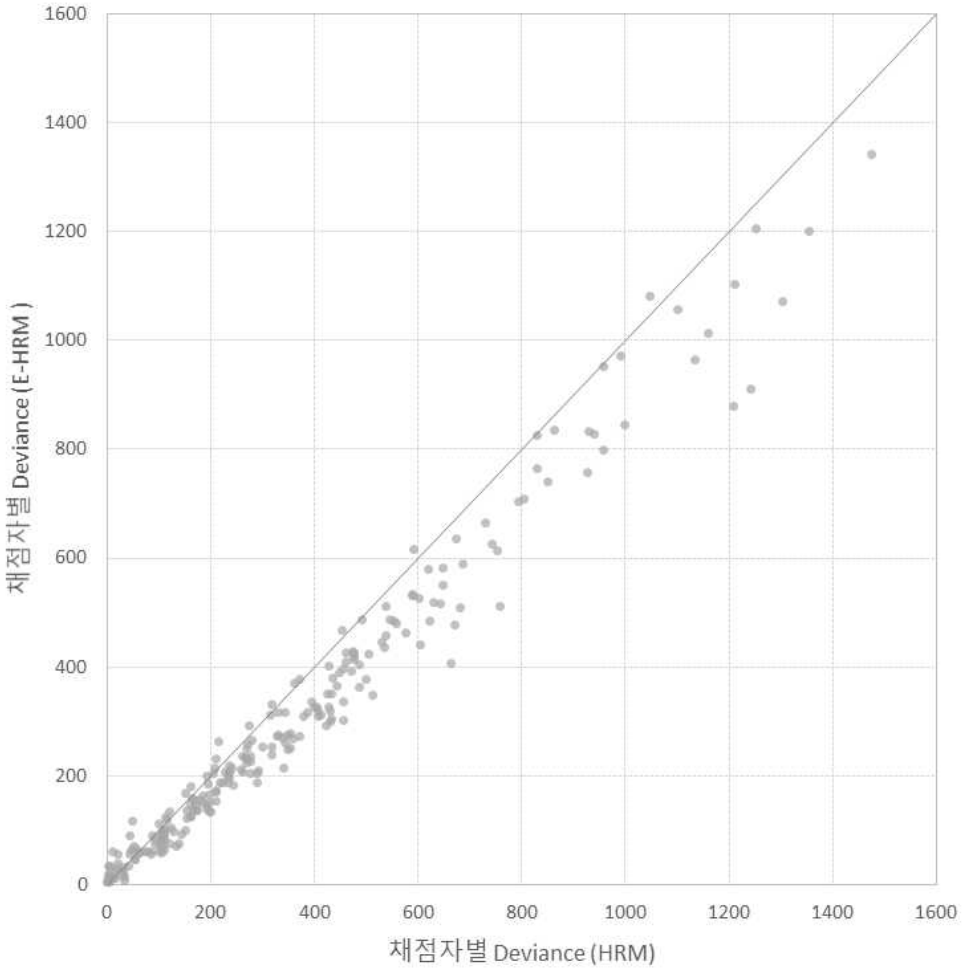
먼저, <표 Ⅳ-6>의 Deviance 통계량의 경우, 두 모형 간의 내재 관계가 있다면 통계적 검정이 가능하다. HRM의 경우 E-HRM에서 채점자 분산성 모수를 제약한 모형이므로 HRM이 E-HRM에 내재되는 모형이다. 이에 Deviance 검정을 실시한 결과, 유의수준 .05에서 15개 국가(조

지아, 아이슬란드, 라트비아, 리투아니아, 룩셈부르크, 뉴질랜드, 노르웨이, 러시아, 슬로바키아, 슬로베니아, 스웨덴, 미국, 벨기에-네덜란드어권, 아이슬란드-5학년, 노르웨이-5학년)에서 E-HRM이 모형의 모수 증가를 감안하더라도 HRM에 비해 통계적으로 유의한 적합도 개선이 있는 것으로 나타났다. 하지만 이 15개 국가들 중 E-HRM의 모든 모수가 적절하게 수렴된 것으로 나타난 국가는 4개 국가에 불과했기에 모형 선택의 기준으로 활용하지 않을 것이다.

<표 Ⅳ-8>의 DIC 지수를 국가 및 모형별로 비교한 결과, 모수 증가로 인한 패널티를 반영하고도 24개국 중 4개(덴마크, 루마니아, 마케도니아, 벨기에-네덜란드어권) 국가를 제외한 20개국에서 모든 문항별로 분산성 모수 ψ_{ir}를 추정한 E-HRM의 적합도가 기존 HRM에 비해서 높은 것으로 나타났다. 한편, Deviance로부터 산출되었다는 점에서 DIC 역시 Deviance를 채점자 단위에서 분해하여 살펴보면 절대적인 모형 선택 기준으로 삼기에는 제한점이 있다.

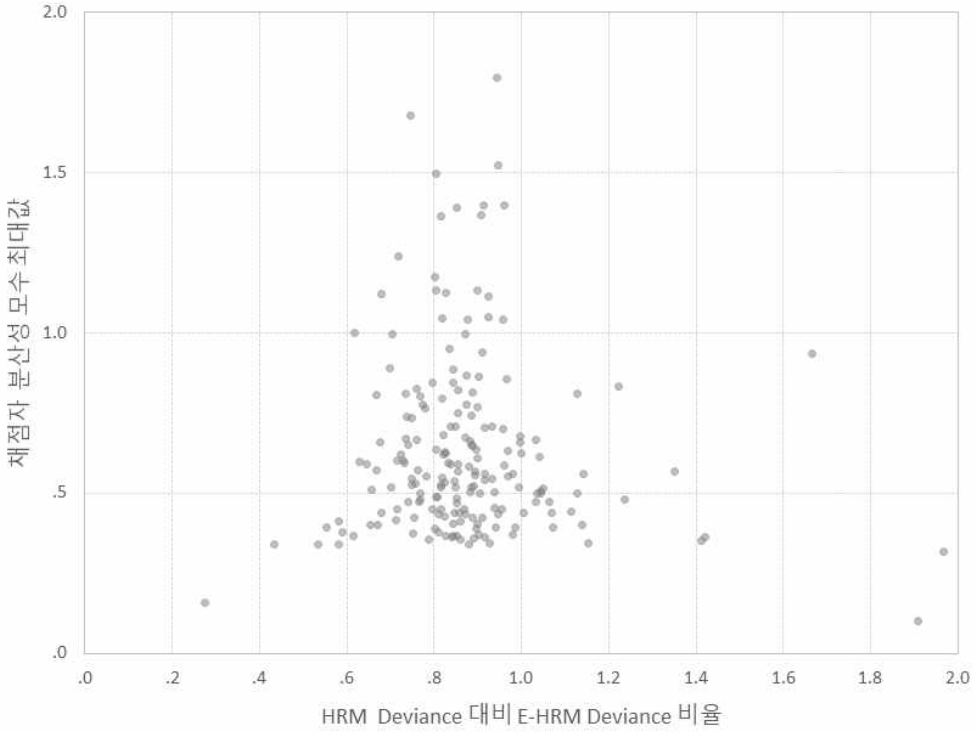
국가별, 모형별로 차이를 보이는 적합도 지수들의 원인을 세부적으로 살펴보기 위하여 채점자 단위의 Deviance를 분석해보았다. 채점자 단위 Deviance는 JAGS 소프트웨어 분석에서 모든 채점 단위의 Deviance를 추출하고, 해당 채점을 한 채점자를 기준으로 합산하여 채점자별 Deviance 통계량을 얻고 모형 간 비교를 진행하였다. 전체 채점자에 대한 Deviance 분석 결과 값은 <부록표 1>에 제시하였고, 간략하게 도표로 제시하면 [그림 Ⅳ-3]과 같다.

HRM의 Deviance 통계량을 x-축으로 E-HRM의 Deviance 통계량을 y-축으로 표시한 [그림 Ⅳ-3]에서 대다수의 채점자들은 $y=x$인 기준선 아래에 위치해있어 E-HRM에서 HRM보다 채점자의 Deviance 통계량이 감소하여 적합도가 향상된 것을 볼 수 있고, 이는 모형이 복잡해지면서 채점 자료에 대한 모형의 적합도가 향상된 것이라 하겠다.



[그림 Ⅳ-3] HRM 및 E-HRM 채점자별 Deviance

HRM에 비하여 E-HRM의 채점자 Deviance 통계량이 높은 비율로 감소했다는 것은 문항별로 채점자 분산성 모수를 추정한 모형으로 인하여 자료가 더 잘 설명된다는 것이고, 이는 문항별로 채점자 분산성 모수에 차이가 존재한다는 것이므로 채점자의 채점 특성이 문항에 따라서 다른가를 판별하는데 유용한 정보가 될 수 있다. 이를 확인하기 위하여 HRM 대비 E-HRM Deviance 비율에 따른 채점자 분산성 모수 최댓값의 분포를 [그림 Ⅳ-4]와 같이 제시하였다.



[그림 Ⅳ-4] HRM 대비 E-HRM Deviance 비율에 따른 채점자 분산성 모수 최댓값

HRM 대비 E-HRM Deviance 비율이 1을 넘는 채점자가 소수이긴 하지만, 이 채점자들에서는 분산성 모수 최댓값이 정상 채점 기준인 .5를 넘는 경우가 많지 않았다. 하지만, HRM 대비 E-HRM Deviance 비율이 1 이하로 내려가면서 채점자들의 분산성 모수 최댓값이 .5를 넘는 경우가 매우 높아졌고, 도표 제시의 편의상 분산성 모수의 최댓값만을 표시하였지만 전체 분산성 모수 ψ_{ir}을 모두 고려한다면 더 많은 분산성 모수가 정상 채점 기준인 .5를 넘는 것을 확인할 수 있다.

그런데 [그림 Ⅳ-3]에서 $y=x$인 기준선 위에 위치한 채점자 혹은 [그림 Ⅳ-4]에서 HRM 대비 E-HRM Deviance 비율이 1을 넘는 채점자들도 일부 관찰할 수 있고, 이 채점자들은 HRM보다 E-HRM에서 오히려 Deviance 통계량이 높아져 적합도가 나빠진 채점자들로 볼 수 있고, 전체 234명의 채점자 중 44명의 채점자가 모수의 증가로 인하여 자료에 대

한 모형의 적합 가능성이 높아졌는데도 불구하고 Deviance 통계량이 높아져 적합도가 나빠졌다. 이렇게 HRM보다 E-HRM에서 적합도가 악화된 채점자를 해당 채점자의 전체 채점건수와 Deviance 통계량 정보와 함께 제시하면 <표 Ⅳ-9>와 같다.

<표 Ⅳ-9> E-HRM 적합도 악화 채점자

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
덴마크	1	230	42.0	56.8
	3	257	0.3	7.6
	6	585	8.2	61.9
	8	226	3.0	18.5
	11	1,131	114.1	121.4
	12	292	16.0	29.2
	15	322	3.8	34.5
	16	392	0.2	13.3
조지아	12	328	206.4	215.3
홍콩	11	236	0.3	12.2
	12	563	0.7	12.1
헝가리	4	1,888	159.5	181.7
아이슬란드	1	543	1.6	35.3
라트비아	7	875	213.0	263.5
룩셈부르크	3	1,498	591.1	616.4
네덜란드	1	551	20.3	57.5
	2	1,198	46.0	64.9
	5	269	0.6	8.8
뉴질랜드	2	1,635	452.9	468.6
	10	334	54.4	66.5
루마니아	1	1,579	98.6	113.7
	3	867	85.0	91.2
슬로바키아	5	903	42.5	90.4
	6	1,081	315.9	331.9
	1	941	118.8	135.6
	9	379	20.6	40.6
마케도니아	4	480	112.2	124.0
	11	283	47.9	116.8
	14	309	14.6	24.3
	15	393	29.8	33.6

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
	17	389	0.1	4.7
	18	323	0.1	5.4
	19	300	11.7	12.3
	20	293	2.9	5.5
미국	3	2,766	1047.1	1082.5
영국	2	340	2.2	19.4
	4	1,954	359.4	371.4
스코틀랜드	1	1,182	274.0	292.9
	5	1,140	51.0	72.4
벨기에(프랑스어권)	2	1,120	149.8	168.8
	3	1,149	208.2	232.0
아이슬란드(5학년)	5	771	191.1	199.6
	9	1,315	369.5	379.3
	10	1,178	204.6	205.5

분석 대상 24개국 중 17개국에서 최소한 1명의 채점자는 HRM보다 E-HRM에서 적합도가 악화된 것으로 나타났고, 채점자가 22명으로 가장 많았던 마케도니아와 채점자가 16명으로 두 번째로 많았던 덴마크에서는 각각 8명의 채점자들이 적합도가 악화된 결과를 보였다. 이와 같이 E-HRM에서 Deviance가 악화된 원인을 찾고자 이 채점자들의 채점 특성을 제시하면 <표 Ⅳ-10>과 같다.

<표 Ⅳ-10> E-HRM 적합도 악화 채점자 채점 특성

국가	채점자	Deviance	채점건수		채점단위 최대 Deviance
		E-HRM	평균	최소	
덴마크	1	56.8	9.6	1	11.5
	3	7.6	18.4	12	1.6
	6	61.9	27.9	2	5.1
	8	18.5	16.1	12	4.9
	11	121.4	39.0	13	7.2
	12	29.2	10.8	1	3.6
	15	34.5	13.4	1	4.3
	16	13.3	14.0	1	.8
조지아	12	215.3	11.3	4	4.9

국가	채점자	Deviance	채점건수		채점단위 최대 Deviance
		E-HRM	평균	최소	
홍콩	11	12.2	8.1	5	3.3
	12	12.1	19.4	8	2.6
헝가리	4	181.7	65.1	51	9.7
아이슬란드	1	35.3	20.9	12	4.1
라트비아	7	263.5	62.5	39	15.2
룩셈부르크	3	616.4	51.7	12	9.7
네덜란드	1	57.5	36.7	4	7.1
	2	64.9	47.9	19	7.6
	5	8.8	14.9	3	4.4
뉴질랜드	2	468.6	56.4	43	7.2
	10	66.5	11.5	1	6.5
루마니아	1	113.7	54.4	1	10.7
	3	91.2	61.9	1	10.4
슬로바키아	5	90.4	31.1	22	5.8
	6	331.9	37.3	30	10.6
	1	135.6	36.2	9	6.8
	9	40.6	47.4	35	6.5
마케도니아	4	124.0	16.6	10	6.6
	11	116.8	9.8	7	3.8
	14	24.3	10.7	8	5.6
	15	33.6	13.6	9	7.0
	17	4.7	13.4	10	.1
	18	5.4	11.1	3	.2
	19	12.3	10.3	8	4.2
	20	5.5	10.1	7	.4
미국	3	1082.5	95.4	69	16.0
영국	2	19.4	18.9	9	2.9
	4	371.4	67.4	8	13.1
스코틀랜드	1	292.9	51.4	6	9.1
	5	72.4	60.0	36	8.2
벨기에(프랑스어권)	2	168.8	38.6	23	14.1
	3	232.0	39.6	26	8.1
아이슬란드(5학년)	5	199.6	30.8	14	13.8
	9	379.3	52.6	1	14.6
	10	205.5	40.6	8	9.6

가장 먼저 눈에 띄는 특징은 E-HRM에서 적합도가 나빠진 44명중 20

명이 문항당 평균 채점건수가 20건 이하였고, 27명은 문항당 최소 채점건수가 10건 이하인 경우였다는 점이다. 이와 같이 채점자-문항당 채점건수가 적은 경우, MCMC 연쇄가 잘 수렴되지 않거나 MCMC 연쇄가 수렴 되었다하더라도 연쇄의 중앙값이 모수를 대표하는 점 추정치로 부적절하거나 자료를 잘 설명하지 못하는 것으로 보인다.

그리고 채점 1건에 지나치게 Deviance 통계량이 큰 경우들이 발견되는데, 네덜란드 5번 채점자의 경우에는 채점자의 Deviance 통계량 총합이 8.8인데 하나의 채점이 가지는 Deviance 통계량이 4.4로 채점자 deviance의 49.6%에 이른다. 게다가 9,981건의 채점건수를 가지는 네덜란드 자료 전체의 Deviance 통계량인 501.6의 1%에 가까운 비율을 모형에 따른 우도(likelihood)가 낮은 단 1건의 채점 사례가 차지한 것이다. 이 사례 외에도 <표 Ⅳ-10>에서 제시된 채점자들의 단일 채점이 가지는 최대 Deviance 통계량은 최대 16을 넘고, 평균이 7.04로 MCMC 추정에서의 잘못된 대푯값 혹은 자료를 잘 설명하지 못하는 대푯값으로 인하여 Deviance 통계량이 얼마나 왜곡될 수 있는가를 보여준다. 따라서 Deviance 통계량과 Deviance 통계량으로부터 얻어지는 DIC의 활용에도 주의를 기울일 필요가 있고, 특히 최적 모형 선택에서 소수의 이상치로 인하여 잘못된 판단을 내리지 않도록 주의해야할 것이다.

Ⅴ. 연구 결과: 국가별 분석

이 장에서는 국가별 자료에 대해 기존의 HRM과 본 연구에서 제안하는 E-HRM을 적용하여 분석한 채점자 효과를 중심으로 살펴봄으로써 HRM과 비하여 E-HRM이 가지는 장점과 유용성을 제시하고자 한다. 그리고 실제 구성형 문항 채점 자료에 대한 E-HRM의 적용 가능성과 제한점을 탐색해보겠다. 이를 위하여 24개국 중 3개 국가에 대한 구체적 분석 결과를 살펴볼 예정인데, 첫 번째로 HRM과 E-HRM 모두 적절하게 수렴되어 분석이 가능한 아이슬란드, 두 번째로 E-HRM만 적용 가능한 벨기에(프랑스어권), 마지막 세 번째로 HRM만 적용 가능한 덴마크의 분석 사례를 제시하였다.

1. 아이슬란드

가. 모형 평가

아이슬란드 자료에 HRM과 E-HRM을 모두 적용해 본 결과, 두 가지 모형이 모두 적절하게 수렴하는 것으로 나타났다. 그리고 모형 적합도에 있어서는 1수준인 반응 수준에서 채점자-문항별로 채점자 분산성(ψ_{ir})를 추정하는 E-HRM의 Deviance(2008.5)와 DIC(2514.6)가 채점자별로만 채점자 분산성(ψ_i)을 추정하는 HRM의 Deviance(2504.5)와 DIC(2898.0)보다 더 낮은 것으로 나타나 E-HRM이 HRM보다 아이슬란드 자료를 좀 더 잘 설명하는 모형으로 판단된다. 따라서 아이슬란드 채점 자료에 대한 분석 결과는 채점자의 특성이 문항별로 차이가 존재한다고 가정하고 문항별로 채점자 분산성 모수(ψ_{ir})를 추정하는 E-HRM의 분석 결과에 중점을 두고, 채점자 특성이 문항별 차이가 없다는 가정하에 채점자 분산성 모수(ψ_r)를 추정하는 HRM 분석 결과와 비교 제시할 것이다.

나. 채점 설계

아이슬란드 자료는 전체 29개 문항에 대한 피험자 674명의 응답 9,032건을 총 11명의 채점자가 채점한 결과이다. 채점 설계에 있어서는 모든 채점자가 모든 문항을 채점하지는 않았지만, 공통 채점자와 공통 피험자를 통해 채점자 모수, 문항 모수, 피험자 능력 모수가 동일한 척도 상에서 비교 가능하도록 채점 설계가 이루어졌다. 채점 설계의 특징은 다음 <표 Ⅴ-1>에서 채점자별 채점인원, 채점건수, 채점문항수를 통해 확인할 수 있듯이 모든 채점자가 동일한 인원, 문항수, 건수를 채점하는 균형 설계(balanced design)가 아닌 불균형 설계(unbalanced design)이다. 이러한 불균형 설계에 따라서 채점자 간에 채점 부담에 차이가 있었는데, 채점 부담이 가장 큰 5번 채점자는 1,189건을 채점한 것으로 나타나 채점 부담이 가장 적은 6번 채점자의 474건의 2배 이상의 답안을 채점하였다.

<표 Ⅴ-1> 채점자별 채점 배정: 아이슬란드

채점자	채점인원	채점건수	채점문항수
1	79	543	26
2	168	1,115	29
3	148	975	29
4	130	857	29
5	174	1,189	29
6	69	474	29
7	86	577	29
8	115	771	29
9	106	689	29
10	157	1,049	29
11	116	793	29

문항-채점자 배정 설계를 살펴보면, 배정 설계상의 전체 319개 셀(문

항 29개 × 채점자 11명) 중 99.9%인 총 316개 셀에 채점이 배정되어 문항-채점자 설계에 있어서는 완전구획설계(complete block design)에 가까운 설계이고, 자세한 문항-채점자별 채점인원은 <부록표 2>에서 확인할 수 있다. 비록 전체 문항-채점자 설계 셀에 채점 배정이 되는 설계 방식은 아니었지만, 모든 문항은 공통 채점자를 통해서 연결되어 문항 모수가 동일한 척도 상에서 비교가 가능하고, 모든 채점자들 역시 앞서 [그림 Ⅲ-2]의 사회관계망 그래프 (a)에서 확인할 수 있듯이 채점자 간에 공통 피험자를 가지는 채점자-피험자 설계를 통해서 채점자 모수 역시 동일한 척도 상에서 비교가 가능하다.

다. 문항 모수

아이슬란드 자료에 대한 E-HRM과 HRM 분석 결과에서 공통적으로 GPCM으로 추정된 문항 모수를 비교하면 다음 <표 Ⅴ-2>와 같다. 앞서 연구 방법에서 언급한 바와 같이 문항 곤란도 모수와 단계 모수는 분리 추정하지 않았기 때문에 단계 모수와 결합된 문항 곤란도 모수(β_k)를 제시하였다.

<표 Ⅴ-2> 문항 모수: 아이슬란드

문항	HRM				E-HRM			
	α	β_2	β_3	β_4	α	β_2	β_3	β_4
1	1.289	-2.126	1.362		1.309	-2.111	1.345	
2	1.300	-2.408	.087		1.245	-2.352	.036	
3	.720	-.444	1.072		.741	-.415	1.026	
4	.623	-.021	.010		.618	-.026	-.016	
5	1.110	-.968	.628		1.104	-.983	.607	
6	.803	.310	-1.177		.831	.530	-1.389	
7	1.342	-2.428	-1.929		1.333	-2.494	-1.899	
8	.839	1.064	.165		.818	1.097	.121	
9	.865	-.438	1.815		.881	-.384	1.883	
10	.954	.147	.961		.945	.165	.904	
11	1.238	.139	2.322		1.125	.164	2.246	

문항	HRM				E-HRM			
	α	β_2	β_3	β_4	α	β_2	β_3	β_4
12	.809	.196	-1.288		.791	.183	-1.314	
13	.968	-1.160	-.555		.921	-1.153	-.593	
14	.716	-.268	2.382		.729	-.288	2.388	
15	.913	-1.204	1.496		.889	-1.213	1.524	
16	.707	-.577	-.787		.724	-.545	-.809	
17	.819	-.288	.723		.820	-.278	.706	
18	.836	.938	.288		.833	.943	.211	
19	1.205	.398	1.676		1.124	.437	1.446	
20	.715	.944	-1.173		.701	.908	-1.137	
21	.996	-1.113	.452		.927	-1.125	.468	
22	.958	-1.656	-1.457		.960	-1.656	-1.435	
23	.872	2.047	-1.033		.910	2.339	-1.357	
24	1.123	2.131	-.109		1.155	2.126	-.160	
25	.868	-1.221	.093	-1.166	.896	-1.224	.073	-1.120
26	.791	-1.091	.503	-.346	.798	-1.113	.568	-.353
27	1.014	-2.055	1.280	1.958	1.056	-2.109	1.229	1.872
28	.674	-1.497	.066	.935	.660	-1.430	.041	.925
29	1.157	-1.481	1.746	1.216	1.160	-1.492	1.732	1.225

문항 모수 분석 결과, 두 모형 HRM과 E-HRM의 문항 모수들 간의 관계는 상관계수가 .998(α모수 .985, β모수 .998)로 1에 가까운 매우 높은 상관을 보여 두 모형의 문항 수준에서 추정되는 모수의 경향성은 일치하는 것으로 볼 수 있겠다. 그리고 두 모형의 문항 모수들의 대소 관계를 살펴보면, α모수의 경우 HRM이 큰 경우가 15개, E-HRM가 큰 경우가 14개로 어느 한 모형의 문항모수가 일관되게 크거나 작은 추정치를 가지는 경향을 보이지는 않았다. 또한 β모수의 경우 모수 최댓값에 모수 최솟값을 뺀 범위값을 비교했을 때, HRM이 큰 경우가 14개, E-HRM가 큰 경우가 15개로 나타나 특정 모형에서 문항모수가 일관되게 작은 추정치를 가지는 수축 경향을 보이지는 않았다. HRM에서 지적되는 문제점 중 하나가 문항의 변별도 계수의 추정값이 자료 생성 과정에서 설정한 참값보다 일관되게 작게 추정되는 수축(shrinking) 문제인데(Kim, 2009), 본 연구에서는 채점자 모수를 문항 단위로 세분화해서 더 많이 추정하는

E-HRM에서 변별도 모수가 더 심하게 수축되는 문제가 우려되었다. 하지만 모형 간 문항 모수 비교를 통해서 변별도 계수의 추정값이 수축이 되었는지를 확인할 수는 없지만, 최소한 두 모형 간에 수축 정도의 차이가 나타나지 않은 것으로 판단할 수 있겠다.

문항 곤란도를 문항 단위에서 살펴보면, 29개 문항 중 12개 문항(6, 8, 12, 16, 18, 20, 23, 24, 25, 26, 29번 문항)에서 높은 점수(β_k)와 낮은 점수(β_{k-1})간의 모수가 역전되는 결과를 확인할 수 있었다. 이와 같이 문항 곤란도 모수의 역전되는 경우, 3점 척도 문항에서 2점과 같은 중간 점수의 확률이 크게 낮아질 수 있고, 모형의 채점자 모수 ϕ와 ψ 추정에도 영향을 줄 가능성이 높아 문제가 될 수 있다.

라. 채점자 모수

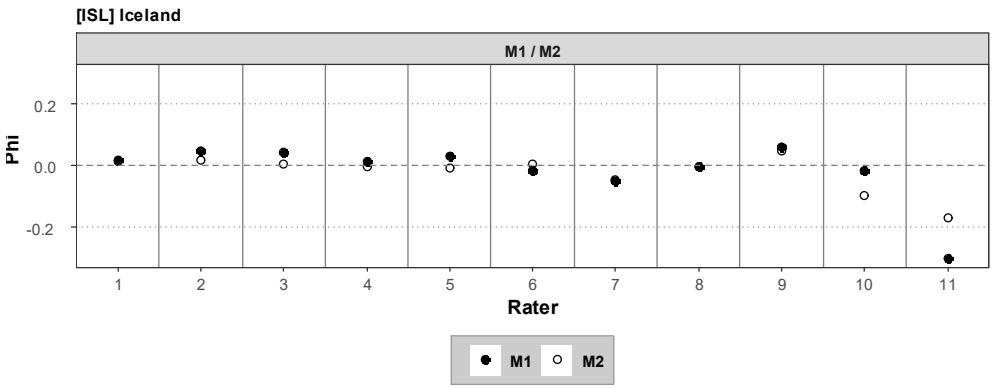
채점자 모수 분석 결과, E-HRM과 채점자 수준 모수에서 E-HRM에 내재되는 모형인 HRM의 채점자 모수를 <표 V-3>과 같이 제시하였다.

<표 V-3> 채점자 모수: 아이슬란드

모형	모수	채점자										
		1	2	3	4	5	6	7	8	9	10	11
HRM	ϕ_r	.017	.047	.045	.015	.032	-.015	-.050	-.002	.062	-.018	-.300
E-HRM	ϕ_r	.013	.019	.007	-.003	-.006	.004	-.048	-.004	.048	-.096	-.171
HRM	ψ_r	.055	.374	.373	.365	.345	.387	.370	.357	.368	.356	.240
E-HRM	$\psi_{1,r}$	.063	.062	.060	.344	.059	.065	.071	.062	.094	.066	.065
	$\psi_{2,r}$	.068	.180	.501	.398	.334	.646	.325	.149	.340	.424	.145
	$\psi_{3,r}$	.353	.436	.062	.425	.074	.065	.140	.356	.449	.068	.067
	$\psi_{4,r}$	.363	.385	.060	.060	.296	.068	.069	.262	.362	.309	.070
	$\psi_{5,r}$	.071	.369	.071	.251	.500	.450	.070	.312	.485	.339	.076
	$\psi_{6,r}$	.069	.573	.408	.336	.095	.845	.414	.177	.068	.297	.765
	$\psi_{7,r}$	.068	.064	.064	.061	.080	.594	.065	.067	.069	.066	.388
	$\psi_{8,r}$	.066	.517	.063	.168	.063	.078	.401	.067	.098	.530	.071
	$\psi_{9,r}$	.407	.064	.077	.059	.474	.138	.072	.066	.066	.066	.072
	$\psi_{10,r}$	.068	.118	.065	.098	.064	.065	.074	.548	.365	.069	.072

$\psi_{11,r}$	.080	.291	.438	.338	.115	.066	.074	.496	.313	.546	.100
$\psi_{12,r}$	.062	.060	.066	.488	.076	.068	.071	.074	.422	.068	.064
$\psi_{13,r}$	.061	.061	.069	.069	.420	.066	.070	.075	.133	.066	.064
$\psi_{14,r}$	.062	.059	.067	.076	.167	.068	.072	.406	.346	.066	.066
$\psi_{15,r}$	.087	.059	.064	.066	.058	.398	.070	.060	.070	.065	.065
$\psi_{16,r}$	.069	.511	.206	.119	.061	.068	.423	.064	.072	.065	.258
$\psi_{17,r}$	.068	.060	.064	.066	.435	.113	.070	.361	.069	.069	.067
$\psi_{18,r}$	.069	.081	.067	.067	.084	.736	.071	.065	.072	.399	.071
$\psi_{19,r}$		.062	.072	.420	.434	.224	.134	.099	.128	.510	.116
$\psi_{20,r}$		.078	.596	.103	.067	.068	.810	.496	.451	.274	.068
$\psi_{21,r}$		.064	.061	.472	.059	.064	.075	.064	.065	.067	.085
$\psi_{22,r}$	.073	.340	.064	.075	.064	.070	.103	.508	.068	.064	.070
$\psi_{23,r}$	.072	.682	.574	.092	.421	.067	.438	.652	.070	.072	.193
$\psi_{24,r}$	.071	.087	.145	.075	.063	.072	.505	.423	.413	.069	.357
$\psi_{25,r}$	.068	.063	.421	.092	.098	.066	.432	.064	.393	.065	.067
$\psi_{26,r}$	.068	.068	.062	.059	.063	.070	.071	.064	.559	.066	.065
$\psi_{27,r}$	.066	.062	.234	.395	.437	.524	.069	.064	.068	.064	.544
$\psi_{28,r}$	.068	.208	.064	.289	.126	.557	.320	.061	.068	.281	.064
$\psi_{29,r}$	.326	.461	.317	.424	.156	.311	.064	.067	.377	.390	.326

먼저 HRM, E-HRM 공통적으로 채점자별로 차이가 있고, 문항 간에는 차이가 없는 모수로 추정된 채점자 엄격성 모수 ϕ를 살펴보기 위하여 그래프로 제시하면, [그림 Ⅴ-1]과 같다.



[그림 Ⅴ-1] 채점자 엄격성 모수: 아이슬란드

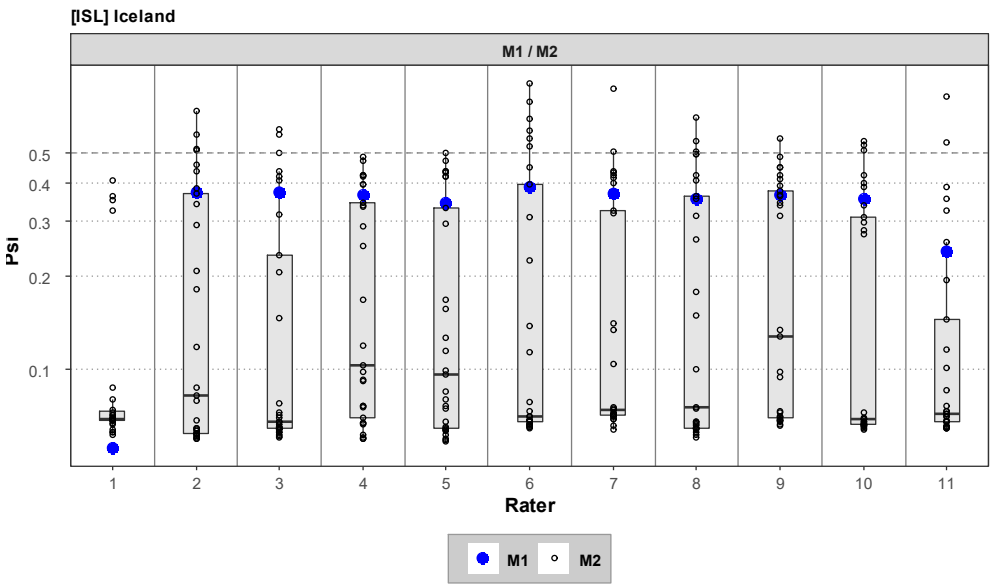
모형간 모수 추정 결과의 경향성을 살펴보면, E-HRM(M2)의 경우 채점자 엄격성 모수는 최대 .048, 최소 -.171로 추정되어 HRM(M1)의 채점자 엄격성(최대 .062, 최소 -.300)에 비하여 좁은 범위를 가졌고, 10번 채점자의 엄격성 모수를 제외하면 E-HRM의 엄격성 모수는 HRM과 비교하여 절댓값이 거의 같거나 작은 결과를 보였다. 한편, 11번 채점자의 HRM에서의 엄격성 모수는 E-HRM에서의 엄격성 모수와 .129 차이로 엄격성 모수 척도의 크기를 감안했을 때 작은 차이가 아니지만 엄격성 모수로 인한 평정 정확성의 변화는 분산성 모수에 따라 달라진다는 점에서 11번 채점자의 낮은 분산성 모수(ψ_{11}=.240)로 인하여 평정 정확성에 미치는 영향은 크지 않을 것으로 보인다.

채점자별로 엄격성 모수를 검토하면, 절댓값이 Patz, Junker, Johnson, & Mariano(2002)과 Casabianca & Wolfe(2017)에서 제시된 기준이자 진점수(η)보다 인접한 점수($\eta-1$ 혹은 $\eta+1$)로 평정할 확률이 높아지는 .5 이상을 초과하는 것으로 추정된 값은 없었고, 대부분 엄격성 모수의 절댓값이 .05 이내의 값을 나타내어 전반적으로 채점자 엄격성 모수는 양호한 수준인 것으로 판단된다. 방향성에 있어서는 11번 채점자의 HRM 엄격성 모수가 -.300로 비록 -.5~.5 범위 이내로 심각한 평정 편의는 아니지만 나머지 채점자들에 비하여 상당히 엄격하게 평정한 것으로 해석할 수 있다. 그리고 10번 채점자는 엄격성 모수가 -.096으로 역시 절댓값이 크지는 않지만 추정치가 음수로 나타나 나머지 채점자들에 비하여 조금 더 엄격한 평정을 한 것으로 볼 수 있다.

채점자별로만 분산성 모수 ψ_r을 추정한 HRM과 아이슬란드의 최적모형이자 채점자-문항별로 분산성 모수 ψ_{ir}를 추정한 E-HRM의 비교 결과를 그래프로 제시하면 [그림 Ⅴ-2]와 같다. ψ 모수 추정치 그래프 제시에 있어서는 ψ 값이 주로 0~.5 구간에 분포해있는데, 일부 최댓값이 .5를 벗어나는 큰 값으로 인하여 그래프상의 0~.5 구간에서 모수 추정치의 비교가 어려운 문제점이 있다. 이를 보완하기 위하여 ψ값에 해당하는 y축을 로그 척도로 변환하여 제시하였기에 동일한 ψ값 차이가 나더라도

그래프 길이가 다르게 표현되었다는 점에 주의해야 한다. 분산성 모수 그래프 상에서 HRM(M1)의 추정치 ψ_r은 큰 원으로 나타내었고, E-HRM(M2)의 추정치는 특정 채점자의 문항별 ψ_{ir}의 분포를 파악하기 쉽도록 상자도표를 제시하였고, 구체적인 문항별 추정치는 작은 원으로 표시하였다.

먼저, HRM에서의 채점자 분산성 모수를 살펴보면 최솟값은 1번 채점자로 .055로 추정되었고, 최댓값은 6번 채점자로 분산성 모수 ϕ가 .387로 추정되어 모든 채점자들의 분산성 모수가 Patz et al.(2002), Casabianca, Junker, & Patz (2016), Casabianca & Wolfe(2017) 등의 연구에서 제시한 평정의 일관성(consistency) 기준인 '$\psi < 0.5$'를 충족하였다. 따라서 HRM 모형의 분산성 모수 추정 결과에서는 채점자들이 평정에 있어서 이상 채점으로부터 크게 벗어나지 않는 안정적인 채점을 하고 있는 것으로 분석되었다.

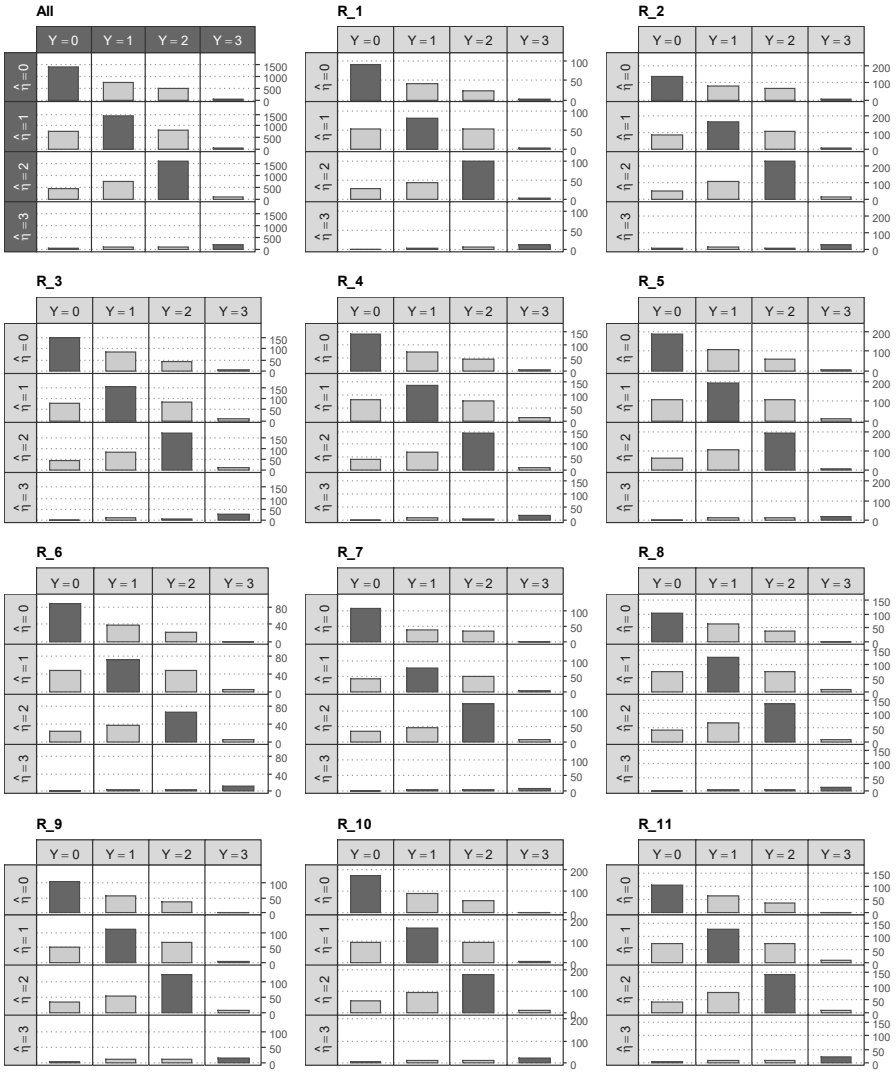


[그림 V-2] 채점자 분산성 모수: 아이슬란드

반면, 채점자-문항 단위에서 채점자 분산성 모수(ψ_{ir})를 추정한

E-HRM의 결과를 살펴보면, 모든 채점자들의 분산성 모수가 .5 이하의 값으로 추정되었던 HRM과 달리 분산성 모수가 .5를 초과하는 결과가 총 24건 발견되었다. 1번, 4번, 5번 채점자의 경우에는 문항 단위에서 분산성 모수의 범위가 [.058, .500]으로 나타나 모든 문항에서 일관성에 큰 문제가 없는 것으로 판단된다. 하지만 2번 채점자 4개 문항, 3번 채점자 3개 문항, 6번 채점자 6개 문항, 7번 채점자 2개 문항, 8번 채점자 3개 문항, 9번 채점자 1개 문항, 10번 채점자 3개 문항, 11번 채점자 2개 문항에서 채점자 분산성 모수 ψ가 .5를 초과하여 채점자 단위에서 분산성을 추정하였을 때와 채점자-문항 단위에서 분산성을 추정하였을 때 차이가 있을 수 있음을 보여준다. 이는 채점자의 평정 일관성 혹은 비일관성이 문항과 무관한 것이 아니라 문항에 따라 달라질 수 있음을 나타내는 것이고, HRM에서 드러나지 않았던 채점자의 비일관성을 E-HRM에서는 채점자-문항별로 채점자 분산성 모수 ψ_{ir}을 추정하여 판별할 수 있었다는 점에서 E-HRM의 유용성을 보여준다. 이와 관련해서는 이 절의 끝 부분에서 [그림 Ⅴ-5]의 이상 평정과 실제 평정 간 관계를 중심으로 구체적으로 살펴볼 예정이다.

이와 같이 반응 수준에서 추정된 채점자 모수들이 어떻게 도출되었는지를 좀 더 직관적으로 살펴보기 위하여 HRM 분석 결과를 토대로 추정된 이상 평정 η에 대한 추정치 $\hat{\eta}$과 관찰된 실제 평정 Y의 분포를 [그림 Ⅴ-3]으로 도식화하여 제시하였다.



[그림 Ⅴ-3] 채점자별 이상 평정 추정치-실제 평정 분포(HRM): 아이슬란드

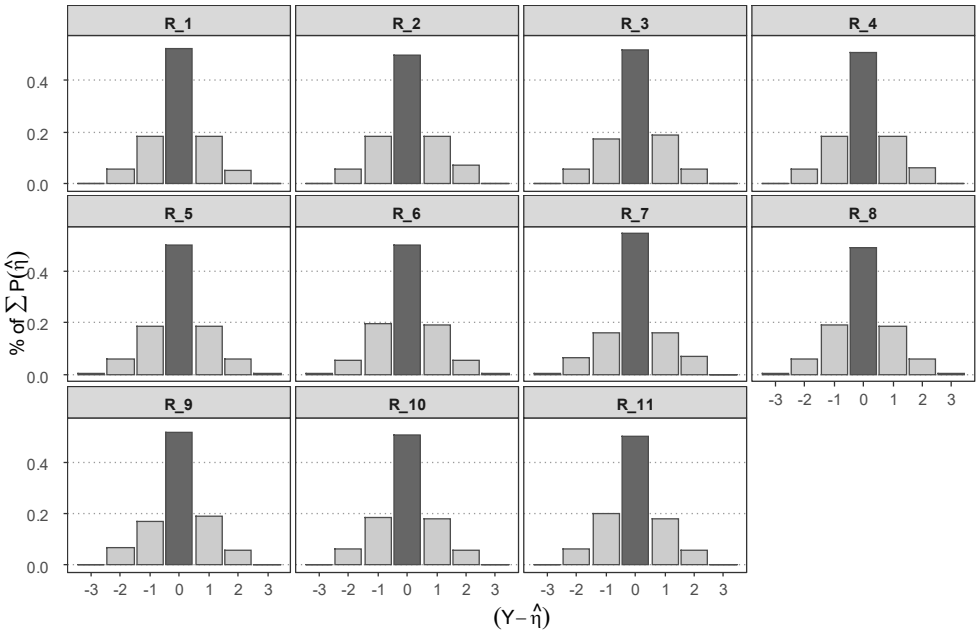
$\hat{\eta}$는 확률적인 잠재 범주 변수이므로 한 정수 값으로 확정할 수 없어 HRM을 통해 산출된 피험자 모수, 문항 모수 추정치를 문항 수준 모형인 GPCM에 투입하여 각각의 관찰된 실제 평정 Y에 대하여 이상 평정 추정치 $\hat{\eta}$가 각 평정 범주 k(범위:$0 \sim K-1$, K: 범주수)일 확률 $P(\hat{\eta}=k|Y)$를 구하였다. 그리고 이 확률을 전체 채점 단위 및 채점자 단위에서 $Y \times \hat{\eta}$ 행렬의 원소(element)별로 합산하여 이상 평정 추정치 $\hat{\eta}$과

관찰된 실제 평정 Y의 분포를 나타내었다.

[그림 Ⅴ-3]에서 막대의 높이는 $P(\hat{\eta}=k|Y)$의 합계로 엄밀하게는 관찰된 실제 평정 Y가 y일 때의 이상 평정 추정치 $\hat{\eta}$가 k로 관찰될 빈도의 기댓값이지만, HRM에서 설정된 관계에 따라서 이상 평정 추정치 $\hat{\eta}$가 k일 때의 실제 평정 Y가 y로 관찰될 빈도의 기댓값으로 바꾸어 해석할 수 있다. [그림 Ⅴ-3]에서 진하게 표시된 막대는 이상 평정 추정치 $\hat{\eta}$와 실제 관찰 평정 Y가 일치하는 경우로 채점자가 해당 채점에서 이상 평정 $\hat{\eta}$을 정확하게 탐지한 채점 빈도의 기댓값이고, 같은 행에서 좌우의 막대들은 이상 평정 $\hat{\eta}$에서 벗어나는 채점을 할 기댓값들이다.

[그림 Ⅴ-3]의 가장 상단 왼쪽의 그래프는 HRM으로 분석한 아이슬란드의 모든 채점 자료에 대한 이상 평정 추정치 $\hat{\eta}$과 관찰된 실제 평정 Y의 분포를 나타내고, 나머지 그래프들은 11명의 채점자로 분포를 제시한 결과이다. 그래프를 살펴보면 대체로 이상 평정 추정치 $\hat{\eta}$와 실제 관찰 평정 Y가 일치하는 경우에 대한 기댓값이 가장 높고, 이를 중심으로 실제 관찰 평정이 정규 분포와 유사하게 퍼지는 것을 확인할 수 있다. 아이슬란드 전체 채점자들의 채점 분포와 개별 채점자들의 분포에 큰 차이가 없어 채점자 모수 추정에서 엄격성 모수가 $|\phi|<0.5$, 분산성 모수가 $\psi<0.5$ 기준을 충족하면서 안정적으로 추정된 분석 결과가 나타났다.

한편, [그림 Ⅴ-3]에서는 문항의 평정 범주수가 3개(0~2점)인 문항과 4개(0~3점)인 문항이 혼재되어 있어서 분포를 다소 왜곡시키고 이상 평정 추정치가 $\hat{\eta}$가 3인 경우의 실제 평정이 어떻게 분포하는지 식별하기가 어려운 문제가 있다. 그리고 HRM의 반응 수준 모형은 $(Y-\hat{\eta})$에 ϕ로 인한 이동이 반영된 후 평정이 흩어지는 정도로 ψ가 추정되므로, $P(\hat{\eta}=k|Y)$의 합계를 이상 평정 추정치와 실제 평정 간의 편차값인 $(Y-\hat{\eta})$에 따라 다시 정리하여 채점자별로 나타내면 [그림 Ⅴ-4]와 같다.



[그림 Ⅴ-4] 채점자별 이상 평정 추정치-실제 평정 편차 분포(HRM): 아이슬란드

그래프의 척도는 $P(\hat{\eta})$의 합계로 표현할 경우, 채점자별로 채점 대상 인원과 답안수가 다르기 때문에 채점자별 분포의 비교가 어려운 점을 감안하여 채점자 단위에서 편차 점수별 $P(\hat{\eta})$의 합계를 비율로 변환하여 제시하였다. 다만, 주의하여야할 점은 HRM의 반응수준 모형은 $P(\hat{\eta}=k|Y=y)$가 아니라 특정 이상 평정일 때의 특정 관찰된 특정 실제 평정들이 나타날 조건부 확률인 $P(Y=y|\hat{\eta}=\eta)$로 추정된다는 점과 그래프에서와 같은 확률의 합이 아니라 확률의 곱인 우도(likelihood)를 활용하여 MCMC 추정이 이루어진다는 점이다. 따라서 [그림 Ⅴ-4]의 '이상 평정 추정치-실제 평정 편차 분포'는 관찰된 실제 평정이 이상 평정에서 이탈하는 정도를 직관적으로 살펴보기 위한 제한적인 목적을 가지고 있다는 것을 주의하여야 한다.

[그림 Ⅴ-4]에서 채점자별 이상 평정 추정치와 실제 평정 간의 편차 분포는 [그림 Ⅴ-3]과 마찬가지로 채점자 간에 눈에 띄는 큰 차이를 보

이지 않았지만 분석된 채점자 모수와의 관련성은 확인할 수 있었다. 먼저, 엄격성 모수와 관련해서는 엄격성 모수가 가장 낮아 엄격한 것으로 나타난 11번 채점자(ϕ_{11}=-.300)의 편차 점수 분포에서 0을 기준으로 대칭적인 분포에서 벗어나 이상 채점보다 실제 채점 점수가 낮은 음수 편차값인 -2점, -1점의 비율이 2점과 1점의 비율보다 좀 더 높은 편포를 보였고, 반대로 엄격성에 있어 가장 관대한 채점자로 분석된 9번 채점자(ϕ_9=.062)의 편차 점수 분포는 0보다 오른쪽으로 쏠려있는 부적 편포임을 확인할 수 있었다.

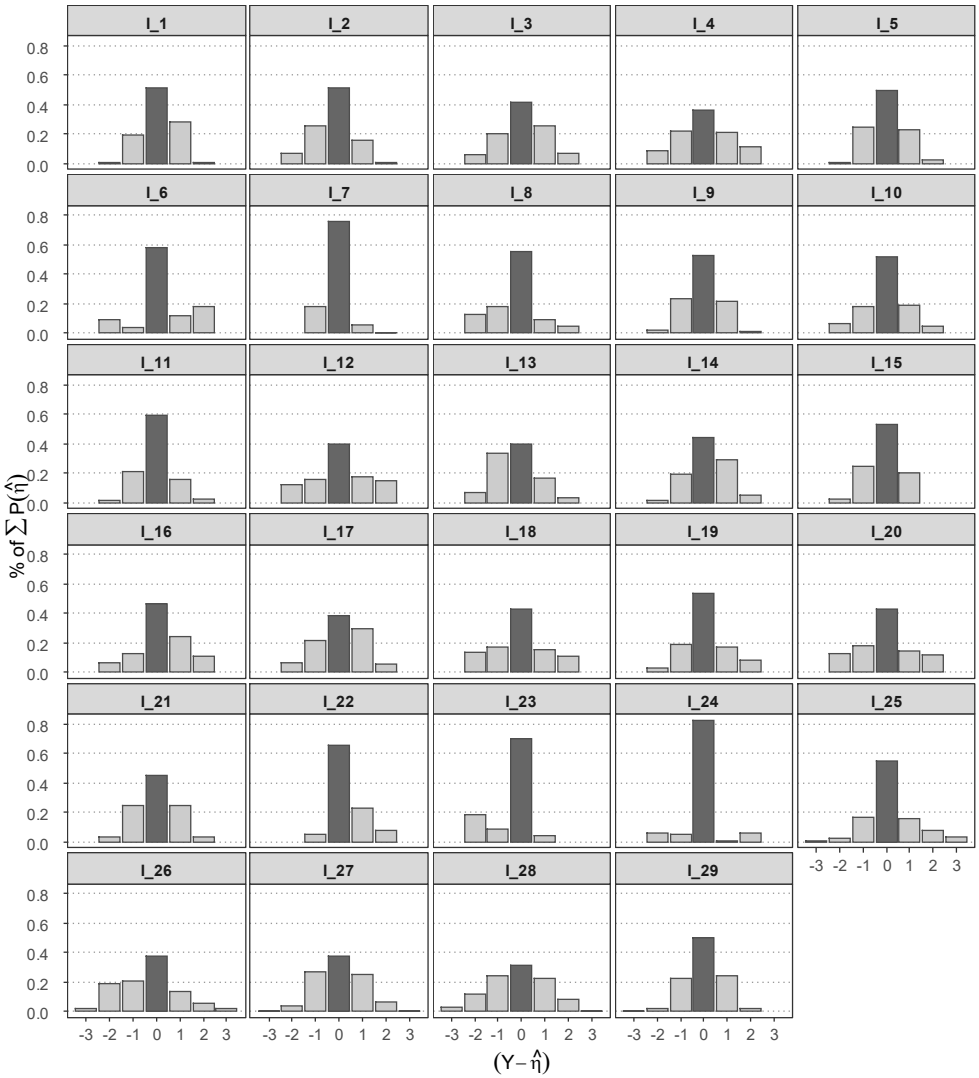
채점자 분산성 모수와 관련해서는 편차 점수 분포에 큰 차이가 나타나지 않아 ψ_r가 $\psi < 0.5$인 범위를 가지는 유사한 값으로 추정된 결과가 나타난 원인을 이해할 수 있다. 채점자별로는 분산성이 가장 낮았던 1번 채점자(ψ_1=.055)가 도표 상으로는 큰 차이를 알아보기 어렵지만 편차 점수의 극단 혹은 극단에 인접한 - 3점, -2점, 2점, 3점에서 다른 채점자들에 비하여 수치상 낮은 비율을 보였다. 그리고 분산성 모수가 가장 높은 11번 채점자(ψ_{11}=.240)의 경우에는 편차 점수가 0점인 비율이 전체 채점자 중에서 가장 낮은 49.6%로 분산성이 높게 추정된 결과로 이어졌다.

이상과 같이 HRM에 대한 분석 결과에서는 채점자 분산성 모수에 큰 차이가 없는 것으로 나타났지만, <표 Ⅴ-3>과 [그림 Ⅴ-3]에서 제시된 바와 같이 E-HRM 분석 결과에서는 채점자 내 분산성 모수들 간에 다소 차이가 있는 것으로 분석되었다. 그리고 모형 적합도 지수 비교에서도 E-HRM가 HRM보다 자료에 좀 더 적합한 것으로 나타나 HRM에서 전체 문항에서는 일관된 것으로 평가할 수 있었던 채점자 분산성 모수가 문항에 따라 차이가 존재함을 어느 정도 뒷받침한다고 하겠다. 좀 더 구체적으로 문항 간 채점자 모수의 차이를 살펴보기 위해서 채점자별 이상 평정 추정치-실제 평정 편차 분포를 살펴본 방식으로 채점자 내에서 문항 간의 편차 분포를 검토해볼 필요가 있겠다. 채점자-문항별로 아이슬란드의 모든 분산성 모수(ψ_{ir})와 관련된 이상 평정 추정치-실제 평정 편

차 분포를 제시하려면 316개의 분포 그래프가 필요하므로 지면상의 문제로 E-HRM 분석 결과에서 $\psi < 0.5$ 기준을 충족하지 못하는 문항이 6개로 가장 많고, 채점자 내 ψ_{ir}의 범위가 넓은 6번 채점자를 기준으로 문항별 이상 평정 추정치-실제 평정 편차 분포를 [그림 Ⅴ-5]와 같이 제시하였다.

[그림 Ⅴ-5]의 결과들 중 문항별 이상 평정 추정치와 실제 평정 간의 편차 분포에서 채점자 분산성 모수 ψ_{ir}이 .5를 초과한 문항들 위주로 살펴보면, 분산성 모수가 가장 크게 추정된 6번 문항($\psi_{6,6}$=.845)은 편차 점수가 0점인 비율이 낮고, 양극단인 -2점과 2점의 비율을 합하여 30%에 가까운 수치를 보여 분산성 모수가 높게 나타난 것을 이해할 수 있다. 두 번째로 분산성 모수가 높았던 18번 문항($\psi_{18,6}$=.736) 역시 편차 점수가 -2점, 2점인 비율이 높았고, 2번 문항($\psi_{2,6}$=.646)의 경우 편차 점수가 -2점인 비율이 높아 분산성 모수가 크게 추정된 것으로 판단된다.

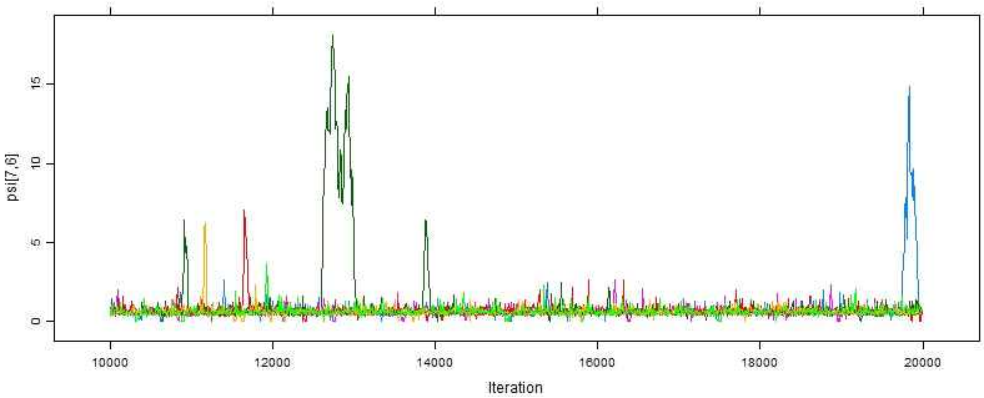
그리고 채점 범주수가 3개인 1~24번 문항들과는 달리 4개 범주를 가지는 25~29번 문항은 3개 범주 문항과 비교가 어렵기 때문에 5개 문항 간에만 비교할 수 있겠는데, 상대적으로 분산성 모수가 컸던 27번 문항($\psi_{27,6}$=.524), 28번 문항($\psi_{28,6}$=.557)은 나머지 문항들과 비교하였을 때, 분포 상에서 상대적으로 첨도가 낮은 납작한 분포 형태를 보여 분산성 모수가 높게 추정된 것을 확인할 수 있다.



[그림 Ⅴ-5] 문항별 이상 평정 추정치-실제 평정 편차 분포(E-HRM): 아이슬란드 6번 채점자

한편 7번 문항($\psi_{7,6}$=.594)의 경우 편차 점수가 0점인 비율이 높은데도 불구하고 분산성 모수가 높게 추정된 부분은 편차 점수 분포로 설명이 되지 않았고, 모수 추정치의 표준오차가 1.058로 매우 높아 MCMC 연쇄를 확인해본 결과, $\hat{R}$=1.018로 수렴 통계량은 “<1.1”을 충족하였으나 [그림 Ⅴ-6]과 같이 6개 체인들에서 일부 구간이 극단적으로 크게 추정되어 충분히 안정적인 수렴에 이르지 못한 모수 추정치라고 볼 수 있다. 만약

burn-in 표본수를 더 늘리고, 체인 길이도 늘린다면 충분히 안정적인 수렴이 예상되고, 분산성 모수 역시 .5 이하로 떨어질 가능성이 높다고 판단된다.



[그림 Ⅴ-6] MCMC Trace plot: 아이슬란드(E-HRM) 6번 채점자, 7번 문항 분산성 모수

2. 벨기에(프랑스어권)

가. 모형 평가

벨기에(프랑스어권)에 대한 두 가지 연구 모형의 분석 결과, 채점자별로만 반응 수준에서 채점자 분산성(ϕ_r)을 추정하는 HRM은 모형이 적절하게 수렴되지 않았고, 채점자-문항별로 채점자 분산성(ψ_{ir})를 추정하는 E-HRM만이 수렴되었다. E-HRM의 적합도 지수는 Deviance 1988.8, DIC 2038.7로 HRM의 적합도 지수인 Deviance 1742.8, DIC 2468.5에 비해 낮기 때문에 E-HRM가 HRM보다 자료를 더 잘 설명할 수 있는 모형이라고 볼 수 있겠다.

나. 채점 설계

벨기에(프랑스어권)는 6명의 채점자가 전체 29개 문항에 대한 피험자 825명의 응답 11,487건을 채점하였다. 채점에 있어서는 모든 채점자가 모든 문항을 채점하여 채점자-문항 설계에서는 완전구획설계(complete block design) 방식을 취하여 모든 모수가 동일한 척도 상에서 비교 가능한 채점 설계를 구성하였다. 채점자별 채점인원, 채점건수, 채점문항수는 다음 <표 Ⅴ-4>에서 확인할 수 있듯이 모든 채점자가 동일한 인원, 건수, 문항수를 채점하는 균형 설계는 아니었지만, 채점자별 채점건수가 최소 1,120건에서 최대 1,346건으로 채점자 간의 채점건수에 큰 차이가 나지는 않았다.

<표 Ⅴ-4> 채점자별 채점 배정: 벨기에(프랑스어권)

채점자	채점인원	채점건수	채점문항수
1	194	1,300	29
2	177	1,120	29
3	174	1,149	29
4	188	1,252	29
5	200	1,346	29
6	205	1,311	29
7	190	1,253	29
8	194	1,294	29

문항-채점자 배정 설계를 살펴보면, 배정 설계상의 전체 174개 셀(문항 29개 × 채점자 8명) 모두에 채점이 배정되어 문항-채점자 설계에 있어서는 완전구획설계(complete block design)이고, 자세한 문항-채점자별 채점인원은 <부록표 3>에서 확인할 수 있다. 채점자-피험자 가교 설계의 경우, [그림 Ⅲ-2]의 사회관계망 그래프 (b)에서 살펴본 바와 같이 채점자 간에 공통 피험자를 가지는 채점자간 설계를 통해서 동일한 척도 상에서 비교 가능하도록 채점 배정이 이루어졌다.

다. 문항 모수

벨기에(프랑스어권) 자료에 대한 E-HRM 분석을 통해 얻은 GPCM 문항 모수를 비록 수렴되지는 않았지만 HRM의 분석 결과로 얻은 GPCM 문항 모수와 함께 제시하면 다음 <표 Ⅴ-5>와 같다.

<표 Ⅴ-5> 문항 모수: 벨기에(프랑스어권)

문항	HRM				E-HRM (비수렴)			
	α	β_2	β_3	β_4	α	β_2	β_3	β_4
1	1.124	-2.868	1.367		1.080	-2.883	1.373	
2	.953	-3.587	-.398		.944	-3.595	-.381	
3	1.030	-2.071	.016		.971	-2.103	-.053	
4	.807	.140	-.133		.798	.118	-.104	
5	.868	-2.001	-.036		.815	-1.942	-.267	
6	.743	-.032	-.300		.754	.176	-.584	
7	1.126	-1.911	-1.893		1.096	-1.965	-1.891	
8	.875	.758	-.284		.815	.853	-.342	
9	1.083	-1.381	2.088		1.105	-1.345	1.979	
10	.778	1.661	-.086		.785	1.623	-.064	
11	.786	-1.269	1.964		.783	-1.153	1.800	
12	.988	-.614	-1.898		.990	-.607	-1.855	
13	.669	-1.640	-.099		.662	-1.547	-.116	
14	.902	-.201	.995		.845	-.071	.905	
15	1.362	-1.349	.790		1.302	-1.369	.732	
16	.922	-1.654	-1.155		.882	-1.725	-1.193	
17	.899	-1.072	.596		.923	-1.041	.656	
18	.947	1.172	.078		.988	1.030	.229	
19	.627	.276	1.794		.616	.305	1.808	
20	.772	.453	-1.239		.735	.498	-1.281	
21	.778	-1.701	-.644		.767	-1.629	-.758	
22	1.080	-1.615	-1.680		1.140	-1.715	-1.507	
23	.766	1.021	-2.193		.752	1.124	-2.293	
24	.615	.365	-1.397		.779	1.336	-2.412	
25	1.097	-.928	-1.332	-.465	1.060	-.918	-1.334	-.494
26	.814	-.704	.579	-.999	.796	-.248	.421	-1.074
27	.879	-2.603	1.722	.840	.843	-2.380	2.062	.459
28	.882	-1.849	-.695	.027	.864	-1.914	-.609	-.062
29	.918	-1.309	2.117	1.286	.942	-1.274	2.095	1.205

문항 모수 분석 결과, 두 모형 HRM과 E-HRM의 문항 모수들 간의 관계는 상관계수가 .998(α모수 간 상관계수 .985, β모수 간 .998)로 1에 가까운 매우 높은 상관을 보여 비록 HRM이 적절하게 수렴되지는 못했지만, 두 모형의 문항 수준에서 추정되는 모수의 경향성은 일치하는 것으로 볼 수 있겠다.

문항 곤란도를 문항 단위에서 살펴보면, 29개 문항 중 14개 문항(4, 6, 8, 10, 12, 18, 20, 22, 23, 24, 25, 26, 27, 29번 문항)에서 높은 점수(β_k)와 낮은 점수(β_{k-1})간의 모수가 역전되는 것으로 나타났다. 문항 곤란도 모수의 역전으로 인하여 중간 점수의 확률이 크게 낮아질 것으로 예상되고, 채점자 모수 추정에서 HRM에서 채점자 분산성 모수 ψ_r의 수렴이 되지 않은 데에 영향을 주었을 것이라고 추측할 수 있다.

라. 채점자 모수

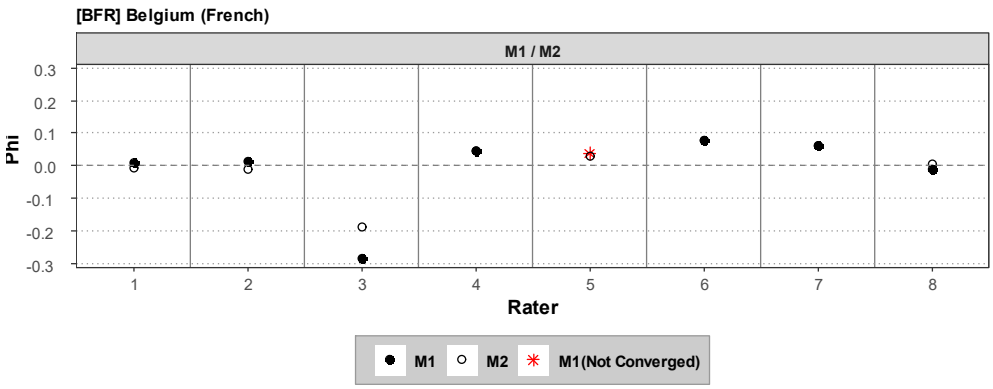
벨기에(프랑스어권)의 채점자 모수 분석 결과에서는 두 연구 모형 중 적합도가 좀 더 양호하면서 유일한 수렴 모형인 E-HRM을 제시하였고, 비교를 위해 수렴되지 않은 HRM의 채점자 모수를 <표 Ⅴ-6>과 같이 제시하였다.

<표 Ⅴ-6> 채점자 모수: 벨기에(프랑스어권)

모형	모수	채점자							
		1	2	3	4	5	6	7	8
HRM	ϕ_r	.010	.012	-.283	.044	.038	.079	.060	-.010
E-HRM	ϕ_r	-.007	-.010	-.188	.045	.029	.078	.063	.006
HRM	ψ_r	.357	.322	.248	.346	.338	.329	.356	.317
E-HRM	$\psi_{1,r}$	.367	.061	.067	.082	.063	.064	.063	.357
	$\psi_{2,r}$	.339	.060	.065	.339	.235	.065	.063	.226
	$\psi_{3,r}$	.147	.061	.264	.382	.397	.259	.152	.085
	$\psi_{4,r}$	.242	.180	.147	.061	.301	.348	.327	.060

$\psi_{5,r}$	.061	.119	.201	.062	.060	.304	.333	.060
$\psi_{6,r}$	.117	.094	.175	.065	.330	.385	.315	.062
$\psi_{7,r}$	.326	.061	.066	.302	.086	.069	.358	.261
$\psi_{8,r}$	.395	.429	.095	.404	.062	.120	.396	.068
$\psi_{9,r}$	.059	.381	.071	.061	.060	.152	.062	.061
$\psi_{10,r}$	.062	.167	.371	.324	.062	.310	.093	.103
$\psi_{11,r}$	.062	.062	.068	.124	.062	.109	.328	.373
$\psi_{12,r}$	.337	.063	.315	.064	.082	.066	.069	.061
$\psi_{13,r}$	.292	.133	.404	.075	.314	.292	.186	.330
$\psi_{14,r}$	.442	.327	.226	.365	.092	.112	.329	.089
$\psi_{15,r}$	.427	.330	.119	.168	.104	.147	.450	.404
$\psi_{16,r}$	.257	.349	.070	.063	.350	.127	.373	.062
$\psi_{17,r}$	.092	.118	.193	.387	.411	.309	.302	.349
$\psi_{18,r}$	.358	.497	.321	.067	.092	.109	.157	.064
$\psi_{19,r}$	.063	.060	.078	.063	.060	.062	.063	.060
$\psi_{20,r}$	.081	.439	.443	.076	.355	.160	.067	.358
$\psi_{21,r}$	.096	.357	.070	.353	.450	.190	.063	.454
$\psi_{22,r}$	.343	.062	.243	.066	.344	.063	.299	.062
$\psi_{23,r}$	.060	.113	.306	.064	.062	.436	.350	.064
$\psi_{24,r}$	.472	.107	.336	.373	.099	.405	.416	.112
$\psi_{25,r}$	.067	.062	.253	.061	.368	.346	.180	.059
$\psi_{26,r}$	.397	.357	.383	.312	.138	.155	.338	.142
$\psi_{27,r}$	.075	.060	.313	.419	.440	.063	.111	.116
$\psi_{28,r}$	.154	.296	.269	.313	.315	.229	.148	.270
$\psi_{29,r}$	.095	.322	.392	.427	.405	.156	.079	.088

먼저 HRM, E-HRM 공통적으로 채점자별로 차이가 있고, 문항 간에는 차이가 없는 모수로 추정된 채점자 엄격성 모수 ϕ를 살펴보기 위하여 그래프로 제시하면, [그림 Ⅴ-7]과 같다.



[그림 Ⅴ-7] 채점자 엄격성 모수: 벨기에(프랑스어권)

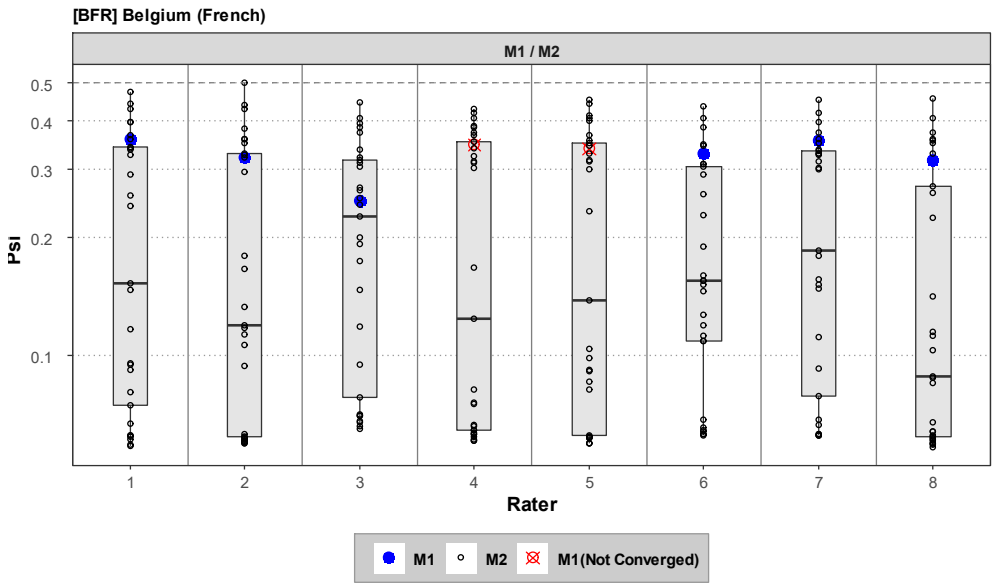
채점자 엄격성 모수 추정 결과를 살펴보면, 수렴되지 않은 모형인 HRM(M1)에서는 채점자 분산성 모수가 수렴되지 않은 채점자 중 5번에서 엄격성 모수가 수렴되지 않은 것으로 나타났다. HRM의 채점자 엄격성 모수의 범위는 [-.283, .079]로 [-.5, .5] 범위 이내에 위치하기 때문에 심각한 평정의 편의는 없었지만, 3번 채점자의 HRM 엄격성 모수 ϕ_3가 -.283으로 절댓값이 .1이하로 0에 근사하여 엄격성으로 인한 편의가 거의 없는 나머지 채점자들에 비해 상대적으로 엄격한 채점을 한 것으로 보인다. 3번 채점자의 E-HRM(M2) 엄격성 모수 ϕ_3 역시 -.188로 다른 채점자들에 비하여 엄격한 편이었는데, 문항별로 추정되는 ψ_{ir}에 따라서 문항별로 엄격성에 의한 영향이 차이가 있음을 엄격성 모수 ϕ_3이 문항별 분산성 모수인 $\psi_{7,3}$, $\psi_{20,3}$과 반응수준 모형에서 결합되어 발생하는 편의를 보여주는 <표 Ⅴ-7>을 통해 확인할 수 있다.

<표 Ⅴ-7> 채점자 모수에 따른 이상평정-실제평정의 편의:
벨기에(프랑스어권) 3번 채점자 7번, 20번 문항

문항		20번			7번		
ϕ		-.188			-.188		
ψ		.443			.065		
이상평정-실제평정 분포 (η내 비율, %)		실제평정 Y			실제평정 Y		
		0	1	2	0	1	2
이상평정η	0	97.1	2.9	.0	100.0	.0	.0
	1	16.6	81.0	2.4	.0	100.0	.0
	2	.0	17.0	83.0	.0	.0	100.0

동일한 피험자의 엄격성 모수였음에도 불구하고 문항에 따라서 분산성 모수가 달라짐에 따라 분산성 모수가 작았던 7번 문항에서는 이상 평정으로부터 편의가 발생할 확률이 낮았던 반면, 20번 문항에서는 이상 평정이 1점일 때, 16.6% 확률로 실제 평정에서 0점을 2.4%확률로 2점으로 평정하게 되는 차이를 보인다.

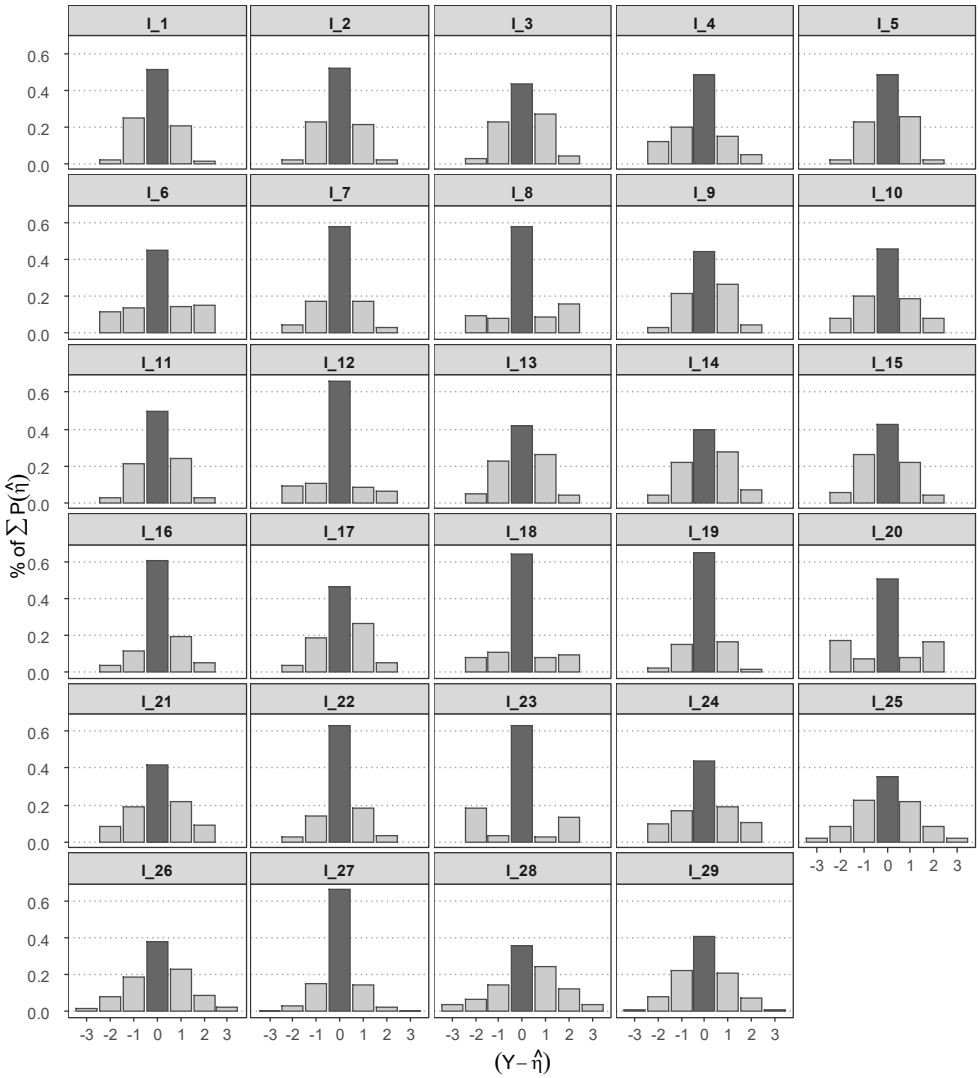
채점자-문항별로 분산성 모수 ψ_{ir}를 추정한 E-HRM(M2)과 수렴되지 않은 모형인 HRM의 분산성 모수 ψ_r을 그래프로 비교 제시하면 [그림 Ⅴ-8]과 같다.



[그림 Ⅴ-8] 채점자 분산성 모수: 벨기에(프랑스어권)

[그림 Ⅴ-8]에서 HRM의 채점자 분산성 모수를 살펴보면 최솟값은 3번 채점자의 .248로 추정되었고, 최댓값은 1번 채점자로 분산성 모수 ϕ가 .357로 추정되어 모든 채점자들의 분산성 모수가 일관성 기준인 '$\psi < 0.5$'를 충족하였다. 그리고 E-HRM의 채점자 분산성 모수 역시 [.59, .497]의 범위로 '$\psi < 0.5$' 기준 이내에 포함되어 있었다. 따라서 HRM의 4, 5번 채점자의 분산성 모수가 수렴되지 않았다는 한계가 있지만, HRM과 E-HRM에서의 분산성 모수 추정 결과를 통해 채점자들이 평정에 있어서 이상 채점으로부터 크게 벗어나지 않는 안정적인 채점을 하고 있는 것으로 해석할 수 있겠다.

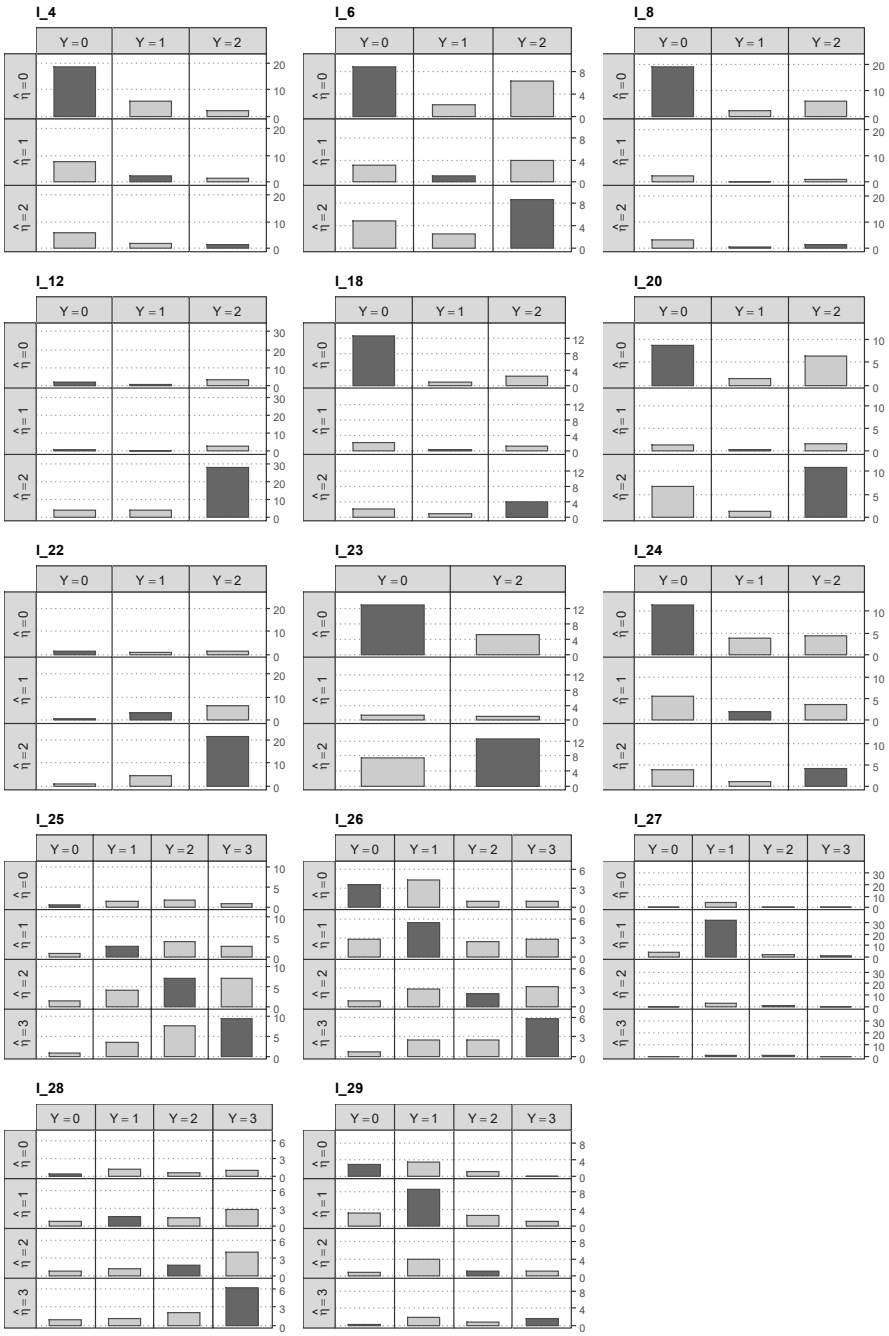
채점자 분산성 모수와 관련하여 HRM에서 채점자 4번과 5번의 분산성 모수가 수렴되지 않은 원인에 대해서 살펴보기 위하여 채점자 4번의 문항별 이상 평정 추정치-실제 평정 편차 분포를 도식화하여 제시하면 [그림 Ⅴ-9]와 같다.



[그림 Ⅴ-9] 문항별 이상 평정 추정치-실제 평정 편차 분포(E-HRM): 벨기에(프랑스어권) 4번 채점자

4번 채점자의 문항별 이상 평정 추정치와 실제 평정 간의 편차 분포에서 눈에 띄는 점은 문항 6번, 8번, 18번, 20번, 23번에서 양 극단의 편차 점수 비율이 인접한 편차 점수의 비율보다 높다는 점이다. 이 문항들은 공통적으로 문항의 변별도 모수가 역전된 경우였다. 이에 대하여 더 구체적으로 확인하기 위하여 벨기에(프랑스어권)의 변별도 모수 추정치가 역전된 14개 문항들의 이상 평정 추정치-실제 평정 분포를 [그림 Ⅴ-

10]으로 제시하였다.



[그림 Ⅴ-10] 문항별 이상 평정 추정치-실제 평정 분포(E-HRM): 벨기에(프랑스어권) 4번 채점자

이상 평정 추정치와 실제 평정의 분포를 살펴보면, [그림 Ⅴ-9]에서 양 극단 편차 점수의 분포가 높았던 6번, 8번, 18번, 20번, 23번, 26번은 중간 점수의 분포가 낮은 경우임을 확인할 수 있었고, 심지어 23번 문항은 실제 평정에서 1점이 없었다. 이와 같이 중간 범주의 확률이 낮은 문항이 다수 혼재되어 있는 답안들을 채점하는 경우, HRM에서는 반응 수준 모형이 기대하는 분포와 차이가 있어 수렴이 어려워지고, 채점자 분산성 모수 추정에 영향을 주게 된다는 점은 HRM이 가진 약점이라고 볼 수 있다. 한편으로 HRM에서 수렴하지 못했던 채점자의 분산성을 문항 단위에서 문제없이 추정해 낼 수 있다는 점에서 E-HRM의 유용성을 보여준다고 하겠다.

3. 덴마크

가. 모형 평가

덴마크 자료에 대해 두 가지 연구 모형을 적용하여 분석한 결과, 채점자별로만 채점자 분산성 ψ_r을 추정하는 HRM만이 적절하게 수렴되는 것으로 나타났다. 모형 적합도 지수에서는 Deviance 통계량은 HRM이 1001.4로 E-HRM의 875.4보다 높고, DIC에서는 HRM이 1121.7로 E-HRM의 1167.8보다 낮아 두 가지 모형 적합도 지수의 평가가 서로 상충되는 결과를 보였다. 이는 덴마크의 채점자수가 16명으로 전체 24개국 중 두 번째로 많기 때문에 채점자 모수가 많고, 이로 인하여 DIC에서 HRM에 비해 E-HRM의 모수수에 따른 패널티가 커진 결과이다. 따라서 단순 Deviance 통계량에서는 E-HRM이 HRM보다 자료에 더 적합하지만 모수수에 따른 패널티가 포함된 DIC 측면에서는 E-HRM이 모형의 간명성을 희생한 것에 비해 자료에 대해 더 잘 설명하고 있지는 못하다고 하겠다.

나. 채점 설계

덴마크의 자료는 16명의 채점자가 29개 문항에 대한 피험자 594명의 응답 8,079건을 채점한 자료이다. 채점에 있어서는 모든 채점자가 모든 문항을 채점하지는 않았고, 공통 채점자 및 공통 피험자를 통해 모수가 동일 척도에서 비교 가능한 설계를 갖추었다. 채점자별 채점인원, 채점건수, 채점문항수는 다음 <표 Ⅴ-8>에서 확인할 수 있듯이 모든 채점자가 서로 다른 인원, 건수, 문항수를 채점하는 불균형 설계 방식으로 채점 부담이 가장 큰 11번 채점자는 1,189건을 채점하여 226건으로 최소 채점건수인 8번 채점자의 5배 분량이다.

<표 Ⅴ-8> 채점자별 채점 배정: 덴마크

채점자	채점인원	채점건수	채점문항수
1	32	230	24
2	142	966	29
3	38	257	14
4	31	235	8
5	172	1,129	29
6	87	585	21
7	92	622	29
8	34	226	14
9	51	322	25
10	85	579	27
11	165	1,131	29
12	40	292	27
13	38	267	15
14	75	524	18
15	47	322	24
11	59	392	28

문항-채점자 배정 설계의 특성으로는 설계상의 총 464개 셀(채점자 16명×문항 29개) 중 361개 셀(77.8%)을 채점하는 불완전구획설계 방식을 취하였고, 구체적인 문항-채점자별 채점건수는 <부록표 4>에서 확인할 수 있다. 그런데 덴마크의 문항-채점자별 채점을 살펴보면, 문항-채점자에 대한 채점건수의 상당수가 10명 미만으로 적은 수였고, 심지어 5건 이하의 채점건수를 가지는 문항-채점자 설계 조건 셀이 81개로 전체 조건의 22.4%(1건 11.9%, 2건 4.2%, 3건 4.4%, 4건 0.8%, 5건1.1%)를 차지하였다. <부록표 4>를 문항당 채점건수를 중심으로 다시 요약 정리하면 <표 Ⅴ-9>와 같다.

<표 Ⅴ-9> 채점자 및 문항별 채점건수: 덴마크

채점자	채점 문항수	문항당 채점건수			5건 이하 채점 문항
		평균	최소	최대	
1	24	9.6	1	30	13
2	29	33.3	2	61	4
3	14	18.4	12	25	-
4	8	29.4	27	31	-
5	29	38.9	5	77	1
6	21	27.9	2	55	3
7	29	21.4	8	39	-
8	14	16.1	12	21	-
9	25	12.9	1	26	12
10	27	21.4	1	32	2
11	29	39.0	13	64	-
12	27	10.8	1	33	15
13	15	17.8	13	22	-
14	18	29.1	16	44	-
15	24	13.4	1	26	10
16	28	14.0	1	54	21

채점자 기준 문항-채점자별 채점건수를 살펴보면 4번 채점자와 같이

문항 간 편차가 거의 없이 30개 정도의 문항을 채점한 경우도 있었지만, 1번 채점자는 문항당 평균 채점건수가 9.6건으로 매우 작고 5건 이하 채점 문항도 13개 문항으로 1번 채점자가 채점하는 문항의 절반을 넘었다. 그리고 16번 채점자의 경우, 채점하는 28개 문항 중 75%인 21개 문항이 5건 이하의 채점건수를 포함하고 있다. 이러한 채점자-문항당 채점건수가 작은 조건은 채점자-문항 단위에서 분산성 모수 ψ_{ir}을 추정하는 연구모형인 E-HRM의 추정을 어렵게 하는 요인이 되고, 이러한 조건에서 E-HRM가 작동할 수 있는 가능성이 있는지를 모의실험 연구에 앞서 확인하고자 하는 것이 덴마크를 국가별 분석에 제시한 이유이기도 하다.

가교 측면에서는 문항-채점자 설계 조건 셀마다 채점 배정이 되는 설계 방식은 아니었지만, 가교 문항과 가교 채점자로 가교가 구성되어 있고, 채점자 간에도 [그림 Ⅲ-2]의 사회관계망 그래프 (c)에서 살펴보았듯이 채점자 간에 공통 피험자를 통해 가교가 이루어져 동일한 척도 상에서 비교 가능하도록 채점 배정 방식이다.

다. 문항 모수

덴마크 채점 자료에서의 수렴된 모형인 HRM의 문항모수는 다음 <표 Ⅴ-10>과 같고, 채점자 모수 추정이 적절하게 수렴이 되지는 못했지만 E-HRM의 문항모수도 비교를 위하여 함께 제시하였다.

<표 Ⅴ-10> 문항 모수: 덴마크

문항	HRM				E-HRM			
	α	β_2	β_3	β_4	α	β_2	β_3	β_4
1	.978	-1.948	1.365		.980	-1.968	1.468	
2	.859	-2.541	.086		.862	-2.545	.089	
3	.686	-1.380	-.499		.712	-1.371	-.484	
4	.743	-1.105	.334		.740	-1.096	.355	
5	1.120	-2.498	.999		1.077	-2.499	.988	
6	.722	.251	-.135		.746	.384	-.193	

문항	HRM				E-HRM			
	α	β_2	β_3	β_4	α	β_2	β_3	β_4
7	1.051	-1.629	-1.882		1.066	-1.624	-1.844	
8	.681	.030	-.248		.672	.050	-.210	
9	1.047	-.982	1.352		1.037	-.973	1.362	
10	1.118	-.330	.392		1.148	-.226	.329	
11	1.234	-1.097	2.046		1.222	-1.089	2.073	
12	1.013	-.728	-1.576		1.005	-.727	-1.585	
13	.879	-1.603	.089		.868	-1.623	.102	
14	.762	-.202	1.199		.772	-.198	1.173	
15	.958	-1.431	.152		.972	-1.432	.188	
16	.969	-.688	-.964		.970	-.709	-.949	
17	1.316	-.749	.627		1.312	-.749	.630	
18	.733	-.269	.800		.722	-.282	.848	
19	.980	.018	1.143		.976	.040	1.130	
20	.707	.032	-2.385		.705	.029	-2.386	
21	.757	-.615	.331		.762	-.610	.321	
22	.963	-1.252	-1.167		.950	-1.279	-1.173	
23	1.127	.143	-1.716		1.170	.185	-1.714	
24	.948	.592	.234		.921	.653	.190	
25	.542	-1.498	-.297	.069	.535	-1.409	-.358	.101
26	.771	-2.213	-.286	-.190	.779	-2.214	-.275	-.203
27	1.131	-2.578	1.751	1.200	1.137	-2.573	1.753	1.202
28	.885	-1.880	-.931	.100	.886	-1.877	-.917	.098
29	.950	-1.494	.864	.196	.930	-1.512	.883	.147

문항 모수 분석 결과, 두 모형 HRM과 E-HRM의 문항 모수들 간의 관계는 상관계수가 전체 모수에서 .999, α모수에서 .962, β모수에서 .987로 매우 높은 상관을 보여 MCMC 연쇄의 수렴 여부를 떠나 두 모형의 문항 수준에서 추정되는 모수의 경향성은 일치하는 것으로 볼 수 있겠다.

그리고 문항 곤란도를 문항 단위에서 살펴보면, 전체 29개 문항들 중 10개 문항들(6, 7, 8, 12, 16, 20, 23, 24, 27, 29번 문항)에서 높은 점수의 곤란도 모수 β_k와 낮은 점수의 곤란도 모수 β_{k-1}간의 모수가 역전되는 결과를 나타냈다. 앞서 2절에서 문항모수 역전이 14개 문항에서 발생하였던 벨기에(프랑스어권)에 대한 문항 모수에 비해서는 양호하지만, 채점

자 모수에 영향을 줄 가능성이 있으므로 채점자 모수 분석에서 이 문항들에 주의할 필요가 있다.

다. 채점자 모수

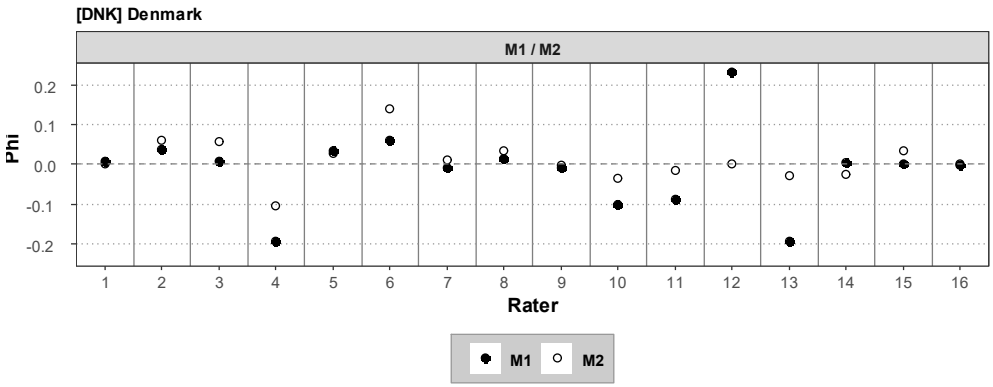
채점자 모수 분석 결과, 자료에 대한 최적모형인 HRM과 채점자 수준 모수에서 E-HRM에 내재되는 모형인 HRM의 채점자 모수를 <표 V-11>과 같이 제시하였다.

<표 V-11> 채점자 모수: 덴마크

모형	모수	채점자											
		1	2	3	4	5	6	7	8	9	10	11	12
HRM	ϕ_r	.017	.047	.045	.015	.032	-.015	-.050	-.002	.062	-.018	-.300	
E-HRM	ϕ_r	.007	.038	.009	-.193	.035	.060	-.010	.013	-.009	-.102	-.088	.232
HRM	ψ_r	.001	.060	.058	-.103	.029	.140	.010	.034	-.004	-.036	-.015	-.001
E-HRM	$\psi_{1,r}$	.326	.328	.054	.262	.306	.066	.344	.064	.372	.324	.295	.222
	$\psi_{2,r}$	.067	.068			.084	.352	.066	.070	.165	.094	.069	.278
	$\psi_{3,r}$	.069	.070			.271	.066	.068	.070	.187	.065	.356	.216
	$\psi_{4,r}$	.067	.070			.071	.068	.069	.072	.189	.062	.068	.206
	$\psi_{5,r}$		.336			.058	.279	.374	.129	.066	.216	.056	.148
	$\psi_{6,r}$		.377			.228	.145	.357	.096	.064	.318	.361	.111
	$\psi_{7,r}$	.225	.392	.096		.377	.353	.380	.078	.519	.083	.473	4.432
	$\psi_{8,r}$	.191	.067	.164		.092	.160	.297	.071	.068	.087	.355	.514
	$\psi_{9,r}$		.067	.084		.344	.189	.297	.070	.227	.081	.184	.085
	$\psi_{10,r}$	.064	.107			.062	.066	.295			.063	.063	.097
	$\psi_{11,r}$	.500	.129			.064	.370	.060			.090	.063	.110
	$\psi_{12,r}$	.377	.101			.062	.067	.363			.113	.063	.101
	$\psi_{13,r}$	.214	.067		.067	.080		.063	.075	.213	.403	.062	.065
	$\psi_{14,r}$	.318	.072		.071	.082		.068	.072	.246	.453	.429	.068
	$\psi_{15,r}$	.565	.064		.070	.065		.310	.072	.177	.067	.060	.230
	$\psi_{16,r}$	.247	.084		.173	.152		.064	.406	6.410	.371	.062	.068
	$\psi_{17,r}$	.238	.375	.070	.066	.059		.062		.208	.173	.059	.064
	$\psi_{18,r}$	.236	.061	.066	.338	.062		.066		.110		.070	.068
	$\psi_{19,r}$		.062	.067	.069	.061		.066		.127		.058	.066

$\psi_{20,r}$	.109	.065	.065		.062	.063	.104		.065	.063	.061	.211
$\psi_{21,r}$	.074	.069	.075		.093	.066	.069		.066	.065	.357	
$\psi_{22,r}$	.435	.066	.066		.063	.062	.384		.064	.062	.087	
$\psi_{23,r}$	.234	.068	.190		.065	.072	.076		.066	.067	.062	.075
$\psi_{24,r}$	.228	.066	.081		.062	.378	.077		.074	.068	.059	.073
$\psi_{25,r}$	.261	.125	.074		.099	.155	.075		.411	.589	.060	.075
$\psi_{26,r}$	.066	.399			.141	.267	.066	.167	.150	.244	.419	.177
$\psi_{27,r}$		.143			.058	.068	.394	.068	.478	.123	.059	.109
$\psi_{28,r}$	.063	.103			.062	.066	.062			.063	.060	.113
$\psi_{29,r}$	.191	.219	.066	.064	.267		.064		.104	.190	.263	.063

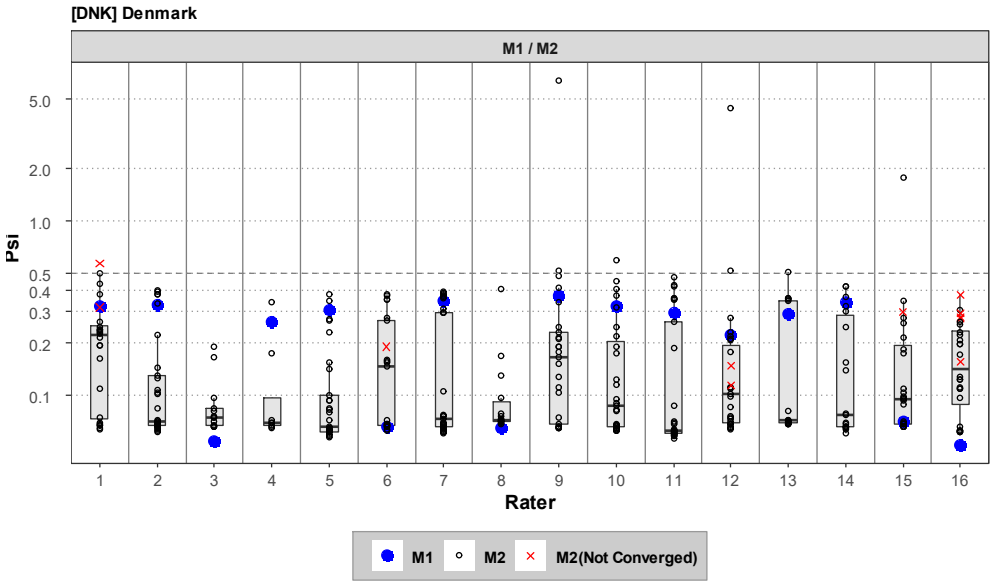
먼저 HRM, E-HRM의 채점자 엄격성 모수 ϕ 추정치를 그래프로 제시하면 [그림 Ⅴ-11]과 같다.



[그림 Ⅴ-11] 채점자 엄격성 모수: 덴마크

채점자 엄격성 모수 추정 결과, HRM의 경우 채점자 엄격성 모수는 [-.193, .232] 범위로 [-.5, .5] 범위 이내에 있었지만, 비수렴된 E-HRM의 엄격성 모수의 범위 [-.103, .140]에 비해서 넓은 범위였다. 엄격성 모수의 절댓값이 가장 크게 나타난 채점자 12번은 엄격성 모수ϕ_{12}가 .232로 상당히 관대한 채점으로 보이는데, 함께 추정된 ψ_{12}가 .222로 작은 편이기 때문에 이상 채점으로부터 실제 채점이 편향될 확률이 심각하게 높지는 않을 것으로 예상할 수 있다.

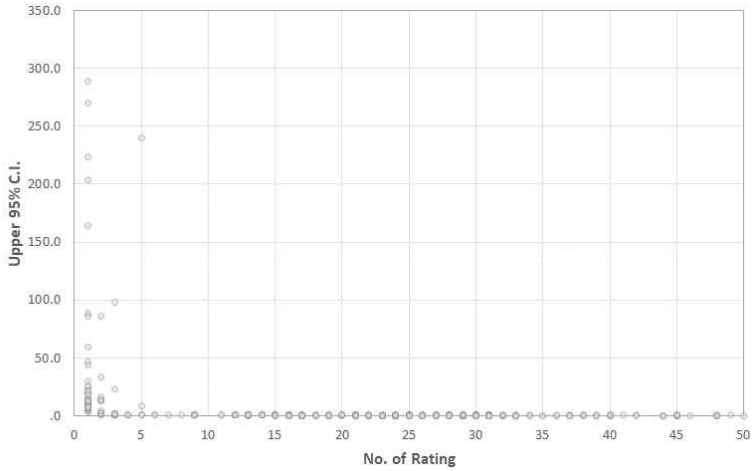
채점자별로만 분산성 모수 ψ_r을 추정한 HRM과 비수렴 모형이지만 HRM과의 비교를 위해 채점자-문항별로 분산성 모수 ψ_{ir}를 추정한 E-HRM의 결과를 그래프로 비교 제시하면 [그림 Ⅴ-12]와 같다.



[그림 Ⅴ-12] 채점자 분산성 모수: 덴마크

먼저, HRM에서는 채점자별 분산성 모수의 최댓값이 .372로 모든 채점자들이 일관성 있는 채점을 하고 있는 것으로 판단된다. 한편, E-HRM에서는 총 11개의 분산성 모수가 수렴이 되지 않은 것으로 분석되었다. 이 모수들은 모두 소수의 채점건수(1개 모수 3건, 3개 모수 2건, 7개 모수는 1건의 채점건수)만 배정된 문항-채점자 설계 셀에 해당하는 경우로 모수 추정에 활용된 채점건수가 적을 때, 대표성을 가지는 MCMC의 추정치의 산출이 어려움을 보여준다. 그리고 수렴 외에 MCMC의 정확도와 채점건수와의 관련성을 살펴보기 위하여 분산성 모수가 .5보다 큰 모수들을 확인한 결과, 일부 모수들에서 지나치게 신뢰구간이 넓어서 정확도가 떨어지는 것을 확인할 수 있었다. 수렴되지 않은 1번 채점자 14번 문항의 분산성 모수는 채점건수가 1건으로 95% 신뢰구

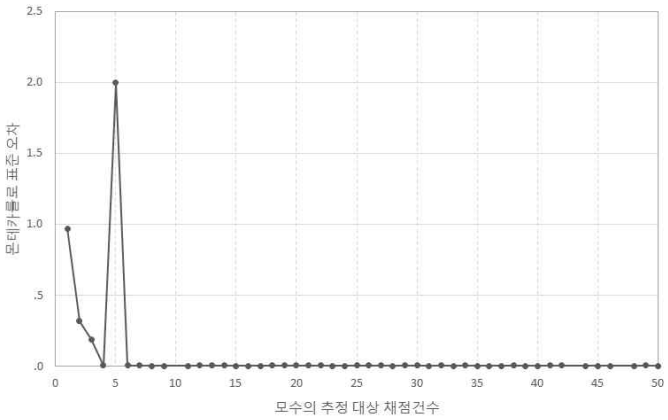
간이 [0.0, 269.7]로 나타났는데, 수렴된 것으로 평가된 분산성 모수들 중에서도 95% 신뢰구간이 9번 채점자 15번 문항(채점건수 1건) [0.0, 289.2], 12번 채점자 6번 문항(채점건수 5건) [0.0, 240.0], 12번 채점자 7번 문항(채점건수 5건) [0.0, 8.7], 15번 채점자 18번 문항(채점건수 3건) [0.0, 97.9] 등과 같이 과도하게 넓은 모수들은 정확도가 낮은 것을 볼 수 있다. 이를 채점건수별로 확인하기 위하여 채점건수와 95% 신뢰구간 상한값 분포를 [그림 Ⅴ-13]과 같이 제시하였다.



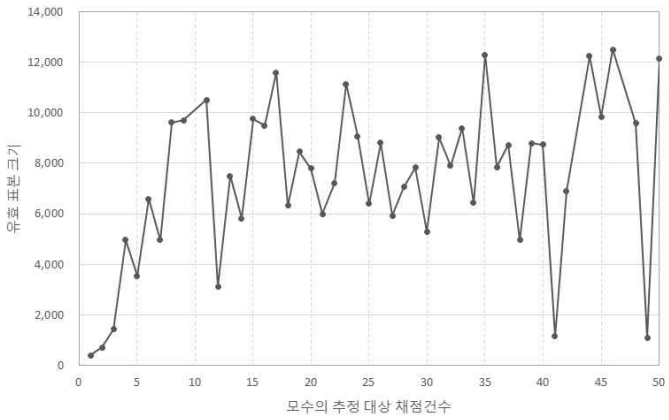
[그림 Ⅴ-13] 분산성 모수의 추정 대상 채점건수 - 95% 신뢰구간 상한값 분포: 덴마크(E-HRM)

범위로 나타나는 채점자 분산성 모수의 95% 신뢰구간을 도표로 제시하기는 어렵고, ψ가 로그 정규 분포를 따르기 때문에 95% 신뢰구간의 하한은 0보다는 크고 실제 값의 최대치는 .4 정도라는 점을 감안하여 채점건수에 따른 95% 신뢰구간의 상한값만 제시하였다. 채점건수에 따른 95% 신뢰구간 상한값 분포를 살펴보면 채점건수가 6건 이상인 경우에는 상한값이 큰 경우를 관찰하기 어려운데 반해서 채점건수가 5건 이하인 경우에는 MCMC 연쇄는 수렴이 되었음에도 불구하고 극단적으로 큰 상한값을 보여 적절하게 추정되지 않았다고 판단되는 분산성 모수들이 일

부 관찰된다. 그리고 채점건수별 몬테카를로 표준오차(Monte Carlo standard error; 이하 MCSE)의 평균을 나타낸 [그림 Ⅴ-14]에서도 95% 신뢰구간 상한값 분석과 유사하게 5건 이하에서부터 오차가 커서 정확도가 낮은 모수가 나타나는 것을 확인할 수 있다.



[그림 Ⅴ-14] 분산성 모수의 추정 대상 채점건수별 MCSE 평균: 덴마크(E-HRM)

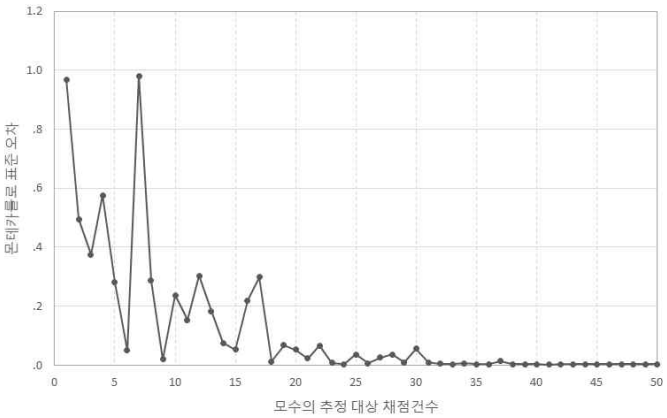


[그림 Ⅴ-15] 분산성 모수의 추정 대상 채점건수별 ESS 평균: 덴마크(E-HRM)

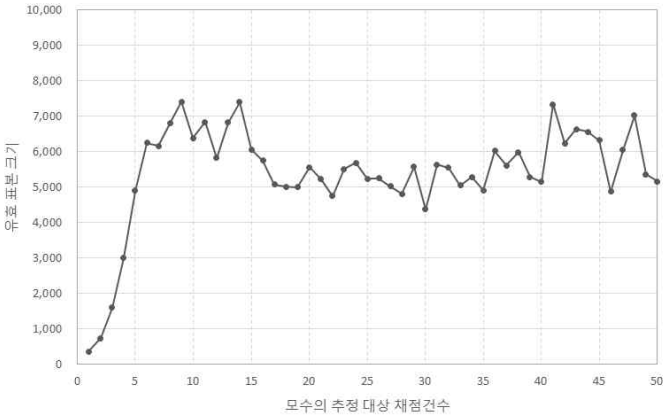
[그림 Ⅴ-15]는 유효 표본 크기(effective sample size; 이하 ESS)로 MCMC 추정의 실제 표본 크기를 자기상관량으로 나눈 값(Kass, Carlin,

Gelman, & Neal, 1998)이다. ESS가 클수록 정확하고 안정적인 사후분포임을 의미하는데, 이에 대한 절대적인 기준은 존재하지 않지만 Kruschke(2014)는 응용 사례에서의 경험에 기초하여 엄밀한 기준으로써 ESS가 10,000을 넘는 것을 추천하고, 경우에 따라서 더 낮은 기준을 활용하는 것을 제안하였다. 이에 따르면 10,000에 근접할 수 있는 채점건수 8~10건 정도가 ESS 관점에서 일정 수준 이상의 정확도를 가질 수 있는 조건일 가능성이 있다.

한편, 국가별로 모수수가 달라 모형의 복잡도가 다르고, 문항, 채점자, 피험자 특성과 채점 설계가 달라 동일한 채점건수 기준을 정하기 어렵지만 덴마크 사례에서 전체 국가로 확장하여 MCMC의 정확도 측면에서 연구 모형인 E-HRM를 적용할 수 있는 문항-채점자별 채점건수 조건을 탐색해보고자 한다. 덴마크 채점자들의 분산성 모수에 대한 사례수별 정확성 분석과 마찬가지 방식으로 전체 국가 자료의 분산성 모수에 대한 사례수별 MCSE, ESS로 나타내면 [그림 Ⅴ-16], [그림 Ⅴ-17]과 같다.



[그림 Ⅴ-16] 분산성 모수의 추정 대상 채점건수별 MCSE 평균: 전체 국가(E-HRM)



[그림 Ⅴ-17] 분산성 모수의 추정 대상 채점건수별 ESS 평균: 전체 국가(E-HRM)

MCSE에서는 채점건수가 30건 이상인 경우, 거의 0에 수렴하는 것을 확인할 수 있었고, 채점자 엄격성 척도의 크기를 감안한다면 MCSE 평균이 0.1을 넘지 않는 17~18건 정도가 E-HRM를 적용하기 위한 모수별 최소 채점건수가 될 가능성이 있다. 그리고 ESS 측면에서는 ESS의 평균이 제시된 구간(50건을 초과한 구간 역시 비슷한 경향이기 때문에 제시하지 않았음) 내에서 Kruschke(2014)가 제시한 10,000을 넘기지 못하고 있지만, 사례수가 충분히 커진 이후에도 5~7,000대의 평균 ESS를 보이고 있다는 점에서 문항-채점자별 채점건수가 5~10건 정도부터 혹은 보수적으로 10건부터 E-HRM를 적용하여도 MCMC의 정확성 확보가 가능할 것으로 추측해볼 수 있다. 물론 이러한 추측은 다음 장의 모의실험 연구를 통한 확인이 필요하고, 문항-채점자별 채점건수를 모의실험 연구 설계의 주요 요인으로 설정한 이유이기도 하다.

Ⅵ. 모의실험 연구 결과

1. 모의실험 채점자 모수

연구 방법에서 계획하였듯이 모의실험 연구에서는 적용 연구를 통해 도출된 채점자 모수를 활용하고자 하였다. 그러나 적용 연구 분석 결과, 연구 방법에서 제시하였던 네 가지 채점 양호도 유형 중 엄격성에 있어서 양호도가 낮은 '엄격'과 '엄격-저신뢰' 유형의 채점자 모수가 나타나지 않았다. 적용 연구 분석 결과, 엄격성 모수의 범위는 [-.355, .467]로 절댓값이 모두 .5 이하로 양호한 값을 가졌고, 분산성 모수가 .5 이하를 만족하는 78명의 채점자들에 대해서만 엄격성 모수를 살펴보면 그 범위가 [-.188, .190]으로 절댓값이 .2 이하로 매우 양호한 수치를 보였다. 때문에 '엄격' 유형의 채점자 모수 설정을 위하여 78명의 '정상' 채점자 중 20명을 무작위로 추출하고, 이들의 엄격성 모수를 균일분포 $U(-.6, -.4)$에서 무작위로 추출한 값으로 대체하여 '엄격' 조건의 채점자 집단을 생성하였다.

'저신뢰' 유형의 채점자는 채점의 대략 1/3이 이상 채점 η에서 벗어나게 되는 .7 이상의 분산성 모수 ψ_{ir}을 가지는 채점자들 중 10명을 무작위로 추출하여 '저신뢰' 채점자 집단을 생성하였다. 한편, 이론적 배경에서 살펴보았듯이 HRM에서는 분산성 모수가 1을 넘을 경우, 중간 범주 점수에서 이상 채점 η가 아닌 인접 범주로 채점될 확률이 54.8%에 이를 정도로 채점의 정확성이 낮아진다. 채점 기준이 세세하게 구성되고, 일정 수준 이상의 채점자 전문성과 추가적인 훈련이 이루어지는 대규모 표준화 검사에서 이런 정도의 심각하게 부정확한 채점은 실제 발생하기 어렵다고 판단되기에 앞서 Ⅳ, Ⅴ장의 적용 연구로부터 도출된 모수 중 1을 초과하는 분산성 모수 값들은 균일분포 $U(0.8, 1)$에서 무작위로 추출된 값으로 대체하였다. 마지막으로 '엄격-저신뢰' 유형의 채점자 집단은 '엄

격' 유형의 채점자 집단 생성 방식과 유사하게 '저신뢰' 집단의 엄격성 모수를 균일분포 $U(-.6, -.4)$에서 무작위로 추출한 값으로 대체하는 방식으로 생성하였다. '정상' 채점자 집단을 포함한 모의실험 연구에 활용된 모든 채점자 모수는 <부록표 2>에 제시하였고, '엄격', '저신뢰', '엄격-저신뢰' 집단의 채점자 모수 참값은 아래 <표 Ⅵ-1>과 같다.

<표 Ⅵ-1> 모의실험 채점자 모수 참값

채점자 특성	번호	ϕ_i	ψ_{ir}									
			문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
엄격	1	-.517	.062	.081	.380	.070	.067	.072	.070	.070	.149	.155
	2	-.475	.061	.067	.059	.059	.088	.062	.352	.362	.059	.151
	3	-.582	.129	.067	.123	.070	.461	.437	.457	.348	.069	.393
	4	-.445	.326	.126	.384	.072	.345	.097	.351	.410	.104	.064
	5	-.410	.059	.061	.073	.336	.341	.060	.059	.058	.058	.316
	6	-.576	.098	.313	.174	.057	.060	.056	.055	.362	.057	.356
	7	-.455	.063	.063	.152	.327	.333	.315	.358	.396	.062	.093
	8	-.542	.069	.356	.068	.056	.361	.473	.355	.184	.063	.063
	9	-.405	.063	.235	.397	.301	.060	.330	.086	.062	.060	.062
	10	-.403	.407	.443	.089	.373	.441	.432	.394	.069	.068	.060
	11	-.468	.371	.134	.261	.444	.353	.410	.061	.156	.345	.402
	12	-.518	.066	.082	.069	.070	.352	.377	.408	.158	.066	.091
	13	-.497	.359	.365	.412	.393	.344	.395	.372	.367	.342	.068
	14	-.451	.067	.065	.342	.363	.059	.085	.063	.201	.349	.064
	15	-.481	.066	.068	.069	.374	.357	.380	.297	.297	.295	.060
	16	-.418	.357	.226	.085	.060	.060	.062	.261	.068	.061	.103
	17	-.415	.069	.070	.115	.190	.420	.251	.387	.066	.068	.346
	18	-.594	.344	.398	.425	.060	.251	.336	.061	.168	.059	.098
	19	-.537	.333	.066	.404	.069	.092	.193	.188	.068	.110	.364
	20	-.504	.367	.339	.147	.242	.061	.117	.326	.395	.059	.062
저신뢰	1	.096	.073	.332	.072	.476	.298	.234	.384	.774	.400	.095
	2	.134	.072	.428	.356	.096	.082	.781	.456	.539	.141	.069

채점자 특성	번호	ϕ_i	ψ_{ir}									
			문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
	3	.007	.370	.663	.577	.494	.500	.814	.109	.550	.093	.369
	4	-.017	.406	.064	.088	.101	.140	.073	.072	.831	.067	.344
	5	.004	.065	.646	.065	.068	.450	.845	.594	.078	.138	.065
	6	.126	.070	.471	.092	.371	.533	.078	.078	.950	.068	.503
	7	.055	.063	.782	.495	.072	.436	.082	.101	.989	.506	.643
	8	-.178	.322	.089	.433	.074	.387	.731	.295	.958	.110	.499
	9	.119	.070	.152	.086	.110	.902	.303	.964	.106	.071	.111
	10	-.075	.116	.368	.304	.484	.422	.861	.570	.958	.250	.387
엄격-저신뢰	1	-.469	.073	.332	.072	.476	.298	.234	.384	.774	.400	.095
	2	-.543	.072	.428	.356	.096	.082	.781	.456	.539	.141	.069
	3	-.470	.370	.663	.577	.494	.500	.814	.109	.550	.093	.369
	4	-.531	.406	.064	.088	.101	.140	.073	.072	.831	.067	.344
	5	-.463	.065	.646	.065	.068	.450	.845	.594	.078	.138	.065
	6	-.444	.070	.471	.092	.371	.533	.078	.078	.950	.068	.503
	7	-.588	.063	.782	.495	.072	.436	.082	.101	.989	.506	.643
	8	-.550	.322	.089	.433	.074	.387	.731	.295	.958	.110	.499
	9	-.568	.070	.152	.086	.110	.902	.303	.964	.106	.071	.111
	10	-.462	.116	.368	.304	.484	.422	.861	.570	.958	.250	.387

※ 음영 표시된 값은 채점의 양호도가 낮은 모수

2. 모형 수렴

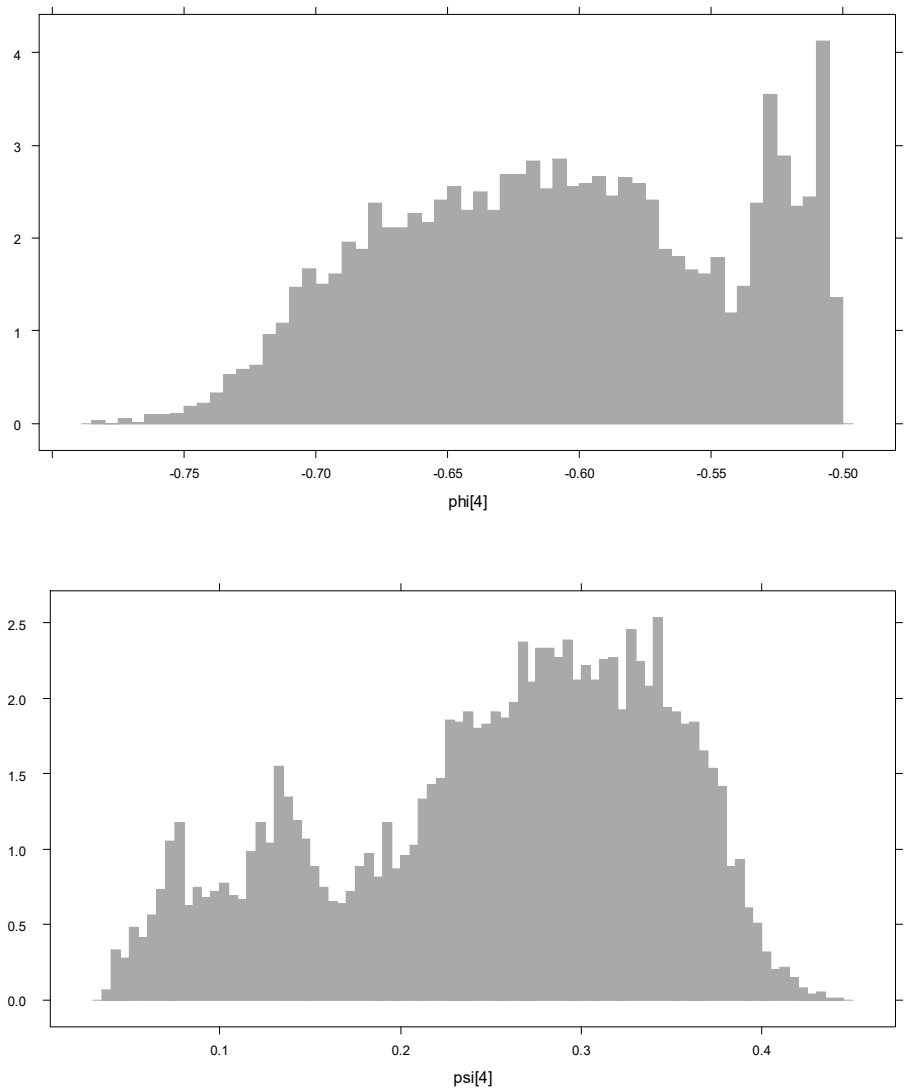
12가지 모의실험 설계 조건별로 E-HRM과 HRM을 적용한 50번씩의 반복 실험에서 정보적 사전분포를 활용하여 추정한 문항 모수 36,000개(변별도 모수 12,000개, 곤란도 모수 24,000개)의 MCMC 연쇄는 모두 적절하게 수렴한 것으로 나타났다. 반면에 무정보적 사전분포를 활용하여 추정한 채점자 모수의 경우, <표 Ⅵ-2>에서 제시된 바와 같이 대부분의 모의실험 조건에서 수렴하지 않는 경우가 발생하였다.

<표 Ⅵ-2> 모의실험 채점자 모수 수렴 비율

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	수렴 비율	모수 수	수렴 비율
E-HRM	100%	200	250	1.000	2,500	.996
		100	500	1.000	5,000	.996
		50	1,000	1.000	10,000	.999
		25	2,000	1.000	20,000	1.000
	70%	200	250	.992	2,500	.996
		100	500	1.000	5,000	.997
		50	1,000	1.000	10,000	.998
		25	2,000	1.000	20,000	.994
	40%	200	250	.996	2,500	.998
		100	500	1.000	5,000	.999
		50	1,000	.999	10,000	.997
		25	2,000	.999	20,000	.989
HRM	100%	200	250	.868	250	.816
		100	500	.906	500	.876
		50	1,000	.980	1,000	.950
		25	2,000	.998	2,000	.984
	70%	200	250	.824	250	.784
		100	500	.942	500	.904
		50	1,000	.976	1,000	.952
		25	2,000	.997	2,000	.984
	40%	200	250	.856	250	.812
		100	500	.946	500	.916
		50	1,000	.981	1,000	.964
		25	2,000	.999	2,000	.987

E-HRM은 엄격성 모수에서는 정상범위 채점자 비율이 '70%'인 조건에서 채점자-문항당 채점건수가 '200건'일 때 2개(0.8%), 정상범위 채점자 비율이 '70%'인 조건에서 채점자-문항당 채점건수가 200건, 50건, 25건 일 때 각각 2개, 1개, 1개의 모수가 적절하게 수렴하지 못한 결과를 보였다. 그리고 분산성 모수에서는 모든 조건에서 MCMC 연쇄가 수렴하지 않는 사례가 발생하였지만 대체로 99% 이상의 수렴 비율을 보여 MCMC 추정에 큰 문제는 없고, 연쇄의 길이를 현재의 10,000개보다 더 늘이면 대부분 적절히 수렴할 것으로 판단된다.

반면, HRM은 모든 정상범위 채점자 비율 조건에서 채점자-문항당 채점건수가 '200건'과 '100건'으로 많을 때, 95%에 못 미치는 다소 낮은 수렴 결과를 보였다. 채점자-문항당 채점건수가 200건일 경우 HRM에서는 채점자 모수 하나의 추정에 2,000건의 채점 자료가 활용된다는 점에서 채점자 모수 추정을 위한 채점건수는 충분한데도 불구하고, 오히려 채점자 엄격성 모수와 분산성 모수 모두 MCMC 연쇄의 수렴이 어려운 것으로 나타난 것이다. 이러한 결과는 모의자료 생성에 활용된 분산성 모수가 문항마다 달랐기 때문에 문항 간 채점자 분산성 모수 ψ_{ir}차이로 인하여 채점자 단위에서의 분산성 모수 ψ_r이 단봉분포로 수렴하지 못하고, 엄격성 모수 ϕ_r도 그 영향으로 수렴되지 못한 것으로 추측된다. 실제 '일부정상범위-채점건수 200건' 조건에서 비수렴 모수인 4번 채점자의 엄격성 모수와 분산성 모수의 사후분포를 확인한 결과, [그림 Ⅵ-1]과 같이 엄격성 모수와 분산성 모수가 단봉 분포가 아닌 다봉 분포(multimodal distribution) 형태를 띄고 있는 것을 확인할 수 있었다.



[그림 Ⅵ-1] HRM 채점자 모수의 MCMC 표본 분포
(정상범위 채점자 70%, 채점건수 200건 조건, 반복1, 채점자 4번)

[그림 Ⅵ-1]의 '정상범위 채점자 70%, 채점건수 200건 조건, 반복1, 채점자 4번'의 모의자료 생성에 활용된 문항별 채점자 분산성 모수 ψ_{ir}의 참값은 10개 문항에 대하여 모수 크기순으로 '.056, .063, .063, .068, .069, .184, .355, .356, .361, .473'으로 설정되어 있었다. 분산성 모수 중 .184를 제외한 나머지 모수들은 크게 두 가지(.056∼.069, .355∼.473) 다른 분산

성 모수 분포를 보이는데 채점건수까지 충분히 많았기 때문에 하나의 분산성 모수 ψ_r로는 이를 충분히 설명하지 못하고, 단봉분포 가정 하에 점추정치를 추정하는 상황에서 적절하게 수렴되지 않은 추정의 불안정성으로 나타난 것이라 판단된다.

3. 모수 복원력

가. 채점자 모수

1) 신뢰구간

채점자 모수에 대한 MCMC 추정의 사후분포가 얼마나 정확한지를 확인하기 위하여 모수 추정치 95% 신뢰구간 내 모수 참값이 포함되는 비율을 확인한 결과는 다음 <표 Ⅵ-3>과 같다. 이때 95% 신뢰구간은 전통적인 빈도주의 접근의 통계학에서의 평균을 기준으로 표준오차로 나타내는 신뢰구간이 아닌 MCMC 추정을 통해 획득한 사후분포의 2.5~97.5 백분위수를 의미하고 참값이 95% 확률로 포함된다고 해석할 수 있다.

<표 Ⅵ-3> 채점자 모수 추정값 95% 신뢰구간 내 모수 참값 포함 비율

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	포함비율	모수 수	포함비율
E-HRM	100%	200	250	.964	2,500	.953
		100	500	.970	5,000	.928
		50	1,000	.966	10,000	.898
		25	2,000	.983	20,000	.881
	70%	200	250	.956	2,500	.948
		100	500	.948	5,000	.924
		50	1,000	.959	10,000	.915
		25	2,000	.975	20,000	.894

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	포함비율	모수 수	포함비율
	40%	200	250	.944	2,500	.944
		100	500	.912	5,000	.927
		50	1,000	.958	10,000	.924
		25	2,000	.959	20,000	.922
HRM	100%	200	250	.836	250	-
		100	500	.914	500	-
		50	1,000	.897	1,000	-
		25	2,000	.922	2,000	-
	70%	200	250	.760	250	-
		100	500	.774	500	-
		50	1,000	.799	1,000	-
		25	2,000	.847	2,000	-
	40%	200	250	.564	250	-
		100	500	.634	500	-
		50	1,000	.728	1,000	-
		25	2,000	.793	2,000	-

12가지 모의실험 조건의 자료에 대한 E-HRM 분석 결과, 엄격성 모수는 '정상범위 채점자 40%, 채점자-문항당 채점건수 100건' 조건에서는 모수 추정치 95% 신뢰구간에 참값을 포함하는 비율이 .912로 .95에 못 미쳤지만 나머지 조건들에서는 모두 .95에 가깝거나 이를 상회하는 것으로 나타났다. 하지만 분산성 모수에 있어서는 채점자-문항당 채점건수가 '200건'인 조건들에서만 .95에 근사하는 수치들을 보였고, 채점자-문항당 채점건수가 적어질수록 95% 신뢰구간 내에 모수 참값을 포함하는 비율이 낮아져 채점건수가 '25건'인 조건들에서는 .9 이하로 다소 낮게 나타났다. 한편, HRM 분석 결과에서는 채점자 분산성 모수 참값은 문항 단위로 존재하기 때문에 HRM 채점자 분산성 모수 ψ_r에 대해서는 모수 추정값 95% 신뢰구간 내 모수 참값의 포함 여부를 평가할 수 없고 엄격성

모수의 신뢰구간에 대한 평가만 가능하다. 엄격성 모수에 있어서는 앞서 MCMC 추정의 비수렴 비율이 높았던 채점자-문항당 채점건수가 200건과 100건 조건들에서 95% 신뢰구간 내 참값의 포함 비율이 .8에 미치지 못했고, 채점건수가 50건과 25건일 때에도 .8에 가까운 낮은 포함 비율을 보여 HRM은 문항에 따라 채점자 모수가 차이가 있는 자료에 대해 추정의 정확성이 현저히 떨어지는 것을 확인할 수 있었다.

2) 편향 및 제곱편향

채점자 모수에 대한 추정치의 정확성을 평가하기 위하여 편향(이하 bias)과 제곱편향(이하 SB)을 분석한 결과는 다음 <표 Ⅵ-4>, <표 Ⅵ-5>와 같다.

<표 Ⅵ-4> 채점자 모수 추정치 편향

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	bias	모수 수	bias
E-HRM	100%	200	250	-.002	2,500	-.025
		100	500	.000	5,000	-.032
		50	1,000	.002	10,000	-.040
		25	2,000	-.004	20,000	-.050
	70%	200	250	.002	2,500	-.026
		100	500	-.007	5,000	-.031
		50	1,000	-.006	10,000	-.033
		25	2,000	-.008	20,000	-.033
	40%	200	250	.003	2,500	-.026
		100	500	-.005	5,000	-.029
		50	1,000	.001	10,000	-.028
		25	2,000	-.006	20,000	-.012
HRM	100%	200	250	.006	250	-
		100	500	.008	500	-

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	bias	모수 수	bias
		50	1,000	.011	1,000	-
		25	2,000	.009	2,000	-
	70%	200	250	.015	250	-
		100	500	-.004	500	-
		50	1,000	.000	1,000	-
		25	2,000	-.003	2,000	-
	40%	200	250	-.003	250	-
		100	500	-.015	500	-
		50	1,000	-.010	1,000	-
		25	2,000	-.009	2,000	-

채점자 모수 추정값의 bias에 대해 분석한 결과, 먼저 E-HRM의 엄격성 모수는 bias의 절댓값이 .01 이하로 매우 낮고, 특정한 방향으로의 편향성이 관찰되지 않았다. HRM의 엄격성 모수에 대한 bias 역시 일부 조건에서 ±.015 수준의 bias가 나타났지만 다른 조건들에 비해서 상대적으로는 절댓값이 크지만 절대적으로 큰 편향으로 보이지는 않는다. 그리고 E-HRM의 분산성 모수에서는 음수로 일관된 방향을 가지는 bias가 관찰되는데 분산성 모수는 항상 양의 값이기 때문에 분산성 모수가 작아지는 방향으로 수축(shrinking)하는 것으로 이해할 수 있고, 채점자-문항 당 채점건수가 적어질수록 수축 정도가 커져 모수 추정의 정확도가 떨어지는 경향이 나타났다.

<표 Ⅵ-5> 채점자 모수 추정치 SB

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	SB	모수 수	SB
E-HRM	100%	200	250	.001	2,500	.005
		100	500	.002	5,000	.007

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	SB	모수 수	SB
		50	1,000	.004	10,000	.012
		25	2,000	.005	20,000	.017
	70%	200	250	.001	2,500	.006
		100	500	.002	5,000	.008
		50	1,000	.004	10,000	.017
		25	2,000	.005	20,000	.058
	40%	200	250	.001	2,500	.005
		100	500	.002	5,000	.008
		50	1,000	.003	10,000	.014
		25	2,000	.006	20,000	.130
HRM	100%	200	250	.011	250	-
		100	500	.007	500	-
		50	1,000	.011	1,000	-
		25	2,000	.013	2,000	-
	70%	200	250	.008	250	-
		100	500	.012	500	-
		50	1,000	.015	1,000	-
		25	2,000	.016	2,000	-
	40%	200	250	.017	250	-
		100	500	.018	500	-
		50	1,000	.017	1,000	-
		25	2,000	.020	2,000	-

SB에 대해 두 가지 채점자 모수를 비교해보면 엄격성 모수는 모수 추정 과정에 분산성 모수의 10배의 채점건수가 활용되기 때문에 분산성 모수에 비해서 상대적으로 작은 SB 값을 가지는 결과를 보였고, 채점자-문항당 채점건수가 줄어들수록 SB가 증가하여 모수 추정의 정확도가 낮아지는 것을 확인할 수 있었다. 그리고 앞서 일방향적 bias가 관찰되었던 분산성 모수는 SB에서도 채점자-문항당 채점건수가 적어질수록 모

수의 수축이 커진 영향으로 SB가 커지고, 모수 추정의 정확도가 떨어지는 것으로 판단된다.

3) 분산

분산(이하 VAR)은 모형을 통해 획득한 추정치가 점 추정치를 중심으로 얼마나 퍼져있는지 즉, 모수 추정치가 가진 추정의 오차를 표현하는 지표이고, VAR이 작을수록 추정의 효율성이 높다고 볼 수 있다. 12가지 모의실험 설계에 따른 모의실험 자료에 E-HRM과 HRM 모형을 적용한 분석 결과에서 나타난 VAR은 다음 <표 Ⅵ-6>과 같다.

<표 Ⅵ-6> 채점자 모수 추정치 VAR

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	VAR	모수 수	VAR
E-HRM	100%	200	250	.003	2,500	.006
		100	500	.005	5,000	.007
		50	1,000	.007	10,000	.009
		25	2,000	.011	20,000	.015
	70%	200	250	.003	2,500	.005
		100	500	.004	5,000	.007
		50	1,000	.006	10,000	6.959
		25	2,000	.010	20,000	65.591
	40%	200	250	.002	2,500	.005
		100	500	.003	5,000	.012
		50	1,000	.005	10,000	3.399
		25	2,000	.009	20,000	237.436
HRM	100%	200	250	.020	250	.003
		100	500	.023	500	.003
		50	1,000	.027	1,000	.005
		25	2,000	.035	2,000	.006

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	VAR	모수 수	VAR
	70%	200	250	.012	250	.003
		100	500	.018	500	.003
		50	1,000	.022	1,000	.005
		25	2,000	.028	2,000	.006
	40%	200	250	.008	250	.003
		100	500	.012	500	.004
		50	1,000	.015	1,000	.005
		25	2,000	.020	2,000	.006

VAR 분석 결과, 엄격성 모수와 분산성 모수 모두에서 채점자-문항당 채점건수가 작아질수록 VAR이 커져 추정의 효율성이 낮아지는 경향이 나타났다. 두 가지 채점자 모수는 정상범위 채점자 비율 '100%' 조건에서는 E-HRM이 전반적으로 낮은 VAR을 보였고, 정상범위 채점자 비율 '70%'와 '40%' 조건에서는 채점건수가 200건, 100건일 경우에 정상범위 채점자 비율 '100%' 조건과 거의 동일한 수준의 VAR을 보이던 E-HRM 분석 결과가 50건부터 커지기 시작하여 25건에서는 매우 높은 값을 보였다. 지나치게 큰 VAR을 가질 가능성이 높은 비수렴 모수들을 제외하고, 수렴된 모수들만 분석하여도 50건과 25건 조건에서는 각각 6.717, 9.624의 VAR을 보여 추정 오차가 지나치게 크고 효율이 떨어지는 것으로 보인다.

그런데 정상범위 채점자 비율 '70%', 채점자-문항당 채점건수가 '50건'인 경우의 VAR 값을 살펴본 결과, 1만개의 모수 중 상위 10개의 모수의 VAR(59592.6, 6324.9, 1577.9, 572.2, 562.8, 335.8, 287.5, 37.6, 36.1, 26.1)이 VAR 총합의 99.7%로 거의 대부분을 차지하였다. 그리고 이를 제외한 VAR은 .024로 매우 낮은 수준이었고, 채점건수가 100건인 조건에서의 모수별 VAR 최댓값인 .042를 넘는 모수는 83개(전체 모수의 0.83%)로 매우 낮은 비율이었다. 마찬가지로 정상범위 채점자 비율 '40%' 조건

하에서도 채점자-문항당 채점건수가 '50건'일 때, VAR 상위 10개 모수가 차지하는 VAR 비율이 91.5%, 상위 20개 모수의 VAR 비율이 97.9%로 매우 높았다. 따라서 전체 VAR 결과는 비록 오차가 크고 효율이 낮은 것으로 평가된 채점건수 50건인 조건에서도 극소수의 모수 추정치를 제외하고는 절대 다수의 모수 추정치가 VAR이 낮기에 충분한 효율성을 가지고 있는 것으로 판단된다.

정상범위 채점자 비율 '70%', 채점자-문항당 채점건수가 25건인 경우에 대해서도 50건과 마찬가지 접근으로 살펴본 결과, VAR 상위 10개 모수가 차지하는 VAR 비율이 79.5%, 상위 20개 모수의 VAR 비율이 86.7%로 50건인 조건에 비해서는 그 비율이 높지 않았다. 그리고 정상범위 채점자 비율이 '70%'로 동일하고 채점자-문항당 채점건수가 '100건'인 조건에서의 모수별 VAR 최댓값인 .042를 넘는 모수는 647개(전체 모수의 6.47%)로 이들을 제외한 VAR 평균은 .013으로 낮은 수준이었지만, 이상치 수준의 오차를 갖는 모수 추정치가 많기 때문에 채점건수 25건 조건에서 E-HRM이 충분한 효율성을 가진다고 판단하기는 어렵다.

4) 평균제곱오차

평균제곱오차 MSE는 정확성과 관련된 SB와 효율성과 관련된 VAR을 합쳐 전체 오차를 평가하는 지표로 모의실험 설계 조건별 채점자 모수 추정치의 MSE는 다음 표 <표 Ⅵ-7>과 같다.

<표 Ⅵ-7> 채점자 모수 추정치 MSE

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	MSE	모수 수	MSE
E-HRM	100%	200	250	.005	2,500	.011
		100	500	.007	5,000	.014
		50	1,000	.011	10,000	.021

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	엄격성 모수		분산성 모수	
			모수 수	MSE	모수 수	MSE
		25	2,000	.017	20,000	.033
	70%	200	250	.004	2,500	.011
		100	500	.006	5,000	.015
		50	1,000	.010	10,000	6.976
		25	2,000	.015	20,000	65.649
	40%	200	250	.003	2,500	.010
		100	500	.005	5,000	.021
		50	1,000	.008	10,000	3.413
		25	2,000	.014	20,000	237.565
HRM	100%	200	250	.032	250	-
		100	500	.030	500	-
		50	1,000	.039	1,000	-
		25	2,000	.048	2,000	-
	70%	200	250	.020	250	-
		100	500	.030	500	-
		50	1,000	.037	1,000	-
		25	2,000	.044	2,000	-
	40%	200	250	.025	250	-
		100	500	.029	500	-
		50	1,000	.032	1,000	-
		25	2,000	.040	2,000	-

모수 추정치의 종합적인 오차인 MSE는 모든 조건에서 SB(.001～.130 범위)보다는 VAR(.003～237.436)의 영향을 더 크게 받았기 때문에 VAR의 경향성과 유사한 결과를 보였다. E-HRM은 정상범위 채점자 비율이 '100%'인 조건들에서 채점자-문항당 채점건수가 줄어들수록 오차가 커졌지만, 다른 정사범위 채점자 비율 조건들에 비해서는 심각한 수준의 오차 증가폭을 보이지는 않았다. 그러나 정상범위 채점자 비율이 '70%', '40%'인 조건들에서는 채점건수가 줄어들수록 MSE가 커지면서 50건부

터는 매우 큰 오차를 보이고, 25건부터는 수용하기 어려운 수준의 오차를 보이는 것으로 판단된다. 그리고 HRM은 모든 조건에서 엄격성 모수의 MSE가 E-HRM에 비해 높은 값을 나타내어 채점자-문항 독립성이 보장되지 않는 채점 자료에 대해서 HRM은 E-HRM보다 큰 오차를 가진다고 결론내릴 수 있다.

5) 정상범위 외 분산성 모수 판별

E-HRM에 기반하여 생성된 모의실험 자료에 대한 HRM 분석에서는 채점자 분산성 모수의 참값을 알 수 없기 때문에 bias, SB, VAR, MSE를 통한 E-HRM과의 비교가 불가능하다는 한계점이 있다. 때문에 HRM이 문항별로 채점자 분산성 모수에 차이가 있는 자료에서 어떻게 작동하는지를 확인하기 위해 '정상범위를 벗어나는 분산성 모수 참값을 가지는 채점자를 HRM의 분산성 모수 추정치로 정확하게 판별해낼 수 있는가'라는 관점에서 분석하고자 하였다. 이를 위해 모의자료 생성 조건으로부터 '채점자 단위에서 분산성 모수 ψ_{ir}의 참값들 중 하나라도 정상범위 기준 범위($\psi < 0.5$)를 초과하지 여부' 더미변수를 생성하고, HRM 분석 결과로부터 '분산성 모수 추정치 $\hat{\psi}_r$가 정상범위 기준(.5)을 초과하는지 여부' 더미 변수를 생성하여 조건별로 두 더미 변수들을 교차 분석하였다. 아울러 E-HRM과의 비교를 위하여 E-HRM 분석 결과 역시 분산성 모수 ψ_{ir}의 참값과 추정치 각각의 정상범위 기준 초과 여부 변수를 생성하고, 교차 분석하여 참값이 정상범위 기준을 넘었을 때 추정치가 이를 판별해내는가를 분석하였다. 그 결과는 <표 Ⅵ-8>에 비교 제시하였다.

<표 Ⅵ-8> 모형별 분산성 모수 이상 기준 판별률

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	모수 참값 판별	전체 모수 수	추정치 판별		정확 판별률 (%)
					정상 범위	정상 범위 외	
E-HRM	100%	200건	정상	2,500	2,488	12	99.5
		100건	정상	5,000	4,951	49	99.0
		50건	정상	10,000	9,845	155	98.5
		25건	정상	20,000	19,542	458	97.7
	70%	200건	정상	2,368	2,337	31	98.7
			비정상	132	4	128	97.0
		100건	정상	4,751	4,684	67	98.6
			비정상	249	18	231	92.8
		50건	정상	9,511	9,302	209	97.8
			비정상	489	59	430	87.9
		25건	정상	18,973	18,349	624	96.7
			비정상	1,027	198	829	80.7
	40%	200건	정상	2,253	2,207	46	98.0
			비정상	247	14	233	94.3
		100건	정상	4,475	4,378	97	97.8
			비정상	525	57	468	89.1
		50건	정상	8,980	8,685	295	96.7
			비정상	1,020	146	874	85.7
		25건	정상	17,989	17,171	818	95.5
			비정상	2,011	429	1,582	78.7
	전체		정상범위	106,800	103,939	2,861	97.3
			정상범위외	5,700	925	4,775	83.8
HRM	100%	200건	정상	250	250	0	100.0
		100건	정상	500	500	0	100.0
		50건	정상	1,000	1,000	0	100.0
		25건	정상	2,000	2,000	0	100.0
	70%	200건	정상	200	200	0	100.0
			비정상	50	45	5	10.0
		100건	정상	400	400	0	100.0
			비정상	100	67	33	33.0

모형	정상범위 채점자 비율	채점자-문항 당 채점건수	모수 참값 판별	전체 모수 수	추정치 판별		정확 판별률 (%)
					정상 범위	정상 범위 외	
		50건	정상	800	800	0	100.0
			비정상	200	140	60	30.0
		25건	정상	1,600	1,599	1	99.9
			비정상	400	261	139	34.8
	40%	200건	정상	150	150	0	100.0
			비정상	100	73	27	27.0
		100건	정상	300	300	0	100.0
			비정상	200	135	65	32.5
		50건	정상	600	600	0	100.0
			비정상	400	277	123	30.8
		25건	정상	1,200	1,197	3	99.8
			비정상	800	545	255	31.9
	전체		정상범위	106,800	103,939	2,861	97.3
			정상범위외	5,700	925	4,775	83.8

※ 음영 표시: 민감도(실제 기준 초과 채점자를 기준 초과 채점자로 판별할 확률)

<표 Ⅵ-8>에서 전체 조건에서 정상범위 외의 참값을 가진 모수를 정상범위 외의 추정치로 판별할 확률을 나타내는 민감도가 E-HRM은 83.8%, HRM은 31.4%로 큰 차이를 보였다. 이 결과는 실험자료 생성이 E-HRM 모형에 따라 이루어졌고, 분산성 모수 참값 역시 채점자-문항 독립성을 완화한 조건이었기 때문에 당연한 결과일 수 있다. 하지만 이는 HRM이 채점자-문항 독립성이 가정되지 않는 상황 즉, 채점자가 문항에 따라 일관된 채점을 하지 못한 채점 상황이라면 특정 문항에 대한 신뢰할 수 없는 채점 결과를 잡아낼 수 없다는 취약점을 보여주는 근거가 될 수 있다.

분석 결과를 구체적으로 살펴보면, 먼저 정상범위 채점자 비율이 '100%'인 조건하에서는 실제로 정상범위를 초과하는 분산성 모수가 존재하지 않기 때문에 민감도는 계산할 수 없고, 정상범위 참값인 경우를 정

상범위 추정치로 판별하는 특이도만 산출 가능하다. 이 때, HRM의 경우 정상범위 채점자 비율이 '100%' 조건에서 모든 분산성 모수를 정상범위로 판단하였기 때문에 특이도가 100%였다. 반면에 E-HRM의 경우에는 정상범위 채점자 비율이 '100%' 조건일 때, 채점자-문항당 채점건수가 '200건', '100건', '50건', '25건'에서 각각 99.5%, 99.0%, 98.5%, 97.7%로 나타나 채점자-문항당 채점건수가 적어질수록 특이도가 낮아지는 결과를 보였다.

정상범위 채점자 비율이 '70%'인 조건하에서 E-HRM 분석 결과는 채점자-문항당 채점건수가 감소할수록 추정치를 통해 정상범위 외의 참값을 가진 분산성 모수를 정확하게 판별해내는 민감도가 낮아지는 것을 확인할 수 있다. 정상범위 채점자 비율이 '70%'인 조건하에서 채점자-문항당 채점건수가 '200건'일 때는 97.0%로 높은 민감도를 보이다가 '100건'에서는 92.8%로 낮아졌고, '50건'과 '25건'에서는 각각 87.9%, 80.7%로 90%에 못 미치는 민감도를 나타내었다. HRM의 경우에는 정상범위 채점자 비율이 '70%'인 조건하에서 채점자-문항당 채점건수가 많은 '200건'이 나머지 채점자-문항당 채점건수 조건들에 비하여 현저하게 낮은 10.0%의 민감도를 보였는데, 이는 앞서 분산성 모수의 수렴 비율 분석 결과에서 나타난 현상과 유사하게 사례수가 충분히 많기 때문에 문항별로 다른 분산성 모수들의 영향이 크게 발현된 결과로 나타난 추정의 불안정이 원인이라고 판단된다.

다음으로 정상범위 채점자 비율이 '40%'인 조건하에서 E-HRM 분석 결과는 채점자-문항당 채점건수가 감소할수록 정상범위 외의 모수를 정확하게 판별해내는 민감도가 떨어지는 경향이 정상범위 채점자 비율이 '70%'인 조건들에서보다 좀 더 뚜렷하게 나타났다. 정상범위 채점자 비율이 '70%'인 조건하에서 채점자-문항당 채점건수가 '200건'일 때는 94.3%로 95%에 가까운 높은 민감도를 보였으나, '100건'에서 89.1%, '50건'에서 85.7%, '25건'에서 78.7%로 낮아지는 결과를 보였다. 그리고 정상범위 채점자 비율이 '40%'인 조건에서의 HRM의 분산성 모수 판별 결

과, 민감도에 대한 채점자-문항당 채점건수에 따른 영향은 관찰되지 않았다.

그리고 채점자 모수 참값을 제시한 <표 Ⅵ-1>을 상기해보면 '저신뢰', '엄격-저신뢰' 유형의 채점자들은 전체 분산성 모수 ψ_{ir} 중 '>.5'로 정상범위 외 참값을 가진 분산성 모수의 수가 1~4개로 설정되어 있다. 따라서 <표 Ⅵ-8>의 HRM 분석 결과를 채점자별 정상범위 외 참값을 가지는 ψ_{ir} 수로 세분화해서 살펴볼 수 있고, 정상범위 채점자 비율이 '70%', '40%'인 조건들에서 E-HRM 분석 결과를 세분화하여 제시하면 <표 Ⅵ-9>와 같다.

<표 Ⅵ-9> HRM 분산성 모수 이상 기준 판별률

정상범위 채점자 비율	채점자-문항 당 채점건수	채점자 내 분산성 모수 정상범위 외 참값 수	전체 모수 수	HRM		정확 판별률 (%)
				정상범위	정상범위외	
70%	200건	0	200	200		.0
		1	4	4		.0
		2	17	17		.0
		3	22	20	2	9.1
		4	7	4	3	42.9
	100건	0	400	400		.0
		1	16	16		.0
		2	31	20	11	35.5
		3	41	28	13	31.7
		4	12	3	9	75.0
	50건	0	800	800		.0
		1	45	44	1	2.2
		2	59	41	18	30.5
		3	58	44	14	24.1
		4	38	11	27	71.1
	25건	0	1,600	1,599	1	.1
		1	74	65	9	12.2
		2	113	75	38	33.6
		3	125	92	33	26.4
		4	88	29	59	67.0

정상범위 채점자 비율	채점자-문항 당 채점건수	채점자 내 분산성 모수 정상범위 외 참값 수	전체 모수 수	HRM		정확 판별률 (%)
				정상범위	정상범위외	
40%	200건	0	150	150		.0
		1	15	15		.0
		2	36	28	8	22.2
		3	36	26	10	27.8
		4	13	4	9	69.2
	100건	0	300	300		.0
		1	35	35		.0
		2	53	38	15	28.3
		3	64	47	17	26.6
		4	48	15	33	68.8
	50건	0	600	600		.0
		1	81	76	5	6.2
		2	110	81	29	26.4
		3	117	88	29	24.8
		4	92	32	60	65.2
	25건	0	1,200	1,197	3	.3
		1	155	146	9	5.8
		2	241	177	64	26.6
		3	242	168	74	30.6
		4	162	54	108	66.7
전체		0	9,000	8,996	4	.0
		1	425	401	24	5.6
		2	660	477	183	27.7
		3	705	513	192	27.2
		4	460	152	308	67.0

※ 음영 표시: 민감도(실제 기준 초과 채점자를 기준 초과 채점자로 판별할 확률)

<표 Ⅵ-9>를 살펴보면 모든 조건에서 채점자 내에서 정상범위 외 참값을 가지는 분산성 모수 ψ_{ir}의 개수가 증가할수록 민감도가 높아지는 결과를 확인할 수 있는데, 이는 모든 문항에서 평균적인 분산성이 높아져 채점자 단위에서의 분산성이 높아진 결과, HRM에서도 정상범위 기준(.5)을 초과하는 채점 분산성을 잡아낼 수 있게 된 것이다. 이를 반대로 해석하자면 전체 문항에서 전반적인 채점자 분산성이 낮다면 채점자가 1~2개 문항에서 채점의 일관성이 떨어져 높은 분산성을 가지더라도

이를 잡아내기 어렵다는 의미가 될 수 있다.

나. 문항모수

조건 및 모형별로 문항 변별도 모수 추정치는 500개씩이었고, 500개씩의 문항 변별도 모수 추정치에 대한 bias와 오차인 SB, VAR, MSE를 제시하면 다음 <표 Ⅵ-10>과 같다.

<표 Ⅵ-10> 문항 변별도 모수 추정치 bias, SB, VAR, MSE

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
E-HRM	100%	200	.026	.011	.011	.022
		100	.032	.010	.011	.021
		50	.029	.010	.011	.021
		25	.028	.010	.011	.021
	70%	200	.027	.011	.011	.022
		100	.032	.010	.011	.021
		50	.036	.011	.011	.022
		25	.028	.010	.011	.021
	40%	200	.023	.011	.011	.022
		100	.035	.010	.012	.021
		50	.038	.012	.012	.023
		25	.033	.010	.012	.021
HRM	100%	200	.026	.011	.011	.022
		100	.031	.010	.011	.021
		50	.030	.010	.011	.021
		25	.027	.010	.011	.021
	70%	200	.028	.011	.011	.022
		100	.032	.010	.011	.021
		50	.035	.011	.011	.023
		25	.026	.010	.011	.021

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
	40%	200	.024	.011	.012	.023
		100	.034	.010	.012	.022
		50	.036	.012	.012	.024
		25	.030	.010	.012	.021

먼저, bias의 범위는 [.023, .038]로 문항 변별도 모수는 평균적으로 증가하는 방향의 bias가 있었던 것으로 나타났다. 그리고 SB는 모든 조건과 모형에 걸쳐 [.010, .012]의 범위를 나타내어 '정상범위 채점자 비율'이나 '채점자-문항당 채점건수'에 따른 차이가 관찰되지 않았다. 또한 VAR 역시 모든 조건과 모형 하의 분석 결과들이 [.011, .012]의 범위를 보여 '정상범위 채점자 비율'과 '채점자-문항당 채점건수' 조건에 영향을 받지 않았다. 그리고 SB와 VAR의 합인 MSE 역시 모의실험 연구 설계 조건에 따른 차이를 보이지 않았다.

<표 Ⅵ-11> 문항 곤란도 모수 추정치 bias, SB, VAR, MSE

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
E-HRM	100%	200	.004	.029	.030	.059
		100	.002	.028	.029	.058
		50	.004	.028	.029	.057
		25	.002	.028	.030	.057
	70%	200	.006	.030	.030	.060
		100	.000	.029	.030	.059
		50	.001	.029	.030	.059
		25	.002	.029	.030	.060
	40%	200	.006	.031	.032	.063
		100	.004	.030	.031	.062
		50	.004	.029	.031	.060

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
		25	.008	.030	.032	.062
HRM	100%	200	.004	.030	.030	.060
		100	.003	.029	.029	.058
		50	.005	.028	.029	.057
		25	.005	.028	.030	.058
	70%	200	.007	.032	.031	.063
		100	.001	.030	.030	.060
		50	.003	.029	.030	.060
		25	.004	.029	.031	.060
	40%	200	.003	.040	.033	.073
		100	.002	.035	.032	.067
		50	.000	.033	.032	.065
		25	.004	.032	.033	.065

문항 곤란도 모수의 bias, SB, VAR, MSE를 제시한 <표 Ⅵ-11>에서도 변별도 모수의 유사한 결과를 보인다. Bias의 범위는 [.000, .008]로 문항 곤란도 모수는 특정한 방향으로는 bias가 없었던 것으로 나타났다. 그리고 SB는 조건과 모형에 따라서 [.028, .040]의 범위를, VAR은 [.029, .033]의 범위를 보여 '정상범위 채점자 비율'이나 '채점자-문항당 채점건수' 조건에 영향을 받지 않았고, 두 지표를 합한 MSE 역시 모의실험 설계 조건에 따른 차이가 없었던 것으로 나타났다.

문항 변별도 모수와 문항 곤란도 모수가 모의실험 설계 조건에 따른 영향을 받지 않은 것과 관련하여 '채점자-문항당 채점건수'의 경우 채점자 모수 추정에서만 추정에 활용되는 채점건수가 줄어들 뿐 문항 모수 추정에서는 이 조건과 무관하게 동일하기 때문에 예상 가능한 결과이다. 하지만 '채점 특성'이 문항 모수들의 추정에 영향을 주지 않았다는 점은 주목할 만한데, 이는 채점자 수준과 피험자 수준으로 나뉘어 추정하는 HRM과 E-HRM의 특성상 '채점 특성'이 엄격 혹은 관대하거나 낮은 신

뢰도를 보이는 채점자를 포함하더라도 채점자 수준에서 이를 보정하여 안정적으로 문항 모수가 추정 가능하도록 하는 모형의 유용성을 잘 보여주는 결과이다.

다. 피험자 모수

모의실험 설계 조건 및 모형별로 피험자는 500명이었기에 피험자 모수 추정치는 500개씩이었고, 50번의 반복으로 인하여 25,000개씩 총 60만개의 피험자 모수 추정치가 분석되었다. 이 피험자 모수 추정치들에 대한 bias와 오차인 SB, VAR, MSE를 제시하면 다음 <표 Ⅵ-12>와 같다.

<표 Ⅵ-12> 피험자 능력 모수 추정치 bias, SB, VAR, MSE

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
E-HRM	100%	200	.004	.288	.269	.557
		100	.004	.285	.267	.553
		50	.004	.286	.268	.554
		25	.004	.288	.269	.556
	70%	200	.004	.290	.272	.562
		100	.004	.287	.270	.557
		50	.004	.287	.269	.555
		25	.004	.290	.273	.562
	40%	200	.004	.296	.279	.575
		100	.004	.296	.276	.573
		50	.004	.293	.275	.569
		25	.004	.295	.278	.573
HRM	100%	200	.004	.289	.270	.560
		100	.004	.286	.268	.555
		50	.004	.287	.269	.556
		25	.004	.289	.269	.558

모형	정상범위 채점자 비율	채점자-문항당 채점건수	bias	SB	VAR	MSE
	70%	200	.004	.293	.274	.568
		100	.004	.290	.272	.562
		50	.004	.289	.271	.560
		25	.004	.292	.275	.567
	40%	200	.004	.301	.283	.584
		100	.004	.302	.281	.583
		50	.004	.299	.280	.579
		25	.004	.300	.283	.583

피험자 모수의 bias부터 살펴보면 전체 모의실험 설계 조건 및 모형별 분석에서 나타난 bias의 범위는 [.004, .004]로 피험자 모수는 평균적으로 양의 방향으로 bias가 나타났지만 큰 편향은 없었던 것으로 판단된다. 별도 표로 제시하지는 않았지만 조건 및 모형별 피험자 능력 모수의 표준편차 추정치의 평균은 [.828, .840]의 범위를 나타내었고, 이러한 수치는 전체 분석 결과의 능력 모수 표준편차인 1에 비하여 0에 가까워진 결과로 적용 연구에서도 분석된 바와 같이 양극 평정 점수가 채점자 분산성으로 인하여 중심으로 쏠리게 되는 HRM의 특성에 따라 피험자 능력 모수는 평균인 0으로 수축한 결과이다.

이러한 능력 모수 수축의 영향으로 모든 설계 조건과 모형 분석에서 결과에서 SB는 [.285, .302]의 범위를 보였는데, 분산성 모수가 크지 않아 능력모수의 수축이 적을 것으로 예상되는 정상범위 채점자 비율 '100%' 조건에서의 SB가 분산성 모수가 상대적으로 크고, 수축이 좀 더 있을 것으로 예상되는 정상범위 채점자 비율 '70%'와 '40%' 조건들 하에서의 SB보다 조금 낮은 경향을 보였다. 그리고 VAR의 경우 [.267, .283]의 범위를 나타내어 모의실험 연구 설계 조건과 모형에 따른 영향이 크지 않았던 것으로 보이고, 이러한 결과는 MSE에서도 설계 조건과 모형에 따른 차이가 크지 않은 결과로 이어진다.

Ⅶ. 결론

1. 요약

이 연구는 구성형 문항 채점에서 문항에 따라 다르게 나타나는 채점자 특성의 영향을 정확하게 파악하기 위하여 E-HRM을 제안하고, 이 모형의 정확성과 효율성을 검증하는 데에 목적을 두고 있다. 이에 따라 아래의 두 가지 연구 문제를 설정하고 연구를 진행하였다.

첫째, 채점자-문항 독립성 제약이 완화된 위계적 채점자 모형은 채점자 모수의 문항별 차이를 판별할 수 있는가?

둘째, 채점자-문항 독립성 제약이 완화된 위계적 채점자 모형의 모수 추정은 얼마나 정확하고 효율적인가?

먼저 첫 번째 연구문제를 해결하기 위하여 HRM 모형의 두 가지 채점자 모수에 대한 채점자-문항 독립성 제약을 완화하는 방식으로 확장한 세 가지 모형을 제시하고, 기존 HRM과 함께 대표적인 국제비교 대규모 성취도 검사인 PIRLS 2006 자료를 대상으로 적용 연구를 실시하였다. 일반적으로 모형을 개발하는 절차는 모형을 제안하고, 모의실험을 통해서 이 모형을 검증하고, 이를 실제 자료에 대한 적용 연구를 통해 유용성을 밝히는 방식으로 진행된다. 하지만 본 연구에서는 E-HRM을 주요 연구 모형으로 고려하였으나, 어떤 채점자 모수에 대한 제약을 완화하는 것이 실제 자료에 더 적합한 방식인지 알 수 없기 때문에 두 가지 대안모형과 채점자-문항 독립성 제약을 유지한 HRM까지 총 4가지 모형을 PIRLS 2006 자료에 적용하는 연구를 먼저 수행하였다. 이를 통해 E-HRM이 실제 자료에서 기저모형인 HRM이나 나머지 대안모형들에 비해 더 적용 가능성이 높은지를 확인하고자 하였다. 또한 적용 연구 과

정에서 E-HRM의 정확성과 효율성을 검증하기 위한 모의실험 연구에서 고려해야할 조건들에 대한 탐색도 함께 진행하였다.

적용 연구에서의 주요결과를 제시하면, 첫째로 기존 HRM 모형과 채점자-문항 독립성 제약을 완화한 3가지 모형으로 PIRLS 2006 자료에 대해 국가별 분석을 진행한 결과, 엄격성 모수에서 채점자-문항 독립성 제약을 완화한 대안모형1과 그에 더해 분산성 모수까지 채점자-문항 독립성 제약을 완화한 대안모형2는 채점자 모수의 MCMC 연쇄가 수렴되지 않는 비율이 높았을 뿐만 아니라 HRM과 E-HRM에서는 안정적으로 추정된 문항 모수마저 수렴이 되지 않는 결과가 빈번히 나타났다. 그 결과, 대안모형들은 최소한 PIRLS 2006 자료와 같은 형태의 다수의 구성형 문항으로 이루어진 검사에는 적합하지 않다고 결론내리고, E-HRM을 연구의 제안 모형으로 적절함을 확인하였다.

둘째, HRM과 E-HRM의 모형 비교를 통하여 채점자-문항 독립성을 제약하는 HRM에 비하여 제약을 완화한 E-HRM이 다수의 국가별 분석에서 모형 적합도(Deviance, DIC)가 향상되는 비율이 높아 실제 자료는 채점자-문항 독립성 가정이 지켜지지 않을 가능성이 높다는 것을 확인할 수 있었다. 그리고 채점 및 채점자 단위 Deviance 분해 결과와 E-HRM으로 획득한 분산성 모수 ψ_{ir}의 관계 분석 결과로부터 E-HRM과 HRM의 채점자 단위 Deviance 비교가 채점자 특성이 문항에 따라 차별적으로 기능하는 것을 탐지하는데 유용할 수 있음을 알 수 있었다. 그리고 이 과정에서 E-HRM에서 모수 추정이 불안정할 경우 Deviance와 DIC가 큰 변동폭을 가질 수 있다는 취약점을 확인하였다.

셋째, HRM에서 채점자별 분산성 모수 ψ_r가 정상범위인 '<.5' 기준을 충족했더라도 E-HRM에서는 채점자-문항 단위의 분산성 모수 ψ_{ir}가 .5를 초과하여 나타나는 경우를 관찰할 수 있었다. 이는 채점자 특성의 영향이 전체 문항에서는 평균적으로 일관된 것처럼 보이더라도, 특정 문항에서는 일관되지 않게 나타날 수 있다는 것을 드러내는 결과로, 본 연구의 제안모형인 E-HRM이 의미 있는 상세한 채점자 정보를 추가로 획득

하게 해준다는 사실을 증명한다. 나아가 적용 연구의 국가별 분석에서는 E-HRM과 HRM의 분산성 모수 비교를 통해 실제로 이질적인 채점 분포를 가진 문항들이 혼재되어 있는 경우에 HRM은 추정에 문제가 발생하지만, 채점자와 문항 간의 독립성 제약이 없는 E-HRM은 이를 문제없이 추정해내어 HRM의 약점이자 E-HRM의 강점인 지점이 드러났다.

마지막으로 E-HRM은 MCMC 연쇄의 수렴이 되지 않고 HRM만이 수렴하는 국가들에 대한 분석 결과에서는 이 국가들이 대체로 채점자-문항당 채점건수가 적은 경우가 대부분이었다. 이론적으로 1개의 관찰값만 있더라도 추정이 가능하다는 MCMC 방법에서도 채점자-문항당 채점건수가 어느 정도 이상 확보되지 못하면 E-HRM이 수렴하지 못하거나 MCMC 연쇄의 길이를 매우 늘려야 제대로 추정이 가능한 것으로 보인다. 이와 관련하여 채점건수에 따른 모수의 수렴 비율, MCSE, ESS를 분석한 결과, 최소 10~30건 이상의 채점건수가 필요할 것으로 추정하였다.

이상의 적용 연구 결과를 토대로 설계한 모의실험 연구는 두 번째 연구문제에 대한 답을 제시하기 위한 과정이다. 모의실험 연구에서는 적용 연구에서 모수의 수렴과 효율성(MCMC 관점에서 연쇄의 정확도로 적용 연구에서는 ESS와 MCSE)에 가장 크게 영향을 미치는 요인으로 나타난 '채점자-문항당 채점건수'를 설계의 주요 요인으로 설정하였고, 또한 적용 연구에서 HRM과 비교하여 E-HRM이 가지는 강점을 확인할 수 있었던 상황들(문항별 채점 분포가 이질적인 상황, 특정 문항에서만 채점이 일관되지 않은 상황 등)에서 정말로 E-HRM이 HRM에 비해 강점을 가지는가를 확인할 수 있도록 '정상범위 채점자 비율'을 설계 요인으로 고려하였다. 최종적으로 '채점자-문항당 채점건수' 요인의 4개 수준과 '정상범위 채점자 비율' 요인의 3개 수준을 조합하여 총 12개의 조건하에서 모의실험 자료를 생성하고, 이 자료에 E-HRM과 HRM을 적용하여 E-HRM의 정확성과 효율성을 분석하였다.

모의실험 연구 결과의 주요 결과를 요약하면 첫째, 채점자-문항당 채

점건수가 적을수록 E-HRM의 모수 추정 정확성 및 효율성과 관련하여 채점자 모수 추정치의 bias, SB, VAR이 커져서 정확도와 효율성이 낮아지고 MSE가 커져 전반적 오차가 증가하는 경향이 나타났다. 정상범위 채점자 비율이 '100%' 조건하에서 E-HRM의 분산성 모수 추정치에 대한 RMSE(MSE의 제곱근)는 채점건수가 가장 많은 '200건' 조건을 기준으로 '100건'일 때는 15.7%, '50건'일 때는 41.0%, '25건'일 때는 75.3% 증가하는 것으로 분석되었다. 그리고 정상범위 채점자 비율이 '70%' 조건의 경우, 채점건수 '200건' 조건일 때를 기준으로 '100건'일 때는 17.3% 증가하였고, '50건'과 '25건' 조건에서는 이상치의 영향으로 RMSE가 급증하여 24배, 77배로 뛰었다. 이러한 채점자-문항당 채점건수 감소에 따른 RMSE 증가 경향은 정상범위 채점자 비율이 '40%' 조건에서 더 심해져서 채점건수 '200건' 조건일 때를 기준으로 '100건', '50건', '25건' 조건에서 각각 1.4배, 17배, 150배로 오차가 급격히 커졌다. 한편, MSE나 RMSE는 상대적인 비교만 가능하고 절대적인 평가가 불가능하기 때문에 오차만으로 E-HRM이 어느 정도의 채점건수까지 잘 작동하는지는 판단 내리기에는 한계가 있다.

둘째, E-HRM의 채점자 모수 추정 정확성과 관련하여 오차보다 좀 더 직관적인 이해가 가능한 베이지안 추정 신뢰구간에 대한 접근에서도 채점건수가 증가함에 따라 모수 복원력이 낮아졌다. 채점자 엄격성 모수에 대한 추정치 사후분포가 95% 신뢰구간 내에 참값을 포함하는 비율은 12개 실험설계 조건 중 정상범위 채점자 비율 '40%', 채점자-문항당 채점건수 '200건' 조건 하나를 제외하면 나머지 모든 조건에서 .95를 넘거나 근사한 수치로 상당히 높은 정확성을 보였다. 분산성 모수에 대한 추정치 사후분포는 채점자-문항당 채점건수 '200건'에서 정상범위 채점자 비율 '100%', '70%', '40%' 조건들에 대해 각각 .953, .948, .944로 .95에 근사한 수치를 보였다. 하지만 채점자-문항당 채점건수가 감소할수록 모수 추정치 신뢰구간 내 모수 참값의 포함 비율이 줄어들어 '25건'에서는 세 가지 정상범위 채점자 비율 조건들에 대해 각각 .881, .894, .922로 채

점자-문항당 채점건수 '200건'에 비해 낮은 비율을 보였다.

셋째, 정상범위 외 채점자 모수에 대한 판별 측면에서 E-HRM은 정상범위 기준을 벗어나는 채점자 분산성 모수에 대한 높은 민감도를 보였다. 문항 분산성 모수의 참값이 .5를 초과할 때, E-HRM의 추정치가 .5를 초과할 확률은 판별분석에서의 민감도이고, 참값이 .5이하일 때 추정치가 .5이하일 확률은 특이도로 해석할 수 있다. 채점자-문항당 채점건수별로 E-HRM의 민감도와 특이도를 분석한 결과, 채점자-문항당 채점건수가 '200건'일 때는, 민감도 .953, 특이도 .988로 E-HRM은 매우 높은 확률로 정상범위 채점 특성과 정상범위 외 채점 특성을 판별해내었다. 채점자-문항당 채점건수 '100건'인 경우에도 민감도 .903, 특이도 .985로 판별 정확성이 높았는데, 세부 정상범위 채점자 비율 조건별로 민감도는 '70%'에서 .928, '40%'에서 .891로 정상범위 채점자 비율이 낮아질수록 민감도가 낮아지는 결과를 보였다. 채점자-문항당 채점건수가 '50건'에서는 민감도 .864, 특이도 .977로 정상범위 외 채점을 판별해내는 기능이 다소 떨어졌고, 채점자-문항당 채점건수가 25건인 경우에는 특이도는 .967로 여전히 높았지만, 민감도는 .794로 다소 낮아진 결과를 보였다.

넷째, E-HRM은 '채점 특성'의 영향을 적절히 보정하여 엄격(혹은 관대)하거나 신뢰도가 낮은 채점자 특성이 피험자 모수와 문항 모수의 정확성 및 효율성에 미치는 영향을 통제할 수 있는 것으로 나타났다. 채점 특성이 '전체정상범위' 조건과 '일부정상범위' 조건에서의 문항 변별도 모수, 문항 곤란도 모수, 피험자 모수들은 bias, SB, VAR, MSE를 비교하였을 때, 거의 차이가 없어 E-HRM이 채점자 특성의 영향을 정확히 탐지해내고 제거하여 문항 모수와 피험자 모수에 영향을 미치지 못하도록 차단한 것으로 판단된다.

다섯째, 문항별 채점 분포가 이질적인 상황인 '일부정상범위' 설계 조건하에서 채점건수가 증가할수록 HRM은 적용 연구에서 관찰한 바와 같이 채점자 모수 추정치의 사후분포가 다봉 분포가 뚜렷한 경우가 빈번해지고 이로 인하여 채점자 엄격성 모수와 분산성 모수의 수렴 비율과 정

확성이 떨어지는 것으로 나타났다. 반면에 E-HRM은 문항별 채점 분포가 이질적인 상황인 설계 조건들에서도 채점자 모수에 대하여 100%에 가까운 수렴 비율을 보여 추정에 큰 문제가 없고, HRM보다 더 나은 모수 복원력을 보였다.

2. 논의

적용 연구와 모의실험 연구를 통해 E-HRM이 채점자 모수의 문항별 차이를 판별할 수 있는가라는 첫 번째 연구 문제에 대해서는 명확하게 긍정할 수 있게 되었다. E-HRM은 HRM과 비교하였을 때, 모형의 간명성을 희생한 것에 비하여 직접적인 문항과 피험자 능력 모수의 정확성의 향상을 가져오지는 못하는 부분은 아쉬운 점이지만, E-HRM은 채점자 분산성에 대해서 HRM보다 상세한 정보를 획득하면서도 문항과 피험자 능력 모수의 정확성에 부정적 영향을 주지 않음을 확인할 수 있었다. 그리고 HRM이 판별하지 못했던 정상범위를 초과하는 채점자의 특정 문항에 대한 분산성 모수를 정확하게 판별해낼 수 있음을 보여 본 연구에서 E-HRM을 제안한 목적을 달성하였다. 그리고 이렇게 획득된 세밀한 정보가 적절하게 피드백된다면 채점절차 개선 혹은 채점자 개인의 전문성 향상에 큰 도움이 되고 결과적으로 피험자 능력 모수의 신뢰성을 높일 수 있을 것이라 기대한다.

E-HRM을 통해 획득된 정보는 채점자 단위가 아닌 채점자-문항 단위의 세밀한 정보라는 점에서 채점자 재훈련이나 재채점에 들어가는 노력, 시간, 비용을 크게 절감할 수 있다는 점에서도 매우 유용하다. 대규모 학업성취도 연구의 채점 과정에서는 채점에 대한 중간 분석 결과에 따라서 채점자 재훈련을 포함하는데, E-HRM을 활용하면 모든 문항에 대해 재훈련을 진행하지 않고 채점자가 정상범위 외의 채점자 모수를 보인 특정 문항에 대해서만 재훈련하는 것이 가능하다. 그리고 PIRLS와 PISA의

경우 채점자간 일치도가 85% 이하일 경우 재채점을 실시한다(Mullis et al., 2004; OECD, 2017). 이 때에도 E-HRM을 통해 획득한 채점자-문항 단위의 채점 특성 정보를 활용한다면 모든 문항에 대해 재채점할 필요 없이 채점자가 정상범위 외의 채점자 모수를 보인 특정 문항만을 재채점하는 방식으로 채점에 소요되는 시간과 비용을 절감할 수 있다.

한편, 두 번째 연구 문제인 E-HRM이 모수를 얼마나 정확하고 효율적인가에 대해서는 '최소한의 채점건수를 확보한다면 정확하고 효율적이다'라는 조건부 답변만이 가능하다. 그리고 '최소한의 채점건수'에 대한 질문을 받는다면 모의실험 연구를 진행하였음에도 명확한 기준을 제시하기가 쉽지 않다. 우선 베이지안 통계학의 강점인 모수 추정치의 사후분포 신뢰구간 내 모수 참값 포함 여부에 대한 분석 결과는 채점자-문항당 채점건수가 '200건'일 때, 95% 신뢰구간에 참값을 포함하는 확률이 정상범위 채점자 비율 조건들을 포괄하여 95% 수준이었고, 채점자-문항당 채점건수가 '200건' 이하인 '100건', '50건', '25건'에서는 각각 92.7%, 91.3%, 89.9%로 95% 이하의 신뢰구간 내 모수 참값 포함 비율을 보였기 때문에 보수적으로 200건의 채점 건수가 필요하다고 결론내릴 수도 있다. HRM과의 분산성 모수 비교를 위해 제시된 <표 Ⅵ-8>에서 정상범위 외 모수 판별에서 제시된 민감도 역시 200건의 채점 건수를 지지한다. 그런데 한편으로 MSE와 관련된 오차의 크기를 본다면 정상범위 채점자 비율 '100%' 조건에서 RMSE 단위의 오차는 200건이 .103, 100건 .119, 50건 .146, 25건 .181로 ψ의 기준 범위 .5를 생각한다면 큰 차이라고 판단되지 않는다. 95% 신뢰구간에 참값을 포함하는 확률이 92.7%로 95%에 조금 못 미치고, 200건에 비해서 15% 정도의 오차 증가만 보인 100건 조건을 분석 가능한 채점건수에 포함하는 것은 충분히 가능하다고 판단된다. 그리고 관대하게는 적용 연구 결과에서 더 큰 사례수와 유사하게 안정적인 ESS와 MCSE를 보인 50건 채점건수도 추가적인 검증이나 분석상의 주의가 필요하겠지만 분석 가능한 채점건수로 고려해볼만하다고 판단한다.

채점자-문항 단위의 모수 추정에 활용되는 채점건수를 일정 건수 이상으로 확보하기가 어려운 규모로 피험자수와 문항수가 이미 결정되어 있다면 E-HRM을 다분 구성형 문항 채점 자료에 활용하기 위해서 채점 설계에서 구획 설계를 하는 방안을 고려해볼 수 있다. 피험자 수, 문항수가 결정되어 있고, 한 답안에 대한 복수 채점 인원 역시 제한되어 있다면 문항별로 채점할 답안의 수는 고정되어 있다. 고정된 채점 답안 수에서 한 채점자가 채점하는 채점 건수를 늘릴 수 있는 가장 간단한 방법은 모든 채점자들이 모든 문항을 채점하지 않고, 특정 채점자가 일부 문항만을 채점하게 하여 한 문항에 대한 채점자 수를 줄이는 것이다. 단, 이러한 채점자-문항 불완전 설계를 계획할 때에는 공통채점 문항을 통해 채점자 간 가교를 유지하는 선에서 전체 문항 중 일부 문항을 채점하도록 가교 설계를 포함해야한다는 것을 주의해야 한다.

그럼에도 불구하고 채점자-문항 단위의 모수 추정에 활용되는 채점건수를 충분히 확보할 수 없는 상황이라면 HRM과 E-HRM 중간 정도의 모수를 갖는 새로운 모형을 만드는 것도 한 방안이 될 수 있다. 예를 들어 PIRLS의 경우, 읽기성취도 검사라는 특성상 제시 글이 주어지고 이에 대한 복수의 문항이 출제되는데 제시 글을 중심으로 이 문항들이 하나의 구획이 된다. 이 구획에 포함된 문항들과 그에 딸린 채점 기준이 비슷한 특성을 공유한다면 구획(b) 단위로 채점자-문항 독립성 가정을 제한적으로 완화하여 채점자 모수 ϕ_{br} 또는 ψ_{br}를 추정하는 방식의 모형을 만들 수 있고, 이 모수의 추정에 활용되는 채점건수는 구획에 포함된 문항의 채점건수의 합이 된다. 그리고 PIRLS 2006은 '문학적 경험', '정보 획득 및 활용' 영역으로 문항이 구분되는데, Ⅲ장에서 살펴본 PIRLS 2006의 채점 기준에서 두 영역의 채점 기준이 다소 다른 성격을 가지고 있음을 확인할 수 있었다. 적용 연구의 E-HRM 분산성 모수를 문항별로 평균한 결과에서 '문학적 경험' 영역 문항들이 '정보 획득 및 활용' 영역 문항들에 비해서 분산성 모수가 큰 경향이 있었다. 이러한 상황이라면 구획 단위 모형화보다 좀 더 큰 규모에서 문항의 영역(a)별로 분산성 모

수 ψ_{ar}을 설정하는 것도 검사의 특성을 반영하면서도 모수 추정에 활용되는 채점건수를 늘리는 좋은 방안이라고 생각된다.

연구 과정 전반에서 문제를 일으킨 사항이자 본 연구의 한계점 중 하나는 수렴 평가에 활용한 $\hat{R}$ 통계량과 모형 적합도 평가에 활용한 DIC 통계량이다. Patz et al.(2002)에서 지적한 바와 같이 IRT 모형을 베이지안 추정으로 분석할 때 모수의 사후분포가 단봉-대칭 분포 형태를 크게 벗어나게 되는 사례는 본 연구에서도 빈번히 발생하였다. 그런데 본 연구에서 수렴 평가에 활용한 $\hat{R}$ 통계량과 모형 적합도 평가에 활용한 DIC 통계량은 사후분포가 정규분포 혹은 대칭형일 때를 가정한다(Gelman & Rubin, 1992; 황진수, 강성찬, 2011)는 점에서 비대칭형 사후분포에 대해 잘못된 결론을 내릴 수 있고(Gelman, Hwang, Vehtari, 2014), 때문에 본 연구 과정에서 관찰되는 모수의 사후분포와 평가 지표 간의 괴리를 자주 마주해야 했다. 이 중 $\hat{R}$의 경우, 기준을 엄격하게 적용하거나 MCSE, ESS 등을 함께 활용하거나 직접 연쇄의 자취 그래프나 자기상관 함수 그래프를 검토하는 방식으로 보완이 가능하였다. 그리고 DIC의 경우, 본 연구에서는 이를 대체 활용하지는 못하였지만, 최근 널리 활용되기 시작한 지표인 WAIC(Widely Applicable Information Criterion)이 대안이 될 수 있다. WAIC은 DIC와 달리 사후분포의 점 추정치를 활용하지 않고 사후분포의 밀도함수를 이용한다는 점에서 DIC에 비해 통계적으로 더 나은 특성을 가지고 있다(Watanabe & Sumio, 2010). 본 연구에서 활용한 JAGS에서는 WAIC 수치가 제공되지 않아 계산과정을 모형에 포함시키고 모니터링해야 한다는 점 때문에 활용하지 못했지만, 향후 HRM에 대한 연구에서는 WAIC을 활용하는 것을 유용할 것이라 판단된다.

마지막으로 후속 연구에 대한 제안으로 본 연구에서 조건이 한정적이었던 모의실험 연구를 조건을 확대하여 E-HRM의 적용을 위한 구체적인 분석 가이드라인을 제시할 수 있는 모의실험 연구가 필요하다. 본 연

구에서는 E-HRM을 적용하고자 하는 관심 상황이 적은 수의 범주를 가지는 구성형 문항으로 이루어진 대규모 검사 상황이었고, 그에 따라 제한된 모의실험을 설계하였다. 대규모 검사인 PIRLS에 대한 적용 연구에서 '채점자-문항당 채점건수'가 모수의 수렴이나 정확성에 가장 크게 영향을 미치는 요인으로 나왔기 때문에 모의실험 연구에서는 채점건수를 중심으로 설계 요인을 구성하였고, 다른 요인들은 고정하거나 무선적인 영향을 갖도록 모의실험을 설계하였다. 하지만 적용 연구에서 약하게 드러난 다른 요인들 예컨대 문항 수, 범주 수, 채점자간 가교 특성, 문항 변별도 모수의 역전 현상, 행렬표집설계 등도 고려할만한 모의실험 설계 요인이라고 판단된다. 그리고 본 연구에서 관심을 가진 검사 상황인 대규모 검사에서는 다분 구성형 문항만 사용하는 것이 아니라 이분 구성형 문항과 선다형 문항도 혼합하여 사용하는 것이 보편적이므로 좀 더 현실적인 상황에 가까운 연구를 위해 다분 구성형 문항, 이분 구성형 문항, 선다형 문항의 혼합 검사에 대한 연구도 향후 필요하다고 하겠다.

참 고 문 헌

교육부 (2017). **2017년 국가수준 학업성취도 평가 결과 발표**. 교육부 보도자료 (2017. 11. 30.).

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

강태훈, 심혜진 (2015). 다분문항반응이론에서 두 가지 사후기대추정법을 통한 IRT 능력모수 추정의 정확성 비교 연구. **교육방법연구, 27**(1).

구자옥, 김성숙, 임해미, 박혜영, 한정아(2015). **OECD 국제 학업성취도 평가 연구: PISA 2015 본검사 시행 보고서**. 한국교육과정평가원연구보고 RRE 2015-6-2.

김성숙 (2001). 채점의 변동요인 분석방법에 대한 고찰: 일반화가능도 이론과 다국면 라쉬모형의 적용과 재해석. **교육평가연구, 14**(1).

김수진, 이재봉, 박지현, 이문수(2015). **수학 · 과학 성취도 추이변화 국제비교 연구: TIMSS 2015 본검사 시행**. 한국교육과정평가원 연구보고 RRE 2015-11-2.

김창환, 이규민 (2008). 주관식 시험을 포함하고 있는 A국가고사의 채점자 효과에 대한 일반화가능도 연구. **교육과정평가연구. 11**(1).

남명호 (1995). 수행평가의 타당성 연구: 과학실기 평가, 실험보고서 평가, 컴퓨터시뮬레이션 평가의 비교. 고려대학교 박사학위 논문.

남현우 (1998). 수행평가의 측정학적인 문제들에 관한 고찰. **교육평가연구, 11**(2).

남현우, 남명호, 양길석, 김진하 (2009). 대규모시험의 행렬표집검사설계에 관한 고찰. **사회과학연구, 20**(4).

남현우. (2015). 비 최적조건의 행렬표집설계에서 베이지안 동등화의 최

소조건 탐색. **교육평가연구, 28**(4), 1177-1201.

민경석, & 양길석. (2013). 의학교육 학생평가의 객관성에 대한 쟁점. **의학교육논단, 15**(3), 105-111.

박도순 (2007). 교육평가-이해와 적용. 서울: 교육과학사.

박정 (2001). 문항반응이론을 활용한 수행형 평가문항 분석방법. **교육학연구, 39**(2).

박정, 박경미 (2004). 수행형 문항과 선다형 문항의 수학적 능력 추정 효율성 비교. **수학교육, 43**(2).

박종임, 노은희, 김유향, 전경희. (2017). **중등학교 수행평가 정책 실행 모니터링 및 개선 방안 연구**. 한국교육과정평가원연구보고 RRE 2017-8.

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

백순근 (1996). **초등학교의 새로운 평가제도에 따른 수행평가의 이론과 실제**. 국립교육평가원편. 대한교과서주식회사

백순근 (2000). **수행평가의 원리**. 서울: 교육과학사.

백순근, 임철일, 김혜숙, 유예림, 김미림, 이소라 (2010). 웹기반 수행형 ICT 리터러시 검사도구 개발 연구. **아시아교육연구, 11**(2).

법무부 (2005). **법무부령 제567호 - 사법시험법시행규칙 일부개정령**. 법무부 (2005.4.29).

이인호, 이상일, 김승현, 서민철, 성경희, 이광상, 임해미, 동효관, 배주경, 김성혜, 최병택, 권경필, 최원호, 이기영, 김희영(2016). **2016년 국가수준 학업성취도 평가 출제 연구**. 한국교육과정평가원연구보고 RRE 2016-8.

이현숙 (2008). 다국면 라쉬모형을 적용한 논술 채점 상황에서 채점 설계 및 채점자 특성이 채점의 정확성에 미치는 효과. **교육평가연**

구, 21(4).

지은림(1996). Many-Facet Rasch모형을 적용한 대입논술고사채점의 객관성연구, **교육평가연구, 9**(2).

지은림 (2008). 논술고사의 신뢰성에 영향을 미치는 채점자 특성 분석. **교육평가연구, 21**(2)

채선희 (1996). 논술시험 채점의 공정성과 효율성 확보방안 - 채점과정의 엄격성과 FACETS 모형에 의한 채점결과 분석. **교육평가연구, 9**(1).

허경철, 김왕근, 조덕주, 양미경, 차우규, 박선미 (2001). **초·중등학교 수행평가 현장 적용 평가 및 개선 방안 연구**. 교육인적자원부.

허명회, 한상태 (1996). 논술 채점의 신뢰도에 관한 연구. **응용통계연구, 9**(2), 45-53.

황지유, & 원미진. (2019). 일반화가능도 이론을 적용한 한국어 말하기 숙달도 평가의 신뢰도 및 타당도 분석. **국제어문, 81**, 321-349.

황진수, & 강성찬. (2011). 베이지안 로지스틱 회귀모형에서의 추론에 대한 연구. **응용통계연구, 24**(6), 1149-1160.

Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika, 43*(4), 561-573.

Baker, F. B., Kim, S. H., (2004). *Item Response Theory - Parameter Estimation Techniques*. 2nd edition. Marcel Dekker.

Bamber, D., & Van Santen, J. P. (1985). How many parameters can a model have and still be testable?. *Journal of Mathematical Psychology, 29*(4), 443-473.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6*(1).

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Brennan, R. L. (2001). Variability of statistics in generalizability theory. *In Generalizability theory* (pp. 179-213). Springer, New York, NY.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics, 7*(4), 434-455.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44(2)*, 131-155.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling, 59*(4), 471-492.

Casabianca, J. M., Junker, B. W., Nieto, R., & Bond, M. A. (2017). A hierarchical rater model for longitudinal data. *Multivariate behavioral research, 52*(5), 576-592.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. In *Handbook of Item Response Theory, Volume One* (pp. 477-494). Chapman and Hall/CRC.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling, 59*(4), 471-492.

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and*

resampling methods for social science. Sage Publications.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research, 37*(4), 423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*(1), 53-76.

DeCarlo, L. T. (2008). SStudies of a Latent-Class Signal-Detection Model for Constructed-Response Scoring. *ETS Research Report Series, 2008(2)*, i-55.

DeCarlo, L. T. (2010). Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs. *ETS Research Report Series, 2010(1)*, i-65.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333-356

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333-356.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*, 1-52.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings - Part I. *Psychological Test and Assessment Modeling, 59*(4), 443-452.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93 - 112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard, G., & Wind, S. A. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11(2)*, 195-208.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science, 7*(4), 457-472.

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics. *Statistical methods in medical research, 5*(4), 339-355.

Gelman, A., Carlin, J., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis crc press. *New York.*

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing, 24*(6), 997-1016.

Gilks, W. R., Best, N. G., & Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the*

Royal Statistical Society: Series C (Applied Statistics), 44(4), 455-472.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied psychological measurement, 26*(1), 3-24.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of research in education, 24*(1), 393-446.

Hombo, C. M., & Donoghue, J. R. (2000, June). *Some properties of the distribution of an IRT measure of item fit.* Paper presented at the Annual Meeting of the Psychometric Society, Vancouver, British Columbia, Canada.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

Hoskens, M., & Wilson, M. (2001). Real time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*(2), 121-145.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*(3).

Kane, M. & Mitchell, R. (1996). *Implementing Performance Assessment.* Lawrence Erlbaum Associates. Mahwah, New Jersey.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician, 52*(2), 93-100.

Kim, S. H. (2013). *Dealing with Sparse Rater Scoring of Constructed Responses within a Framework of a Latent Class Signal Detection Model* (Doctoral dissertation). Teachers College, Columbia University, New York.

Kim, Y. K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doctoral dissertation). Teachers College, Columbia University, New York.

Kolen, M. J.(2006). Scaling and norming. In Robert L. Brennan(Eds.), *Educational Measurement*. Preager.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48(4)*, 399-418.

Lee, W. C., & Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48.

Linacre, J. M. (1989). *Many-faceted Rasch measurement* (Doctoral dissertation, The University of Chicago).

Linacre, J. M. (2003). The hierarchical rater model from a Rasch perspective. *Rasch Measurement Transactions*, 17.

Linacre, J. M. (2008). *A user's guide to FACETS.*

Linn, R. L. (1994). Performance assessment: policy promises and

technical measurement standards. *Educational Researchers 23*(9).

Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine, 31*(18), 2010-2026.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. E*valuation & the Health Professions, 13*(4), 425-444.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*.

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

Martin, M. O., Mullis, I. V., & Kennedy, A. M. (2007). Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 Technical Report. *International Association for the Evaluation of Educational Achievement.*

McNamara, T. F. (1996). *Measuring second language performance.* Addison Wesley Longman.

Mullis, I. V., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2004). PIRLS 2006 *Assessment Framework and Specifications: Progress in International Reading Literacy Study. TIMSS & PIRLS, International Study* Center, Lynch School of Education, Manressa House, Boston College, 140 Commonwealth Street, Chestnut Hill, MA 02467.

Mullis, I. V., Martin, M. O. (Eds). (2013) *TIMSS 2015 Assessment*

Frameworks. TIMSS & PIRLS International Study Center.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1).

Muraki, E. (1993). *Variations of polytomous item response models: Raters' effect model, DIF model, and trend model*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46(4)*, 371-389.

Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement, 56*(3), 547-581.

OECD (2017), *PISA 2015 Technical Report*, OECD, Paris.

Patterson, B. F. (2013). *Examining the impact of examinee-selected constructed response items in the context of a hierarchical rater signal detection model* (Doctoral dissertation, Columbia University).

Patterson, B. F., Wind, S. A., & Engelhard, G. (2017). Incorporating criterion ratings into model-based rater monitoring procedures using latent-class signal detection theory. *Applied psychological measurement, 41*(6), 472-491.

Patz, R. J. (1996). *Markov chain Monte Carlo methods for item response theory models with applications for NAEP* (Doctoral dissertation, Carnegie Mellon University).

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of educational and behavioral statistics, 24*(4), 342-366.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4).

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics, 9*(3), 523-539.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*(2).

Raju, N.S., Price, L.R., Oshima, T.C., & Nering, M.L. (2006). Standardized Conditional SEM: A Case for Conditional Reliability. Applied Psychological Measurement, 30(X), 1-12.

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101-138.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology), 64*(4), 583-639.

Stan Development Team (2019). *STAN reference manual.* Version 2.19.

Sturtz, S., Ligges, U., & Gelman, A. E. (2005). R2WinBUGS: a package for running WinBUGS from R.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In *Essays on item response theory* (pp. 89-108). Springer New York.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109-128.

Watanabe, Sumio (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". *Journal of Machine Learning Research. 11*: 3571-3594.

Wei, R. C., Pecheone, R. L., & Wilczak, K. L. (2014). *Performance assessment 2.0: Lessons from large-scale policy and practice.* Stanford Center for Assessment, Learning, and Equity.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*(3), 283-306.

Yen, N. T. Q. (2016). Rater consistency in rating L2 learners' writing task. *VNU Journal of Foreign Studies*, 32(2).

부록표

<부록표 1> 채점자별 Deviance 통계량

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
덴마크	1	230	42.0	56.8
	2	966	143.5	93.9
	3	257	.3	7.6
	4	235	33.3	14.5
	5	1,129	108.8	64.3
	6	585	8.2	61.9
	7	622	137.3	76.1
	8	226	3.0	18.5
	9	322	107.6	72.4
	10	579	119.4	77.3
	11	1,131	114.1	121.4
	12	292	16.0	29.2
	13	267	53.6	47.3
	14	524	110.0	82.9
	15	322	3.8	34.5
	16	392	.2	13.3
조지아	1	577	454.1	303.8
	2	1,001	1160.1	1013.7
	3	666	681.1	509.0
	4	893	660.9	407.9
	5	619	498.9	378.6
	6	703	668.6	479.4
	7	1,003	849.1	740.8
	8	596	485.5	404.4
	9	718	587.1	533.6
	10	1,079	1100.6	1058.6
	11	763	802.8	710.7
	12	328	206.4	215.3
홍콩	1	1,279	434.4	381.9
	2	475	120.9	106.7
	3	494	195.3	151.4
	4	965	314.1	313.5
	5	1,294	474.3	420.4
	6	1,020	325.7	272.5
	7	359	102.0	70.3
	8	275	91.1	83.4
	9	499	160.6	128.7
	10	1,133	394.0	336.4

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
	11	236	.3	12.2
	12	563	.7	12.1
헝가리	1	1,799	151.4	121.5
	2	1,797	289.0	205.9
	3	1,840	153.0	137.1
	4	1,888	159.5	181.7
	5	1,839	107.0	96.2
	6	1,808	267.2	225.9
아이슬란드	1	543	1.6	35.3
	2	1,115	386.3	317.8
	3	975	339.9	214.1
	4	857	258.9	208.6
	5	1,189	261.4	236.2
	6	474	196.1	165.3
	7	577	190.0	139.9
	8	771	207.8	154.2
	9	689	226.1	207.0
	10	1,049	276.1	206.7
	11	793	159.3	124.2
라트비아	1	2,016	997.7	847.0
	2	1,337	642.0	518.0
	3	1,088	537.0	459.6
	4	387	193.7	186.1
	5	688	544.4	489.0
	6	1,860	793.2	704.1
	7	875	213.0	263.5
	8	240	425.4	402.0
	9	1,099	377.1	309.8
	10	1,951	1240.1	912.1
리투아니아	1	5,591	1354.4	1202.0
	2	5,595	1473.7	1342.1
룩셈부르크	1	1,602	622.0	485.9
	2	416	98.7	60.8
	3	1,498	591.1	616.4
	4	525	231.4	198.1
	5	1,463	629.3	520.3
	6	2,081	740.3	627.9
	7	2,233	752.0	613.8
	8	1,120	475.7	426.0
네덜란드	1	551	20.3	57.5

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
	2	1,198	46.0	64.9
	3	2,327	182.1	153.1
	4	581	30.9	24.3
	5	269	.6	8.8
	6	1,519	132.0	70.4
	7	1,912	103.2	60.0
	8	1,624	69.5	61.9
뉴질랜드	1	1,677	491.1	487.8
	2	1,635	452.9	468.6
	3	1,627	728.5	667.0
	4	1,723	827.4	826.3
	5	1,800	926.0	758.5
	6	1,507	603.7	441.3
	7	1,592	427.8	320.4
	8	1,802	956.4	800.6
	9	360	128.2	97.4
	10	334	54.4	66.5
	11	211	100.5	93.6
노르웨이	1	723	474.4	414.7
	2	681	553.2	485.3
	3	675	645.7	550.8
	4	858	827.6	765.7
	5	730	534.7	437.1
	6	543	432.9	305.5
	7	308	256.3	213.4
	8	640	672.2	636.9
	9	555	431.3	352.3
	10	254	103.4	88.6
	11	832	574.5	462.7
	12	636	398.0	326.4
	13	519	618.4	581.2
	14	475	486.4	364.5
루마니아	1	1,579	98.6	113.7
	2	1,592	169.6	137.6
	3	867	85.0	91.2
	4	952	90.9	75.2
	5	1,705	161.6	158.3
	6	1,705	164.4	140.1
	7	885	105.8	89.2
	8	971	52.7	48.3

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
러시아	1	6,017	647.0	583.4
	2	6,017	461.0	427.6
슬로바키아	1	780	191.4	147.0
	2	1,111	402.5	327.9
	3	1,004	422.9	352.6
	4	890	197.4	134.2
	5	903	42.5	90.4
	6	1,081	315.9	331.9
	7	802	234.9	209.1
	8	1,072	298.0	254.7
	9	887	240.2	214.4
	10	966	241.3	184.1
	11	407	182.6	163.6
슬로베니아	1	941	118.8	135.6
	2	1,413	233.9	199.3
	3	1,674	426.1	327.6
	4	1,387	74.8	61.1
	5	1,237	315.2	253.8
	6	1,180	265.5	235.4
	7	932	176.9	156.9
	8	1,872	216.0	188.4
	9	379	20.6	40.6
스웨덴	1	872	341.0	318.5
	2	848	601.6	526.3
	3	865	470.4	393.6
	4	858	410.4	312.2
	5	834	352.3	251.7
	6	861	442.8	366.4
	7	841	472.1	429.7
	8	836	369.8	272.7
	9	837	288.4	189.2
	10	839	446.8	389.9
	11	796	269.0	251.1
	12	799	535.9	512.9
	13	819	335.6	270.1
	14	788	421.6	294.2
마케도니아	1	648	60.6	58.7
	2	716	266.7	232.4
	3	657	208.5	174.1
	4	480	112.2	124.0

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
	5	550	207.6	171.1
	6	665	159.2	149.6
	7	575	291.3	210.9
	8	514	329.4	316.8
	9	292	109.2	98.2
	10	280	104.1	82.8
	11	283	47.9	116.8
	12	280	197.3	135.3
	13	319	14.4	12.7
	14	309	14.6	24.3
	15	393	29.8	33.6
	16	354	29.1	17.0
	17	389	.1	4.7
	18	323	.1	5.4
	19	300	11.7	12.3
	20	293	2.9	5.5
	21	325	38.8	35.9
	22	392	32.2	8.9
미국	1	1,030	427.8	300.0
	2	2,035	928.5	834.3
	3	2,766	1047.1	1082.5
	4	556	233.6	208.9
	5	2,059	955.7	952.3
	6	2,380	1134.3	965.3
	7	2,261	990.7	972.2
영국	1	847	110.6	104.2
	2	340	2.2	19.4
	3	2,601	328.2	276.4
	4	1,954	359.4	371.4
	5	1,904	351.5	279.6
	6	2,152	169.6	146.0
	7	689	108.2	91.7
스코틀랜드	1	1,182	274.0	292.9
	2	1,793	233.7	221.3
	3	2,763	686.2	589.5
	4	1,491	274.5	238.1
	5	1,140	51.0	72.4
	6	409	84.1	56.4
벨기에 (네덜란드어권)	1	2,042	1209.5	1104.3
	2	1,930	1251.1	1207.3

국가	채점자	채점건수	Deviance	
			HRM	E-HRM
	3	1,769	862.5	837.0
	4	1,667	757.5	512.2
	5	2,042	1302.0	1071.5
	6	2,037	1208.2	880.1
벨기에(프랑스어권)	1	1,300	341.3	261.1
	2	1,120	149.8	168.8
	3	1,149	208.2	232.0
	4	1,252	275.5	227.0
	5	1,346	270.8	258.9
	6	1,311	232.2	187.9
	7	1,253	348.0	248.9
	8	1,294	168.2	157.9
아이슬란드(5학년)	1	372	106.4	100.4
	2	1,122	455.0	337.0
	3	565	60.8	59.9
	4	768	315.9	239.3
	5	771	191.1	199.6
	6	1,337	220.7	189.2
	7	383	149.7	100.0
	8	551	82.5	61.9
	9	1,315	369.5	379.3
	10	1,178	204.6	205.5
	11	1,457	357.7	268.2
노르웨이(5학년)	1	385	347.4	277.0
	2	536	461.0	410.1
	3	333	454.6	398.3
	4	489	503.5	425.9
	5	707	530.0	447.6
	6	570	511.5	348.1
	7	400	406.0	319.7
	8	739	937.7	827.9
	9	321	173.8	136.6
	10	509	278.0	266.2
	11	797	588.5	531.8
	12	649	404.8	310.8
	13	416	557.6	482.0

<부록표 2> 아이슬란드 문항-채점자 채점건수 배정 설계

문항	채점자									
	1	2	3	4	5	6	7	8	9	10
1	24	34	44	31	50	24	10	29	32	24
2	25	35	46	32	50	24	12	28	33	24
3	22	34	43	31	48	24	11	24	31	23
4	12	55	41	44	31	12	16	21	35	42
5	12	51	39	40	29	11	15	20	30	40
6	21	39	38	42	35	10	35	20	31	33
7	22	39	40	42	37	10	35	21	32	33
8	19	35	36	37	35	9	29	18	30	30
9	19	22	33	45	28	27	14	17	19	37
10	18	19	28	40	25	25	11	14	18	32
11	15	13	24	30	22	20	11	8	11	26
12	43	50	13	15	51	14	15	58	17	31
13	42	53	14	15	50	14	14	58	15	34
14	41	52	14	15	50	13	13	58	15	34
15	36	48	14	15	46	12	12	53	13	32
16	16	49	28	16	56	14	19	37	15	43
17	16	39	26	15	53	11	18	32	13	42
18	14	32	18	14	44	12	16	27	12	29
19	-	29	45	23	52	20	8	24	23	49
20	-	29	42	22	48	17	6	24	22	49
21	-	31	45	24	52	20	8	24	25	51
22	13	44	42	34	35	13	48	15	31	44
23	15	45	41	34	36	15	49	15	32	43
24	14	45	41	34	36	14	49	15	32	43
25	24	33	46	32	51	24	12	28	32	23
26	13	55	41	41	30	12	13	21	34	41
27	20	22	34	47	28	27	15	17	18	40
28	15	42	22	16	50	14	19	31	11	37
29	12	41	37	31	31	12	44	14	27	40

<부록표 3> 벨기에(프랑스어권) 문항-채점자 채점건수 배정 설계

문항	채점자							
	1	2	3	4	5	6	7	8
1	46	36	41	56	51	41	42	56
2	47	35	39	55	52	40	39	55
3	43	34	39	55	48	40	39	55
4	41	46	42	47	41	46	42	47
5	38	46	39	45	38	46	39	45
6	53	34	41	42	53	41	48	42
7	58	35	44	45	58	42	52	45
8	51	31	39	36	51	36	46	36
9	47	28	31	47	48	37	39	47
10	39	28	29	45	42	35	35	45
11	36	23	27	39	36	29	32	39
12	52	46	54	46	57	57	54	52
13	52	47	55	46	57	58	55	52
14	51	44	52	47	56	54	52	52
15	46	37	46	39	49	45	47	44
16	39	61	36	38	40	61	35	38
17	37	45	32	38	38	45	31	38
18	34	30	27	27	35	30	26	27
19	36	45	26	42	37	53	33	42
20	39	42	29	39	40	51	38	39
21	39	45	31	42	40	54	38	42
22	55	41	56	41	54	53	63	47
23	54	40	53	41	53	55	64	46
24	53	38	52	40	52	53	62	44
25	47	32	40	56	52	37	40	56
26	39	43	39	44	39	43	39	44
27	47	30	35	51	48	39	43	51
28	33	47	35	28	34	48	34	28
29	48	31	40	35	47	42	46	40

<부록표 4> 덴마크 문항-채점자 채점건수 배정 설계

문항	채점자							
	1	2	3	4	5	6	7	8
1	16	16	-	-	5	30	15	13
2	17	19	-	-	12	29	15	13
3	17	19	-	-	11	30	15	13
4	-	56	-	-	77	32	13	21
5	-	45	-	-	76	30	13	21
6	1	36	13	-	53	2	26	20
7	1	37	13	-	57	2	27	21
8	-	33	13	-	57	2	25	19
9	28	3	-	-	31	29	38	-
10	29	2	-	-	31	29	38	-
11	28	3	-	-	29	27	34	-
12	1	39	-	31	27	-	30	13
13	1	40	-	28	27	-	15	13
14	1	40	-	31	26	-	30	13
15	1	39	-	29	23	-	27	12
16	1	61	24	31	53	-	28	-
17	1	57	25	28	45	-	24	-
18	-	54	23	27	45	-	22	-
19	12	29	25	-	33	52	20	-
20	11	28	24	-	30	53	20	-
21	12	31	25	-	34	55	21	-
22	1	35	13	-	42	23	9	-
23	1	37	13	-	48	24	9	-
24	1	37	12	-	48	24	9	-
25	17	19	-	-	12	30	15	13
26	-	54	-	-	75	32	12	21
27	30	3	-	-	33	29	39	-
28	1	59	21	30	49	-	25	-
29	1	35	13	-	40	21	8	-

문항	채점자							
	9	10	11	12	13	14	15	16
1	1	28	14	1	-	16	17	1
2	1	28	14	1	-	17	17	1
3	1	29	13	1	-	17	17	1
4	23	31	51	2	-	31	18	1
5	23	28	51	2	-	30	18	1
6	22	6	29	5	22	-	25	2
7	24	7	30	5	22	-	26	2
8	24	6	29	5	20	-	25	2
9	-	31	29	3	-	41	25	1
10	-	31	29	3	-	42	24	-
11	-	29	26	3	-	40	22	1
12	1	16	44	33	16	-	3	1
13	1	16	45	31	14	-	3	1
14	1	16	45	32	16	-	3	2
15	1	14	44	29	15	-	3	2
16	1	1	64	26	16	31	4	54
17	2	-	56	23	14	29	4	52
18	2	-	56	21	13	30	3	45
19	23	32	36	1	-	25	1	53
20	24	31	33	-	-	26	-	52
21	26	32	38	-	-	27	1	54
22	24	19	48	9	20	-	-	3
23	25	20	53	9	22	-	-	3
24	25	19	53	9	21	-	-	3
25	1	29	14	1	-	17	17	1
26	23	30	50	2	-	30	18	1
27	-	31	31	3	-	44	24	1
28	2	1	60	23	15	31	4	48
29	21	18	46	9	21	-	-	3

<부록표 5> 모의실험 채점자 모수 참값

채점 양호도	번호	ϕ_i	ψ_{ir}									
			문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
정상	1	-.013	.069	.070	.072	.073	.074	.073	.070	.072	.076	.075
	2	-.018	.075	.076	.077	.077	.074	.071	.071	.073	.071	.070
	3	-.014	.075	.074	.073	.074	.074	.076	.074	.076	.073	.077
	4	-.012	.070	.074	.074	.072	.076	.073	.074	.075	.078	.075
	5	-.015	.076	.074	.075	.076	.073	.078	.077	.077	.072	.077
	6	.000	.065	.079	.063	.067	.069	.063	.064	.077	.072	.070
	7	-.005	.078	.075	.083	.073	.077	.072	.070	.075	.074	.071
	8	.025	.058	.063	.082	.068	.096	.066	.064	.069	.064	.063
	9	.001	.068	.075	.074	.073	.074	.077	.074	.073	.072	.102
	10	-.023	.057	.055	.238	.059	.066	.083	.061	.161	.159	.106
	11	-.034	.062	.063	.246	.129	.078	.274	.064	.151	.065	.123
	12	.020	.315	.122	.064	.065	.213	.097	.068	.153	.259	.083
	13	.020	.058	.320	.146	.059	.059	.110	.204	.063	.058	.059
	14	-.033	.076	.084	.065	.108	.069	.073	.331	.067	.075	.070
	15	.001	.058	.062	.056	.058	.059	.332	.331	.059	.327	.058
	16	.035	.285	.185	.120	.056	.054	.059	.070	.333	.058	.094
	17	.035	.059	.243	.336	.066	.068	.167	.066	.076	.058	.064
	18	.002	.059	.061	.073	.336	.341	.060	.059	.058	.058	.316
	19	-.011	.341	.073	.077	.072	.074	.074	.073	.077	.076	.079
	20	.064	.069	.331	.317	.057	.317	.312	.351	.301	.304	.324
	21	.006	.357	.226	.085	.060	.060	.062	.261	.068	.061	.103
	22	-.014	.063	.063	.064	.061	.066	.358	.064	.072	.064	.065
	23	.035	.061	.067	.059	.059	.088	.062	.352	.362	.059	.151
	24	-.012	.098	.313	.174	.057	.060	.056	.055	.362	.057	.356
	25	-.066	.067	.065	.342	.363	.059	.085	.063	.201	.349	.064
	26	-.188	.067	.065	.264	.147	.201	.175	.066	.095	.071	.371
	27	.029	.084	.271	.071	.058	.228	.377	.092	.344	.062	.064
	28	-.002	.062	.081	.380	.070	.067	.072	.070	.070	.149	.155
	29	.010	.066	.068	.069	.374	.357	.380	.297	.297	.295	.060

채점 양호도	번호	ϕ_i	ψ_{ir}									
			문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
	30	.078	.064	.065	.259	.348	.304	.385	.069	.120	.152	.310
	31	.060	.068	.070	.070	.336	.377	.392	.067	.067	.107	.129
	32	.045	.126	.099	.061	.329	.368	.393	.108	.363	.060	.378
	33	-.007	.367	.339	.147	.242	.061	.117	.326	.395	.059	.062
	34	.063	.063	.063	.152	.327	.333	.315	.358	.396	.062	.093
	35	.029	.063	.235	.397	.301	.060	.330	.086	.062	.060	.062
	36	.045	.082	.339	.382	.061	.062	.065	.302	.404	.061	.324
	37	-.057	.333	.066	.404	.069	.092	.193	.188	.068	.110	.364
	38	.190	.066	.082	.069	.070	.352	.377	.408	.158	.066	.091
	39	.067	.326	.126	.384	.072	.345	.097	.351	.410	.104	.064
	40	.045	.359	.365	.412	.393	.344	.395	.372	.367	.342	.068
	41	.013	.077	.123	.141	.064	.412	.296	.065	.237	.062	.064
	42	-.048	.071	.325	.140	.069	.070	.414	.065	.401	.072	.074
	43	.091	.416	.353	.065	.063	.070	.093	.119	.412	.311	.064
	44	-.046	.069	.070	.115	.190	.420	.251	.387	.066	.068	.346
	45	-.010	.358	.122	.242	.229	.420	.184	.375	.120	.362	.150
	46	-.003	.344	.398	.425	.060	.251	.336	.061	.168	.059	.098
	47	.033	.415	.356	.163	.063	.427	.067	.062	.063	.063	.063
	48	-.010	.061	.060	.061	.180	.119	.094	.061	.429	.381	.167
	49	-.011	.078	.088	.060	.059	.100	.085	.338	.431	.068	.083
	50	.003	.432	.068	.064	.411	.066	.066	.068	.070	.327	.429
	51	.015	.057	.387	.435	.144	.373	.073	.060	.400	.059	.063
	52	.010	.370	.062	.072	.084	.273	.436	.063	.108	.061	.065
	53	-.011	.331	.404	.073	.087	.408	.430	.059	.437	.061	.086
	54	.106	.067	.066	.281	.421	.345	.439	.071	.370	.067	.327
	55	.006	.140	.240	.441	.293	.317	.154	.064	.147	.059	.433
	56	.013	.407	.443	.089	.373	.441	.432	.394	.069	.068	.060
	57	-.074	.430	.363	.139	.444	.074	.125	.067	.068	.348	.280
	58	.044	.371	.134	.261	.444	.353	.410	.061	.156	.345	.402
	59	.122	.065	.359	.379	.070	.446	.075	.076	.162	.332	.411
	60	-.004	.288	.451	.278	.240	.336	.062	.060	.093	.064	.370

채점 양호도	번호	ϕ_i	ψ_{ir}									
			문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
	61	.046	.073	.457	.075	.071	.084	.080	.085	.193	.219	.074
	62	.114	.077	.457	.407	.073	.168	.300	.157	.132	.152	.290
	63	.007	.061	.061	.061	.457	.060	.081	.330	.070	.064	.389
	64	.087	.129	.067	.123	.070	.461	.437	.457	.348	.069	.393
	65	.035	.267	.065	.063	.066	.074	.113	.064	.469	.388	.417
	66	-.015	.069	.356	.068	.056	.361	.473	.355	.184	.063	.063
	67	-.007	.324	.381	.415	.473	.339	.383	.165	.420	.315	.440
	68	.104	.145	.394	.475	.092	.447	.334	.066	.318	.067	.415
	69	.018	.477	.105	.327	.073	.076	.090	.085	.083	.070	.448
	70	.092	.321	.399	.369	.440	.479	.416	.372	.378	.336	.350
	71	.084	.067	.075	.329	.419	.420	.355	.205	.485	.180	.239
	72	.048	.094	.340	.449	.362	.485	.068	.069	.098	.066	.365
	73	.006	.061	.207	.486	.085	.067	.121	.396	.386	.076	.069
	74	-.035	.082	.281	.341	.407	.491	.427	.066	.273	.067	.069
	75	.131	.070	.188	.397	.071	.069	.403	.076	.495	.077	.281
	76	.066	.066	.146	.356	.129	.346	.402	.314	.065	.499	.445
	77	-.006	.059	.334	.074	.296	.500	.095	.080	.063	.474	.064
	78	.037	.500	.420	.320	.067	.059	.067	.066	.430	.463	.068
엄격	1	-.517	.062	.081	.380	.070	.067	.072	.070	.070	.149	.155
	2	-.475	.061	.067	.059	.059	.088	.062	.352	.362	.059	.151
	3	-.582	.129	.067	.123	.070	.461	.437	.457	.348	.069	.393
	4	-.445	.326	.126	.384	.072	.345	.097	.351	.410	.104	.064
	5	-.410	.059	.061	.073	.336	.341	.060	.059	.058	.058	.316
	6	-.576	.098	.313	.174	.057	.060	.056	.055	.362	.057	.356
	7	-.455	.063	.063	.152	.327	.333	.315	.358	.396	.062	.093
	8	-.542	.069	.356	.068	.056	.361	.473	.355	.184	.063	.063
	9	-.405	.063	.235	.397	.301	.060	.330	.086	.062	.060	.062
	10	-.403	.407	.443	.089	.373	.441	.432	.394	.069	.068	.060
	11	-.468	.371	.134	.261	.444	.353	.410	.061	.156	.345	.402
	12	-.518	.066	.082	.069	.070	.352	.377	.408	.158	.066	.091
	13	-.497	.359	.365	.412	.393	.344	.395	.372	.367	.342	.068

채점 양호도	번호	ϕ_i	ψ_{ir} 문항 1	문항 2	문항 3	문항 4	문항 5	문항 6	문항 7	문항 8	문항 9	문항 10
	14	-.451	.067	.065	.342	.363	.059	.085	.063	.201	.349	.064
	15	-.481	.066	.068	.069	.374	.357	.380	.297	.297	.295	.060
	16	-.418	.357	.226	.085	.060	.060	.062	.261	.068	.061	.103
	17	-.415	.069	.070	.115	.190	.420	.251	.387	.066	.068	.346
	18	-.594	.344	.398	.425	.060	.251	.336	.061	.168	.059	.098
	19	-.537	.333	.066	.404	.069	.092	.193	.188	.068	.110	.364
	20	-.504	.367	.339	.147	.242	.061	.117	.326	.395	.059	.062
저신뢰	1	.096	.073	.332	.072	.476	.298	.234	.384	.774	.400	.095
	2	.134	.072	.428	.356	.096	.082	.781	.456	.539	.141	.069
	3	.007	.370	.663	.577	.494	.500	.814	.109	.550	.093	.369
	4	-.017	.406	.064	.088	.101	.140	.073	.072	.831	.067	.344
	5	.004	.065	.646	.065	.068	.450	.845	.594	.078	.138	.065
	6	.126	.070	.471	.092	.371	.533	.078	.078	.950	.068	.503
	7	.055	.063	.782	.495	.072	.436	.082	.101	.989	.506	.643
	8	-.178	.322	.089	.433	.074	.387	.731	.295	.958	.110	.499
	9	.119	.070	.152	.086	.110	.902	.303	.964	.106	.071	.111
	10	-.075	.116	.368	.304	.484	.422	.861	.570	.958	.250	.387
엄격-저신뢰	1	-.469	.073	.332	.072	.476	.298	.234	.384	.774	.400	.095
	2	-.543	.072	.428	.356	.096	.082	.781	.456	.539	.141	.069
	3	-.470	.370	.663	.577	.494	.500	.814	.109	.550	.093	.369
	4	-.531	.406	.064	.088	.101	.140	.073	.072	.831	.067	.344
	5	-.463	.065	.646	.065	.068	.450	.845	.594	.078	.138	.065
	6	-.444	.070	.471	.092	.371	.533	.078	.078	.950	.068	.503
	7	-.588	.063	.782	.495	.072	.436	.082	.101	.989	.506	.643
	8	-.550	.322	.089	.433	.074	.387	.731	.295	.958	.110	.499
	9	-.568	.070	.152	.086	.110	.902	.303	.964	.106	.071	.111
	10	-.462	.116	.368	.304	.484	.422	.861	.570	.958	.250	.387

※ 음영 표시된 값은 채점의 양호도가 낮은 모수

Extension and Application of Hierarchical Rater Model Using Bayesian Estimation

A Dissertation for the Degree

of

Doctor of Philosophy in Education

By

Minho Park

Major Advisor : Hyun-Jeong Park, Ph.D.

Department of Education

The Graduate School

Seoul National University

2021

Abstract

Extension and Application of Hierarchical Rater Model Using Bayesian Estimation

Minho Park
Department of Education
The Graduate School
Seoul National University

The purpose of the study was to propose an extended hierarchical rater model(HRM) for identifying differential rater effects on rating construction items of a large-scale test and evaluating its accuracy and efficiency. Construction items, a simple performance evaluation form used widely in recent years, are scored by the rater, so rating reliability is crucial to their effectiveness. To ensure rating reliability, the rubric must be clearly made during the test development process so that they are not interpreted differently depending on the raters. Since the raters must be professional, rigorous, and consistent, there must be sufficient training prior to rating process. Despite the efforts of these test development and rating process, it is difficult to completely control the rater effects. Rater effect models including HRM based on item response theory have emerged to eliminate rater effects and accurately evaluate examinee performance.

Most item response theory models, from the traditional models to HRM, assumed the influence of rater effect was constant regardless of the items. It is however difficult to keep the rater-item

independence assumption in reality, and it is frequently violated in generalizability studies that use ANOVA. In this context, studies that removed the contraint due to the assumption of rater-item independence were performed on the Hierarchical Rater Signal Detection Model(HRM-SDT), a method commonly used to analysis rating data of essay-type items. However because of their complexity, those studies were only used in a few essay-type items, not in tests comprising a large number of construction items.

Thus, in this study, to explore the reparameterize model by mitigating the rater-item independence constraints of HRM, which has advantages in regards to parsimony, as the consistency of the rater will be modeled as a random effect. Then application study and simulation study were examined to verify the accuracy and efficiency of the model. Instead of general research that develops, verifies, and applies models, this study leverages an existing model and first applies all possible models to actual data, following a mitigation of the rater-item independence constraints, followed by a procedure to verify accuracy and efficiency.

A total of four models were analyzed using the rater-item independence constraints mitigation of rater severity parameters and rater variability parameters included in HRM. By applying these models to the PIRLS 2006 data, an international comparison large-scale performance test, and using the Bayesian-based MCMC method, the applicability of model was explored. In addition, the simulation study was conducted to verify the accuracy and efficiency of the proposed model.

Based on the results of the application study, E-HRM(Extended HRM), assuming the rater's consistency varies across items and

reducing the rater-item independence constraints only for the variance parameters, showed the best results of convergence in MCMC chains and interpretability. The analysis of E-HRM revealed that rater effects could be inconsistent in certain items even though they seem to be consistent on average in overall items. The analysis also found that further meaningful information of differential rater effects on each item can be acquired through E-HRM. In addition, E-HRM was better for analyzing data consist of items with heterogeneous rating distributions compared to HRM. However, even with MCMC, E-HRM could not adequately converge without enough number of rating case per item by rater.

In the simulation study, 12 conditions with two factors (normal range rater's rate, number of raters for each rater item) were constructed to evaluate the accuracy and efficiency of E-HRM. The simulation study revealed the following results. First, the study showed that the parameter estimation was less accurate and efficient when there were fewer number of rating per rater-item, as well as larger errors. The proportion of rater severity parameters having true values within the 95% confidence interval showed significantly high accuracy, except for one simulation design condition. The proportion of rater variability parameters having true values within the 95% confidence interval is calculated under the conditions of the normal range rater rate '100%', '70%', and '40%', for number of rating per rater-item '200 cases' .953, .948, and .944 respectively, which is about 95%. Nevertheless, as the number of rating per rater-item decreased, the corresponding rates for '100%', '70%' and '40%' decreased, at .881, .894, and .922, respectively.

Second, regarding the discriminate abnormal rating, E-HRM

demonstrated high sensitivity on rater variability parameters. The sensitivity and specificity of E-HRM by the number of rating per rater-item were analyzed. The results showed that with '200 cases', sensitivity and specificity were .953 and .988, respectively, and different properties between normal raters and abnormal raters could be discerned with a high degree of probability. With '100 cases', the sensitivity and specificity were .903 and .985, indicating more than 90% accuracy in discrimination. However the sensitivity decreased as the number of rating per rater-item decreased. There was slightly lower sensitivity for '50 cases' and '25 cases' at .864 and .794, respectively.

Last E-HRM demonstrated its capability to discriminate rater effects clearly and to appropriately compensate for the effects of rater severity or rater inconsistency on the accuracy and efficiency of examinee and item parameters. Further, whereas HRM showed poor convergence rate and accuracy due to the multimodal posterior distributions of the rater parameter estimates, E-HRM showed stable estimation and better recovery of parameter than HRM.

The study is not without limitation. The evaluation of E-HRM using Bayesian estimation was conducted only for an actual dataset and covered limited simulation designs. However, the data from 24 countries of PIRLS 2006 with varying rating designs were analyzed with E-HRM to confirm its performance and interpretability. Moreover, through the simulation study, the accuracy and efficiency of E-HRM were evaluated under specific conditions. Since the usefulness of E-HRM has been confirmed, detailed information of rater effects by each rater-item offers a benefit to raters in terms of improving the professionality of raters and significantly reduce the

cost, effort, and time of re-training and re-scoring in large-scale tests.

* Keywords: hierarchical rater model, bayesian estimation, MCMC, Gibbs sampling, rater effect

* Student Number: 2016-30397