Rasch 모형을 통한 <독서와 문법> 평가 문항 분석*

이 정 애** (전북대) · **심 재 우***** (전북대)

< 목 차 >

국문초록

본 연구는 현실적 차원에서 여전히 중요성을 갖는 선택형 문항들을 Rasch 모형을 통해 분석함으로써 교사가 출제하는 평가 문항의 신뢰도, 학생들의 능력에 따른 평가 적절성 여부, 평가 문항에 대한 정보와 판단의 타당성 등을 기술하고자 하였다. 우선 Rasch 모형의 분석을 위해 J 시 인문계 고등학생 총 179명이 치른 <독서와 문법>의 평가 문항을 대상으로 Winsteps 3.68 소프트 프로그램을 사용하였으며 결과는 다음과 같다. 첫째, 이 평가지는 학생의 능력 신뢰도가 문항 난이도의 신뢰도에 비해 낮으며 이를 통해 시험 문항들이 학생능력의 상위집단과 하위집단의 능력에 맞게 제작되지 못했음을 알 수 있다. 둘째, 평가의 적합도의 분석 결과, 점이연상관계수가 .20이하를 보이는 5개의 문항들은 국어 성적의 분포 양상에 기여하지 못하는 부적

* 이 논문은 2014년도(이정애)· 2016년도(심재우) 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.

** 제1저자(전북대학교 사범대학 국어교육과 교수)
*** 교신저자(전북대학교 사범대학 영어교육과 교수)

합 문항으로 판정하였다. 셋째, 문항 난이도가 1.5이며, 전체 문항의 33%가 너무 쉽게 출제되어 난이도상 적합하지 않았다. 넷째, 문법 영역은 '상, 중상, 중하, 하'의 난이도 위계를 고루 포함하나 5개의 문항은 문법 지식을 측정하는 비교적 어려운 문항으로 확인하였다. 반면에 독서 영역은 '상'에 해당하는 문항이 없으며 '중하'와 '하'에만 집중된 대체로 쉬운 문제였다. 본 연구가 갖는 의의는 Rasch를 통해 교사들에게 선택형문항에 대해 신뢰도 및 적합도, 문항난이도, 타당도 등을 판단할 수 있는 정보를 제공하며, 이를 바탕으로 객관적이고 발전적인 평가 방향을 모색할 수 있다는 점이다.

주제어 : Rasch 분석, 성취도 검사, 문항 반응 이론, 문항의 신뢰도, 적합도, 문항의 적절성, 문항 난이도의 위계성

1. 문제 제기

본 연구는 지방 소재 한 인문계 고등학교의 1차 고사에 투입된 국어과 선택과목 중 <독서와 문법>의 평가 문항을 대상으로 Rasch 모형을 통해 분석하고자 한다. Rasch 모형은 학교에서 치러지는 평가의 결과들이 얼마나 믿을 만한지에 대해 판단할 수 있는 일종의 문항반응이론(item response theory)으로 최근 널리 사용되고 있는 측정이론의 하나이다. 문항반응이론은 피험자들의 잠재적 특성(latent traits)을 문항들에 대한 그들의 반응에 의하여 예측할 수 있다고 가정함으로써, 교사가 평가 자료를 통해 측정하려는 특성에 대해 추정하고 그 추정 결과에 따라 학생들의 능력을 단순하게 비교하는 위험성을 상당히 해소할 수 있다는 장점을 갖고 있다(성태제, 2001 참조).

본 연구는 지방 소재 J시 인문계 고등학교 2학년 총 179명이 치른 국어교과 선택과목 중 중간고사 <독서와 문법>의 평가 문항을 수집하였으며 이

중 선택형 지필 문항만을 대상으로,[1] Rasch 모형에 적용함으로써 시험 결과의 원점수를 두고 학생들의 국어 능력을 서열척도로 하여 단순 비교하기 전에, 출제 문항에 대한 객관적 결과를 산출할 수 있을 것으로 생각한다.

먼저 본 연구가 인문계 고등학교의 중간고사를 대상으로 한 것은 사실 일반 학교의 정기고사에 대한 논의의 중요성 때문이다. 학교에서 이루어지는 정기적 시험 평가는 개별 학교와 교실, 그리고 개별 교과를 다루는 교사의 특성과 개성을 담아내고 있으며 이는 엄격한 표준화와 객관도를 지향하는 수능과는 다른 목적과 기능을 갖게 된다(박인기, 2000: 95 참조).

한편으로 본 연구가 선택형 문항을 분석대상으로 한 것은 그동안 많은 논의를 통해 선택형 평가가 지닌 최대의 장점과 단점들이 동시에 지적되어 왔음에도 불구하고,[2] 여전히 학교 현장에서 선택형 문항들이 확고부동한 평가 유형으로 사용된다는 점이다.[3] 다만 선택형 지필 평가 방식이 일선 학교에서 실시되는 정기 고사는 물론이며 수능이나 국가 수준 학업성취도에서 여전히 유효하지만 학교의 시험 문항들에 대한 논의가 상대적으로 수능이나 국가 수준 학업성취도 평가에 비해 많지 않았다.

이제 본 연구가 Rasch 모형을 통한 문항 분석으로 다음과 같은 문제들을 논의하는 일은 현재 특정 학교의 현실적 차원에서 선택형 지필평가로 이루어지는 1차 또는 2차 고사의 평가 의의를 손상하지 않은 채, 국어 교과의 시험 문항에 대한 몇 가지의 의미 있는 일반화를 가능하게 할 것으로 본다. 다시 말해 교사가 출제한 평가 문항은 믿을 만한지, 학교의 평가가 학생들의

---

1) 선택형 지필평가만을 분석대상으로 한 것은 단지 편의상의 문제이며 논술형 수행평가의 채점 방식도 Rasch 모형에 적용된다(지은림, 1996, 2000). 국어의 경우, 1차(중간)고사와 2차(기말)고사가 지필 평가로 실시되며, 평가 형식은 선택형이 70%, 서답형이 전체 배점의 약 30% 정도로 부가되며, 그 중 소위 주관식이라고 하는 단답형 20%와 논술식의 서술형 10%로 출제되도록 권장된다.

2) 무수한 국어과 평가관련 연구에서는 선택형 문항이 갖는 평가의 단점을 주로 지적하였으며 대안으로 수행평가가 논의되거나(최미숙, 1998, 2000, 2006), 그 한계와 가능성이 논의되었다(최지현, 2000; 최미숙, 2004; 김중신, 2008).

3) 절대적 공평성을 갖는 지식의 습득 여부를 판별할 수 있으며, 학력 관리를 객관적이고 효율적으로 할 수 있다는 점이 선택형 평가의 가장 큰 매력일 것이다.

능력을 제대로 그리고 적절하게 평가하고 있는지, 교사들은 자신이 직접 출제한 국어 문항에 대한 정보를 제대로 얻을 수 있는지, 그리고 이를 바탕으로 한 교사의 판단과 기술은 타당하고 믿을 만한지에 대해 어느 정도 설명해 줄 것으로 기대한다.[4)]

1) 교사가 출제하는 국어과 평가 문항의 신뢰도는 어느 정도인가?
2) 국어과 평가 문항이 학생능력을 제대로 평가하고 있으며 그렇지 못한 부적합 문항은 무엇인가?
3) 국어과 평가 문항은 학생능력 수준에 따라 적절하게 평가되었는가?
4) 국어과 평가 문항의 난이도 위계에서 발견되는 특징인 무엇인가?

2. 이론적 배경

선다형 지필 평가는 거시적 수준의 대규모 표준화 검사뿐만 아니라 학교 수준의 총괄 평가에 이르기까지 여전히 광범위하게 사용되어 온 평가방식이다. 그동안 국어 교과의 평가 과정과 평가 도구의 논의 속에서 선다형 지필 평가에 대한 많은 비판과 문제점들은 다음과 같은 몇 가지 흐름으로 국어교육의 발전과 국어과 평가 체제의 개선을 지향하고 있음을 알 수 있다.

첫째, 국어교과교육의 차원에서 평가론은 노명완(1988, 1990, 2004), 김광해(1994), 이삼형(1999), 김창원(1999), 박인기(2000, 2008) 등에 이르러 국어교육 평가의 방향이나 원칙, 방법 등이 충분히 모색되었으며 이를 바탕으로 국어과 평가에서 선다형 지필평가는 새로운 평가 패러다임을 모색하려는 관점에서 검토되고 논의되었다는 점이다. 즉, 국어과는 사회 변화의 필연적 산

4) 최근 일반 학교 NEIS(교육행정정보시스템)에는 과목마다 문항반응이론에 기반하여 문항의 개별적 속성을 분석하고 이를 프로그램화하여 제시하고 있다. 즉 지필평가 문항 분석표, 지필평가 교과목별 학생 답 정오표, 지필평가 정답률비교표 등을 쉽게 조회할 수 있으며, 이를 바탕으로 교사는 출제한 평가 문항에 대한 정보와 이해를 얻을 수 있다. 그러나 본 연구와 같은 정밀한 분석과 해석을 얻기에는 부족하다고 생각한다.

물인 교육목적과 교수·학습 과정에 대한 관점의 변화가 새로운 평가 패러다임을 요구하는 맥락과 맞물려, 전통적 평가관에 따른 선다형 지필 평가의 한계와 가능성이 논의되었으며(최지현, 2000; 류수열, 2004), 수행평가를 중심으로 한 대안적 평가론이 지속적으로 제기되었다(최호성, 1996; 최미숙, 1998, 2000, 2006; 김혜정, 2008).

둘째, 성취도 평가 중심으로 평가 내용의 타당도, 평가와 교육과정 내용과의 일치도, 문항 설계의 타당도, 채점의 신뢰도 등의 많은 논의들을 통해[5] 평가에 대한 내용과 방법과 도구제작에 대한 정교화가 요구되었다는 점이다. 그 결과 국어과는 국가 수준의 학업성취도 평가에서, 그리고 대학수학능력시험과 학교 수준의 총괄 평가에서 선다형 평가를 대상으로 국어과의 하위영역인 '듣기·말하기', '읽기', '쓰기', '문법', '문학' 별 특성과 평가 개선점 등의 논의가 활발히 전개되었다(김은애·주세형, 2010; 김혜정, 2011; 이은희, 2011; 주세형, 2010; 최지현, 2011 등).[6]

셋째, 학교의 평가에 대한 내용과 방법과 도구의 논의가 평가의 주체인 교사를 제외할 수 없는 문제이니만큼 평가문항 개발 및 문항 분석에 관련하여 평가 전문성의 연구들도 활발히 이루어졌다. 한국교육과정 평가원(2004)의 국어과 교사의 평가전문성 신장 모형과 기준이 제시되었으며, 뒤이어 많은 연구에서 교사의 평가 문식성과 교사평가 전문성 신장의 중요성이 평가문항의 분석을 바탕으로 다루어졌다(김종윤, 2006; 주세형, 2011; 정혜승, 2008; 임천택, 2010; 박준용, 2013).

지금까지의 논의들을 통해 결국 전통적 평가관으로 오랫동안 사용되어 온 선다형 중심의 지필고사는 평가 과정과 평가 도구 제작에 대한 검토를 바탕으로 평가 도구로서의 타당성과 정교성을 획득하면서 진화하고 있음을 알 수 있다. 본 연구는 학교에서 이미 이루어진 평가 결과 중심으로 하여 Rasch 모

5) 최근 국가 수준 학업성취도 평가를 대상으로 많은 연구들이 이루어졌다(조재윤, 2011; 김영란·김인숙, 2012 등)

6) 물론 지필평가의 문항 분석도 활발히 이루어졌다(유재영, 2004; 김진주, 2006; 하상수 2009; 이영범, 2011 등).

형을 적용하여 평가 문항을 분석함으로써 선다형 지필평가에 대한 평가 도구로서의 신뢰도와 타당도를 검토하는 것은 물론이고 이러한 논의가 교사의 문항 개발에 대한 방향과 제언을 제공할 것으로 기대할 수 있다.

Rasch 이론은 관찰되는 평가 점수를 그대로 사용하는 것(원점수) 대신에 관찰된 피험자들의 반응으로부터 그들의 실제 능력을 추정해내는 것이 더 신뢰할 수 있다고 본다(지은림·채선희, 2000 참조). 즉, 관찰된 피험자들의 반응으로부터 그들의 능력을 추정할 수 있도록 하는 측정이론이 바로 문항반응이론이며 Rasch 모형도 역시 이에 근거한다. 이 측정이론을 사용하여 피험자들의 능력을 추정하고 비교한다면 원점수를 그대로 사용하는 것보다 더 많은 객관성을 확보하게 되므로, 평가에 사용되는 문항들이 어려운 것이든지 쉬운 것이든지에 관계없이 학생들의 능력비교가 일정하게 이루어질 수 있고, 난이도 역시 학생들의 능력수준과 관계없이 일정하게 비교될 수 있다는 것이다(지은림·채선희, 2000 참조).

다만 이러한 문항반응의 이론의 장점에 비해, 이 이론이 갖는 수리적 복잡성 때문에 그다지 활발히 적용되지 못했다가 비로소 복잡한 계산과정을 위한 컴퓨터프로그램들(Wright and Mead, 1976; Wright, Rossner, Schulz, and Congdon, 1988; Smith, 1991; Wright & Linacre, 1992; 지은림·채선희, 2000 참조)이 개발되어 그 실용성이 크게 높아지게 되었다. 본 연구도 그에 힘입은 바가 크다.

3. 연구 방법

3.1. 분석 대상 : <독서와 문법> 평가 문항

<독서와 문법>은 <화법과 작문>과 함께 교육과정 개편에 따른 변화로 인문계 고등학교 국어 교과목들 간의 결합을 통해 생성된 과목의 하나이다. 본 연구 자료는 J 고등학교 <독서와 문법> 과목 1차 고사의 평가문항이며,

총 선택형 24문항을 대상으로 하였으며 100점 만점을 기준으로 선택형 70점이다. 30점을 차지하는 서답형(4문항) 및 서술형(3문항)은 제외하였다.

자료를 제공받은 J 고등학교는 지방의 인문계 여자고등학교로서 오랜 역사를 지닌 지역 명문고로 알려져 있으며 학업성취도 수준은 도내 7개 국·공립고등학교 중에서 언제나 2~3위이다. 학생 정원은 1,074명이며 학년당 10학급, 총 30학급의 규모로서, 2학년부터 자연이공계열(5학급)과 인문사회계열(5학급)로 나뉘게 된다. <독서와 문법>은 <문학>과목과 함께 인문사회계열 학생들이 필수로 받는 과목이며, 본 연구의 피험자인 학생들은 이 학교 2학년 인문사회계열 179명이다. 2학년 <독서와 문법>의 담당 국어교사는 2명으로 출제는 대단원별로 맡아서 하지만 최종적으로 서로 협의하고 검토하는 공동 출제 방식임을 확인하였다.[7]

3.2. Rasch 모형의 문항 분석

본 연구는 Rasch 모형의 문항분석을 위해 Winsteps 3.68 소프트 프로그램을 사용하였다. 실제 평가의 측정에서 능력이 낮은 학생이 추측이나 다른 요인들에 의하여 어려운 문항에서 정답을 하는 경우도 있고, 그 반대로 능력이 높은 학생이 실수나 다른 요인에 의하여 쉬운 문항에서 오답을 하는 경우도 일어날 수 있는데, Rasch 모형은 이처럼 학생의 반응을 완벽하게 예측할 수 없으므로 학생들의 반응을 일으키는 조건으로 학생과 문항이 상호작용하는 것을 확률로 나타내려고 하는 데에서 출발하였다(지은림·채선희, 2000; 성태제, 2001 참조). Rasch 모형을 다음과 같은 산술적 공식으로 표현할 수 있다.

7) 교재는 윤여탁 외 <고등학교 독서와 문법>, ㈜ 미래엔(2013)이다.

$$Pni(Xni = 1/Bn, Di) = e^{(Bn-Di)}/1+e^{(Bn-Di)}$$

Pni = 피실험자가 정답을 제공할 확률
Bn = 피실험자의 능력
Di = 문항의 난이도

위의 공식에서 볼 수 있듯이, 문항 난이도보다 학생들의 능력이 더 높으면 높을수록 학생들이 정답을 할 확률이 높아질 것으로 기대하고, 문항 난이도보다 학생의 능력이 더 낮으면 낮을수록 학생이 정답을 할 확률은 낮아지고 오히려 오답을 할 확률이 더 높아질 것을 기대한 것이다. 학생의 반응은 학생의 능력과 문항 난이도에 의하여 완벽하게 결정되는 것이 아니라 확률적으로 설명할 수 있다고 보는 것이 Rasch 모형의 장점 중 하나이다.[8)]

고전검사이론에서 문항의 난이도(item difficulty)는 피실험자 집단의 평균적 능력(ability)에 따라 편차가 형성된다. 하지만 Rasch 분석은 학생이 어떤 한 문항을 두고 정답을 할 확률값을 구하고 그 확률을 로지스틱(logistic) 변환으로써 학생 능력을 위한 측정치를 계산하게 된다. 한편 문항 난이도를 위해서는 오답할 확률을 로지스틱 변환하여 측정치를 구하게 된다. 로지스틱 변환 후에 산출되는 '로짓(logit)' 단위의 특정치들은 선형적인 조건을 만족시키면서도 동간 척도를 이루게 되어 피험자 능력과 문항 난이도에 따라 정답 활 확률과 그 확률을 로짓으로 변환시켜 특정치를 산출하는 것이다. 이 확률적 모델은 피실험자가 각각의 문항에 대하여 정답을 제공할 수 있는 여부를 일괄되게 측정할 수 있으므로, Rasch 분석을 통해 산출되는 값들을 통하여, 연구자는 모집단(population)에서의 모수들을 유추 또는 일반화할 수 있게

8) 이러한 Rasch 모형의 장점은 국어의 논술 채점이나 쓰기 평가에서 채점의 정확성에 영향을 주는 채점자의 특성을 분석하고 이를 효과적으로 통제하여 평가의 신뢰성 연구에 기여하였으며(이현숙, 2008; 박영민·최숙기, 2010), 쓰기 지도 시 학생들의 쓰기 불안과 같은 정의적 요인들의 척도가 탐색되거나(최숙기, 2011ㄱ), 또는 쓰기평가의 수행 시 교사의 평가 준거의 타당도 분석 및 평가 전문성을 판단할 수 있는 척도 개발에 적용되어 왔다(최숙기, 2011ㄴ; 최숙기, 2015).

된다. 이 장점을 Rasch 분석은 불변성(invariance)라고 명하고 있다.

Rasch 분석이 갖는 또 하나의 장점은 피실험자의 능력(즉, 문항의 정답을 맞힐 수 있는 능력)과 문항의 난이도를 같은 로짓이라는 척도를 사용하여 상대적 비교를 가능하게 한다는 것이다. 예를 들어, 학생의 능력과 문항의 난이도의 위치를 같은 척도를 사용하여 비교함으로써 학생인 피실험자가 어떠한 문항의 정답을 제시할 수 있는 가능성을 예측할 수 있다. 즉, 학생의 능력 로짓값이 문항의 난도 로짓값과 같으면, 학생이 그 문항의 정답을 맞출 수 있는 가능성은 50%임을 암시하며, 학생의 능력 로짓값이 문항의 난도 로짓값보다 높으면, 당연히 학생은 그 문항의 정답을 제시할 능력이 있다고 판단할 수 있다. 마찬가지로 학생의 능력 로짓값이 문항의 난이도 로짓값보다 낮으면, 학생은 그 문항의 정답을 맞힐 수 있는 능력이 부족하다고 볼 수 있다. 이러한 상대적 비교를 통하여, Rasch 분석은 학생 각각의 능력을 파악할 수 있을 뿐만 아니라 문항간의 난이도의 서열을 알 수 있게 해준다(Bond·Fox, 2007).

이처럼 Rasch 모형은 학생의 능력을 추정하기 위하여 로지스틱 함수를 이용하여 학생들이 정답으로 반응할 확률을 모형화한 것으로, 기본적으로 피험자의 능력을 문항 난이도와의 관계에 의해서만 설명할 수 있다. 그럼으로써 학생의 반응이 정답인가 오답인가에 대해 문항 난이도 이외에 추측이나 문항 변별도와 같은 다른 요인들에 의해 영향을 받을 수 있는 점을 배제할 수 있는 것이다. 본 연구는 Rasch가 갖는 이러한 장점에 기대어 평가지 문항의 신뢰도, 평가의 적합도와 부적합 문항, 평가의 적절성, 난이도의 특징들에 대하여 분석하기로 한다.

4. 분석 결과

4.1. 학생능력 신뢰도와 문항난이도 신뢰도

본 연구는 우선 학생능력 신뢰도와 문항난이도 신뢰도를 분석하였다. 그 이유는 학생의 능력 추정치의 일관성과 문항의 난이도 추정치의 일관성을 통해 출제된 문항의 질에 대한 유추를 할 수 있기 때문이다. Rasch 분석을 통하여 나타난 피실험자인 학생능력 신뢰도와 문항난이도 신뢰도는 다음의 <표 1>과 <표 2>이다. 먼저, 학생능력 신뢰도란 동일 학생이 이번 연구에 사용된 국어 평가지와 비슷한 수준의 다른 시험 문제를 풀었을 경우, 학생의 국어능력 분포가 어느 정도 일관성 있게 재현될 수 있는가를 말한다. 다시 말해 상위권 학생에게 이번 평가지와 유사한 시험지를 주었을 때 그 시험에서도 이 학생의 능력 분포가 역시 상위권임을 보여줌으로써, 해당 시험지가 학생의 국어능력 분포에 일관성 있게 재현된다는 것을 Rasch 모형이 측정하고 있음을 알 수 있다.

반면에 문항난이도 신뢰도는 이 시험을 치른 학생과 비슷한 수준의 국어능력을 가진 다른 집단의 학생이 똑같은 문항의 시험지를 풀었을 경우, 문항간의 난이도가 어느 정도 일관성 있게 재현될 수 있는가를 나타낸다. 문항난이도 신뢰도는 다른 평가 분석에서 많이 측정되지만 학생능력 신뢰도는 Rasch 분석에서만 보여줄 수 있다.

일반적으로 학생능력 신뢰도는 학생의 피험자 점수의 분산, 출제문제의 수, 문항의 범주의 길이, 능력수준과 문항난이도의 적절한 일치정도 등의 복합적 요인에 영향을 받는다. 먼저 <표 1>에 의하면, 본 평가지 경우 학생능력 신뢰도는 0.62로서, 최소 수준인 0.60을 가까스로 넘는 수준임을 보여주고 있다.[9] 학생능력 신뢰도가 0.9 정도일 때는 학생능력별 집단을 3개 또는 4개의 수준으로, 학생능력 신뢰도 0.8 정도일 때는 학생능력별 집단을 2개 또는

9) 해당 수치는 표에서 진한 이텔릭체로 표시됨.

3개의 수준으로, 학생능력 신뢰도가 0.5 정도일 때는 학생능력별 집단을 1 또는 2개의 수준으로 구분할 수 있다는 것을 뜻한다. 따라서 이 평가지의 학생능력 신뢰도가 0.62라는 것은 학생들의 능력 정도를 2개 수준, 즉 최대 상위집단과 하위집단 정도로만 나누고 있음을 알 수 있다.

다음으로 <표 1>는 학생능력 수준이 평균 1.5임을 보여준다. 이는 <표 2>의 문항난이도 수준 평균 0보다 1.5 이상 높으며, 이처럼 난이도 평균보다 높다는 것은 학생의 능력신뢰도가 문항의 난이도 신뢰도에 비해 낮게 나타남을 뜻한다. 또한 학생 능력 분리도지수도가 1.27을 보여 줌으로써 기준치 2 이하를 나타내고 있다. 분리도지수가 기준치 2 이하라는 것은 이 평가지가 상위집단과 하위집단을 정확하게 구분할 수 있도록 제작되지 않았다는 것을 뜻한다. 결국 이 평가지가 상위집단과 하위집단의 학생을 정확하게 구분하여 평가하기 위해서는 교사가 출제된 문항의 수준을 재조정할 필요가 있다고 볼 수 있다.

한편 <표 2>는 국어 평가지의 문항난이도 신뢰도가 .96으로 높게 나타났으며, 문항의 분리도지수도 5.13으로, 기준치 3 이상임을 보여주고 있다. 이때 기준치 3 이상이란 시험지의 문항 난이도가 5개의 수준, 예를 들면 상, 상중, 중, 중하, 하 등의 그룹으로 분리될 수 있으며 피실험자수도 충분하다는 것을 뜻한다.

<표 1> 학생능력 신뢰도

	로짓값	표준오차	내적합도 (MNSQ)	표준화된 내적합도 (ZSTD)	외적합도 (MNSQ)	표준화된 외적합도 (ZSTD)
평균	*1.50*	.59	.99	.0	.99	.1
표준편차	1.02	.12	.26	.9	.85	.7
최대값	4.20	1.11	1.99	3.0	9.90	4.8
최소값	-.72	.48	.48	-2.0	.11	-1.0
분리도지수 *1.27* 학생능력의 신뢰도 *.62*						

<표 2> 문항난이도 신뢰도

	로짓값	표준오차	내적합도 (MNSQ)	표준화된 내적합도 (ZSTD)	외적합도 (MNSQ)	표준화된 외적합도 (ZSTD)
평균	*.0*	.25	1.00	.0	1.0	-.1
표준편차	1.60	.17	.09	1.0	.25	1.2
최대값	3.60	1.01	1.22	2.3	1.62	1.9
최소값	-4.12	.17	.85	-2.4	.68	-2.3
분리도지수 *5.13* 문항난이도 신뢰도 *.96*						

4.2. 평가의 적합도

평가의 적합도, 즉 평가지의 개개 문항이 제대로 학생의 능력을 평가하고 있는가를 알 수 있는 방법으로 Rasch 모형의 기댓값을 살펴볼 수 있다. 분석 결과 개개 문항이 기댓값을 충족하면 문항의 문제점이 크게 발견되지 않지만, 기댓값을 벗어나 있다면 이는 문제가 있는 문항으로 수정하거나 검토가 요구된다. Rasch 모형은 문항이 이 모형의 기댓값을 잘 따르고 있는가를 살피는 통계값으로 내부적합도와 외부적합도 수치를 사용한다.

본 연구는 Boone, Staver, Yale(2013)의 문항적합도 기준에 따라, 외적합도(MNSQ)가 0.7에서 1.3 사이와 표준화된 외적합도(OUTFIT ZSTD) -2이하와 +2 사이값을 적합, 부적합 기준으로 삼았다. 또한 점이연상관계수(PT-Measure Correlation) .20이하의 문항을 판별도가 낮은 부적합 문항으로 분류하였다.[10] 본 연구의 평가지에 대해 위의 기준을 사용하였을 때, 모든 문항은 외적합도와 표준화된 외적합도를 충족하였다. 하지만 아래의 <표 3>에

10) 학자들에 따라 점이연상관계수의 기준을 .20 이하, 0.30 이하, 또는 0.40 이하를 기준으로 삼고 있지만, 본 연구는 .20 이하를 기준으로 삼았다.

서 문항 1, 2, 17, 15, 23은 점이연상관계수가 .20 이하를 보여 평가 변별도에 문제가 있음을 보여준다.[11]

점이연상관계수를 통해 나타나는 부적합 문항의 공통점은 이 문항들이 국어 성적의 분포 양상에 크게 관계가 없다는 것이다. 바꾸어 말하면, 이 문항들은 변별도가 매우 낮은 문항들이라고 할 수 있다.[12]

<표 3> 문항 적합도

문항 번호	문항 측정치	표준 오차	내적합도 (MNSQ)	표준화된 내적합도 (ZSTD)	외적합도 (MNSQ)	표준화된 외적합도 (ZSTD)	점이연 상관계수
23	3.60	.23	1.16	1.1	1.27	1.1	***.20***
24	2.79	.19	1.08	.9	1.18	1.1	.32
11	1.57	.17	.98	-.3	.99	-.1	.44
20	1.35	.17	.92	-1.4	.87	-1.5	.50
5	1.29	.17	1.07	1.1	1.16	1.7	.34
7	1.24	.17	1.15	2.3	1.17	1.8	.28
22	.98	.17	1.01	.2	.95	-.5	.41
9	.84	.17	.85	-2.4	.76	-2.3	.54
21	.84	.17	.92	-1.3	.82	-1.7	.49
8	.08	.19	.90	-1.1	.75	-1.5	.45
13	.08	.19	1.09	1.0	1.19	1.1	.24
3	.01	.19	.90	-1.0	.80	-1.1	.43
4	-.22	.20	1.04	.4	.91	-.4	.30
15	-.26	.20	1.22	1.9	1.45	1.9	***.08***

11) 점이연상관계수란 정답/오답과 같은 명목적 데이터와 등간척도를 사용한 성적 데이터간의 상간도를 말한다. 점이연상관계수는 문항을 맞게 답한 학생들의 평균과 틀리게 답한 학생들의 평균의 차이를 전체성적의 표준편차로 나눈 후에 그 값을 정답을 한 학생의 비율과 오답을 한 학생의 비율의 루트값에 곱하여 계산한 것이다.

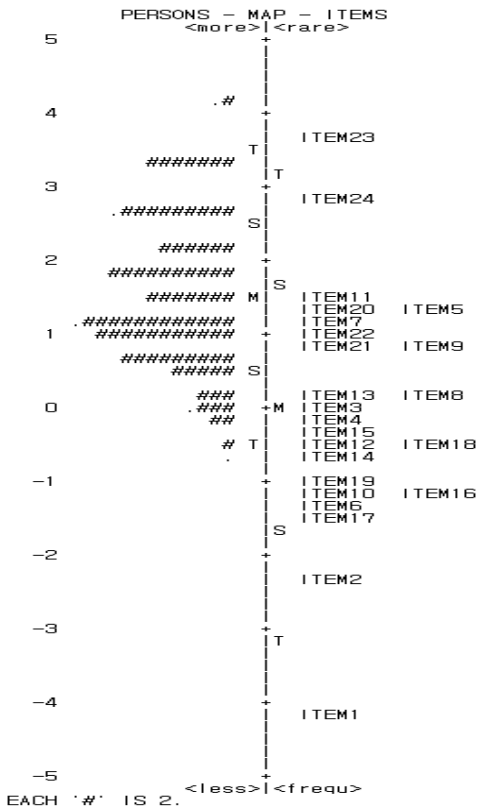
12) 이 문항들의 변별도 문제 발생과 원인의 기술은 4.4 참조.

12	-.58	.22	.91	-.6	.69	-1.2	.39
18	-.58	.22	1.01	.1	.99	.0	.28
14	-.63	.23	.97	-.1	1.11	.5	.29
19	-.97	.25	.96	-.2	.84	-.4	.30
16	-1.11	.26	.97	-.1	.84	-.4	.27
10	-1.18	.27	.99	0	.82	-.4	.25
6	-1.33	.29	.95	-.2	.68	-.8	.29
17	-1.42	.30	1.01	.1	1.38	1.0	***.14***
2	-2.27	.42	.97	.1	.75	-.2	***.18***
1	-4.12	1.01	1.02	.3	1.62	.9	***.00***

4.3. 평가의 적절성

Rasch 분석을 통하여 우리는 국어 평가 문항이 학생 능력의 수준에 맞게 적절하게 평가하고 있는가를 판단할 수 있다. <그림 1>에서 학생의 능력을 왼쪽으로, 문항의 난이도는 오른쪽으로 배치하여 능력과 문항의 배치도 관계를 알 수 있는 문항지도로서, 학습자의 능력과 문항의 난이도의 차이를 자연로그값, 즉 로짓값으로 변환한 것이다. 이때 변환된 로짓값은 각각 등급 간 차이가 진정한 의미의 등간척도(interval scale)를 형성하게 된다. 따라서 <그림 1>에서 로짓값 2와 3의 간격은 1 로짓 간격이고, 로짓값 -2와 -1의 간격도 1 로짓 차이로 나타나므로, 학생의 능력 또는 문항의 난이도의 간격이 1의 간격으로 대등하다고 할 수 있다.

<그림 1> 문항 지도



<그림 1>의 문항 지도에 의하면,[13] 다음과 같은 정보들이 해석될 수 있다. 먼저 <그림 1>에서 우리는 왼쪽의 M(능력평균), 1.5 로짓과 오른쪽의 문항의 난이도 M(난이도평균), 0을 확인할 수 있다. 이는 문항의 난이도가 국어능력 수준에 비해 1.5 로짓값이 낮음을 나타낸다. 보통 1로짓의 차이가

13) 왼쪽 능력배치도에 표시된 #와 .은 각각 2명과 1명의 학생수를 나타내며, 모두 179명의 학생이 문항별 반응에 따라 분포함을 보여 준다. 또한 학생과 문항이 같은 로짓값을 갖는 경우, 학생이 그 문항을 정답으로 맞출 수 있는 확률이 50%이며, 이 50%는 정답을 맞출 확률이 존재하는 것으로 파악할 수 있다(3.2참조).

날 때는 한 학년의 수업양으로 극복될 수 있다고 보며, 학생능력과 문항난이도 차이가 1.5 로짓일 경우 80%의 높은 정답률이 나올 수 있다고 판단한다. 이 사실을 고려한다면, 이 평가지의 문항 난이도가 1.5라는 것은 문항이 매우 쉽게 출제되었다는 것을 의미한다.

다음으로 왼쪽 점 (.)으로 표시된 문항지도 오른쪽 14번 문항(ITEM 14)이하로 배치된 19번, 10번, 16번, 6번, 17번, 2번, 1문항까지 총 8문항은 모든 학생이 이 문항들을 정확하게 맞출 가능성이 매우 높다. 이는 평가지의 총 24개의 문항 중 8문항, 즉 33%의 문항이 너무 쉽게 출제되어 학생들에게 난이도상 적합(mis-targeted)하지 못하다는 해석이 가능하다.

마지막으로 단지 2개의 문항만(23번과 24번)이 상위 19명의 학생들의 능력 수준에 배치된 결과를 보였다. 이는 상위권 학생들에게 적합한 국어능력 평가 문항은 단지 2개뿐인 것으로, 결과적으로 상위권 학생들의 국어능력을 측정하는 문항이 심각하게 결여된 것을 볼 수 있다. 즉 상위권 학생들의 능력이 충분히 평가될 수 없다는 것을 의미한다.

결국 <그림 1>의 문항 지도는 전체 평가 문항이 학생들의 능력에 비해 너무 쉽게 출제되었으며, 상위권 학생의 능력을 평가할 수 있는 문항이 적어 평가의 적절성이 충분히 확보되지 못함을 보여 준다.

4.4. 문항 난이도의 위계성

난이도의 특징을 살펴보기에 앞서, 먼저 본 연구의 평가문항들이 국어능력만을 측정하고 있는가를 확인해야 한다. 만일 이 평가문항이 국어 능력만을 적절한 측정 수준으로 평가하고 있다면, 이를 바탕으로 평가지의 문항 지도는 어떤 난이도의 특징을 보여주는가를 살펴볼 수 있다. 먼저 본 연구에 사용된 국어 평가의 문항들이 국어능력만을 측정하고 있는가를 살펴보기 위하여 표준화 잔차(residuals)의 주성분분석을 적용한 결과, 이 시험을 통하여 설명된 분산은 다음 <표 4>처럼 34%로 나타났다.

Reckase(1979)의 기준에 따르면, 40% 이상의 설명량은 높은 측정수준,

30% 이상은 적절한 측정수준, 20% 이상은 최소측정 수준으로 판단할 수 있으므로 본 연구에 적용된 문항들은 설명된 분산은 높지는 않지만, 적절한 측정 수준에 속한다고 할 수 있다. 따라서 '34%의 설명된 분산'이란 첫 번째 고유치 분산의 설명된 분산 5.3%와 비교했을 때 6배 이상으로 높으며, 결과적으로 문항들이 국어능력의 요인만을 측정하고 있다는 것을 충분히 알려주고 있다.

<표 4> 표준화 잔차에 의한 분산

	분산고유값	%
전체분산	36.3	100
Rasch 모형에 의해 설명된 분산	12.3	***34***
응답자에 의해 설명된 분산	4.0	11.1
문항에 의해 설명된 분산	8.3	22.8
Rasch 모형에 의해 설명되지 않은 분산	24	66
첫 번째 고유치 분산	1.9	***5.3***
두 번째 고유치 분산	1.6	4.4
세 번째 고유치 분산	1.5	4.1
네 번째 고유치 분산	1.4	3.8

그렇다면 이 시험의 문항들이 국어 능력만을 측정한다고 할 때, 이 결과를 통하여 이번에는 문항 난이도의 위계성(hierarchy)을 조사함으로써 학생들의 국어능력 측정 요인을 살펴볼 수 있다. 본 연구에서 <그림 1>의 문항지도가 나타낸 난이도를 살펴본 결과 <표 5>와 같은 난이도별 위계성을 발견할 수 있다.[14]

14) 4.2에서 점이연상과계수가 .20이하를 보여 변별도에 문제가 있음을 보여준 문항 1, 2, 15, 17, 23은 괄호로 처리함.

<표 5> 난이도별 측정 요인

난이도	영역	문항수	해당문항	측정 요인
상	문법	2	(23), 24	문법 지식
	독서	0	없음	
중상	문법	5	5, 7, 20, 21, 22	문법지식 및 추론적 독해
	독서	2	9, 11	
중하	문법	3	3, 4, 8	내용 확인 및 이해
	독서	4	12, 13, (15), 18	
하	문법	3	(1), (2), 6	내용 확인 및 이해
	독서	5	10, 14, 16, (17), 19	

<독서와 문법>의 총 24개의 문항은 문법영역 13개, 독서영역 11개로 구성되어 비교적 두 영역 간 문항 비율이 균형 있게 배치되었다. 문법 영역 13개 문항 중에서 8개 문항(1~8번)은 언어의 본질을 설명하는 교과서의 지문이 사용된 국어 지식 영역을 평가하고 있으며(주 15 참조), 나머지 5개 문법 문항(20~24번)은 지문 제시 없이 국어 음운의 체계와 음운 변동에 대한 내용을 평가하고 있다(주 16 참조). 반면에 독서 영역 11개 문항 중에서 9~17번의 9개 문항은 글의 구성 원리와 관련된 것이고, 18번과 19번의 2개 문항은 독서의 방법에 관한 문항이다.

문법 영역은 '상, 중상, 중하, 하'의 난이도 위계를 고루 포함하는데, 그 중 변별도에 문제가 있는 부적합 문항 23번, 1번, 2번을 제외하면, 학생들이 가장 어렵게 여긴 상 수준의 24번과 중상 수준의 20번, 21번, 22번 등이 모두 전형적인 지식 능력을 측정하는 문제로서, 제시된 지문이나 예시가 없이 학생들은 문제를 읽어가면서 답지의 설명이 적절한지를 판단하도록 출제되었다.[15] 이는 해당 단원이 음운의 체계와 음운 변동을 알아야 접근이 가능하

15) 특히 이 평가지의 문법 문항 20~24번 문항은 다음과 같은 질문의 형식으로 이루어지고 있어서 단편적인 문법 지식을 묻고 있다(밑줄은 원문 대로임). 이러한

다는 특성이 작용했기 때문으로 보인다. 또한 중상 수준의 난이도를 보인 5번과 7번은 언어 현상과 기호의 개념과 관련한 추상적 개념의 이해력을 측정하고 있어 어려운 문항으로 볼 수 있으나 중하 수준의 3번, 4번, 8번은 교과서의 지문이 그대로 출제에 사용되어 문법 영역이라기보다 읽기 영역의 독해력을 요구하는 있으며, 아주 쉬운 문항으로 판별된 6번도 역시 언어의 본질과 관련한 내용을 제시 지문에서 확인하는 문제이다.[16]

평가 문항을 통해서 여전히 학교의 지필 고사는 1)문항 유형의 단순성, 2)이원 분류표에 의한 체계적인 평가 계획의 부족, 3)문법 평가 도구 개발의 적극성 부족 등을 그대로 노출하고 있다(주세형·정지은, 2007; 구본관, 2010).

20. 음운과 음성에 대한 설명으로 바른 것은?
21. 국어의 자음과 모음체계에 대한 설명으로 바르지 <u>못한</u> 것은?
22. 다음 중 각각의 음운 변동의 유형에 해당하는 예를 <u>모두 바르게</u> 제시한 것은?
23. 음운의 첨가 현상에 대한 설명으로 바르지 <u>못한</u> 것은?
24. 다음 중 음운 변동 현상에 대한 설명으로 바르지 <u>못한</u> 것은?

16) 문법 영역은 부적합으로 판명된 1번과 2번을 포함하여 8번 문항까지 제시하면 다음과 같다. 이 문항들은 문제의 제시문과 <보기> 내용이 교과서의 지문을 사용하였기 때문에 관련된 언어지식(언어의 본질)을 수업 시간에 배운 셈이 된다. 결국 교실 수업의 내용만 기억한다면 지문 없이도 쉽게 답을 찾을 수 있으며, 문법 영역이지만 주어진 답지에서 정답을 확인하는 '지식의 재인(recognition)' 수준으로 기본적 독해만 있다면 쉽게 정답을 찾을 수 있도록 출제가 된 것을 알 수 있다.

1. 위 글에서 다루고 있는 화제가 <u>아닌</u> 것은?(변별도 없음)
2. 위 글을 읽고 <보기>에 대해 파악한 내용으로 적절한 것은?(변별도 없음)
3. <보기>의 내용이 위 글에 대해 반박할 수 있는 내용은?(중하)
4. (라)를 뒷받침하는 사례로 적절한 것은?(중하)
5. (가)~(라)를 통해 파악할 수 있는 언어 현상에 대한 설명으로 적절하지 <u>않은</u> 것은?(중상)
6. <보기>를 바탕으로 동물 언어와 구별되는 인간 언어의 특성을 가장 잘 이해한 것은?
7. <u>㉠기호</u>의 속성을 가장 적절하게 설명한 것은?(중상)
8. 다음 글과 <보기 1>을 참고하여 <보기 2>의 ㉠~㉣을 의미 변화의 유형과 원인에 맞게 분류했을 때 적절하지 <u>않은</u> 것은?(중하)

독서 영역은 난이도 '상'에 해당하는 문제가 없으며, 중상 수준의 9번과 11번, 중하 수준의 12번, 13번, (15)번, 18번, 하 수준의 10번, 14번, 16번, (17)번, 19번을 포함한다. 중상 수준의 9번과 11번을 제외하면 대체로 쉬운 문제가 총 9개를 차지한다. 이 중 변별력을 갖지 못한 (15)번과 (17)번을 제외하고 보면, 나머지 7개 문항들은 교과서에서 지문과 <예시>가 그대로 인용되었기 때문에 독서 능력을 측정하는 문항이 단지 내용 확인의 문제가 되어 답을 쉽게 찾을 수 있었으며, 고등 수준의 사고력과 창의력을 평가할 수 있는 상 수준의 문항이 없다는 것을 알 수 있다.[17]

따라서 <문법> 과목이 아주 어렵거나 아주 쉬운 문제로 구성된 것에 비해, <독서>는 아주 어려운 난이도 '상'에 해당하는 문항이 없으며 '중하'와 '하'에만 집중되어 대체로 쉬운 문제였다는 점이다. 독서 영역의 문제가 쉬웠던 것은 평가에 사용된 지문이 100% 교과서에 제시된 것으로써,[18] 대부분 제시문에 드러나는 정보들 간의 관계를 확인하는 내용 확인의 문항으로 이루어졌기 때문이다. 생략된 정보를 유추하거나 새로운 정보로 재구하는 추론 능력이 요구되는 문항도 있었지만(9번, 11번),[19] 대다수 문제들이 교과서에

17) 독서 평가는 문제를 풀기 위해 학생들이 지문에 제시된 지식이나 정보들을 이해하고 활용하여 문제 풀이에 필요한 새로운 지식이나 정보를 창조해 낼 수 있도록 문항이 출제되어야 하므로 교과서 밖의 지문을 원칙으로 해야 함에도 불구하고, 이처럼 교과서 지문에 대한 단순한 '지식의 회상' 수준으로 문제가 출제되거나 지문을 읽지 않고도 답지에서 정답을 찾을 수 있는 문항이 독서 영역의 평가가 된 것은 일선 학교에서 쉽게 관찰되는 총괄평가의 관행이다.

18) 기존 연구에서 확인된 내용이다(유재영, 2004).

19) 9번과 11번의 문항은 다음과 같다.

9. 위 글을 바탕으로, 다음 <보기>를 읽은 것으로 적절하지 않은 것은?
11. '㉠ 비교우위 이론이 그리고 있는 장맛빛 구도'의 내용으로 적절한 것은?

이 두 문항은 같은 제시문으로 <독서와 문법> 교과서 지문의 일부가 사용되었다. 9번에서 제시한 <보기>는 교과서에서 언급되지 않은 내용으로 '자유 무역의 비교 우위 이론'에 대한 충분한 이해가 있어야 정답을 맞힐 수 있으며, 11번도 '㉠ 비교우위 이론이 그리고 있는 장맛빛 구도' 의 내용을 묻는 문제이기 때문에 역시 비교 우위 이론에 대한 이해와 지식과 관련된다. 다만 교과서에 제시

서 다루어진 내용을 다루고 있어 쉽게 답을 찾을 수 있는 것으로 분석된다.

이제 나머지 변별도가 매우 낮은 문항으로 판명되어 국어 성적의 분포 양상에 기여하지 못한 문법영역의 1번, 2번, 23번과 독서영역의 15번, 17번 등 5개 문항을 살펴보기로 한다. 문법 1번과 2번은 너무 쉽게 답을 찾을 수 있어서 변별력을 갖지 못하였다. 전체 177명의 피험자 중에서 1번은 오답자가 1명, 2번은 오답자가 6명일뿐이다.[20] (23)번은 '음운의 첨가'에 대한 문법 지식의 이해를 묻는 문항으로서 여러 답지를 읽어나가면서 문법 지식이나 정보를 기억하고 이를 예시를 통해 정답임을 확인해야 하는데, 답지의 내용이 정확하게 문법 내용을 기술하지 못한 것으로써 학습자들이 제대로 정답을 찾지 못해 '정답 찍기'로 나타난 것이다.[21]

반면에 변별도가 낮은 독서영역 (15)번과 (17)번은 둘 다 글의 구성원리인 '통일성'과 '응집성'을 묻고 있으나,[22] 답지 속에서 통일성과 응집성에 대한

된 지문과 달리 이 문항의 제시문은 그 일부가 사용되어 비교우위 이론에 대한 이해가 충분하지 않으면, 이 지문의 내용으로 정답을 찾을 수 없게 된다. 따라서 추론능력이 요구되기도 하지만, 제시문의 양이 문제 풀이에서 충분히 제공되지 않았음을 알 수 있다.

20) 1번은 제시문의 화제가 아닌 것을 고르는 문제인데 답지의 정답인 "⑤ 인간과 동물의 언어와의 관계"는 제시문에 한 번도 언급된 바가 없으며, '동물'이라는 어휘도 등장하지 않아, 쉽게 정답임을 찾을 수 있다. 2번은 <보기>에 제시된 내용이 언어와 문화의 관계를 찾는 문제인데, 고등학교 인문계 학생의 수준으로는 너무 쉽게 알 수 있는 내용이며, 정답 외의 다른 답지들은 너무 동떨어진 내용이 기술되어 있다.

21) 23번의 정답은 5번이지만 다른 답지에도 고르게 오답자가 분포되어 있다.

답안번호 / 문항번호	1번	2번	3번	4번	5번	무 표기	응시 자수
17	59명	19명	49명	22명	28명	2명	179명

22) 15번, 17번은 모두 '통일성과 응집성'의 이해 정도와 관련되며, 중복된 내용을 질문하고 있다.

어휘의 의미만으로도 쉽게 정답임을 확인할 수 있는 문항으로서 변별력을 갖지 못하였다.[23]

5. 결론

본 연구가 일반 인문계 고등학교의 <독서와 문법> 평가 문항을 대상으로 Rasch 모형을 적용함으로써 애초 제시한 목적을 어느 정도 성취했는지는 다음과 같이 정리함으로써 검토할 수 있다.

먼저, 본 연구에서 논의된 국어 평가지는 학생의 능력 신뢰도가 문항 난이도의 신뢰도에 비해 낮게 나타났으며, 분리도지수가 기준치 이하였다. 즉, 시험 문항들이 학생능력의 상위집단과 하위집단의 학생을 정확히 구분할 수 있도록 학생 능력에 맞게 제작되지 못했으며, 교사는 문항의 수준을 학생들의 능력에 따라 재조정해야 한다는 것을 뜻한다.

둘째, Rasch 분석을 통해 평가의 적합도의 분석 결과, 이 시험 문항들은 모두 외적합도와 표준화된 외적합도를 충족하였으나, 5개의 문항들이 점이연상관계수가 .20이하를 보여 변별도에 문제가 있음을 보여주었다. 즉, 5개의 문항들은 변별도가 매우 낮아 전체 국어 성적의 분포 양상에 기여하지 못했으므로 부적합 문항으로 판정하였다.

셋째, 평가의 적절성에 대한 분석 결과, 문항 난이도가 1.5로서 문항이 매우 쉽게 출제되었으며, 전체 문항의 33%가 너무 쉽게 출제되어 난이도상 적합하지 않았다. 또한 상위권 학생들에게 적합한 평가 문항은 단지 2개만 출제되었으며, 그 중 1개는 부적합 문항으로 판정되었다.

---

15. ㉠통일성의 측면에서 글을 쓰거나 퇴고할 때 고려해야 할 요소로 적절하지 않은 것은?
17. ㉢통일성과 응집성에 대한 내용으로 적절하지 않은 것은?

23) 왜 변별도가 낮은 문항이 발생하는가? 가령, <독서>와 <문법>이 갖는 교과 내용의 차이인지, 영역별 문항에 대한 학생들의 능력 차이지, 아니면 출제 교사의 문제인지 등을 본 연구에서 깊이있게 살펴보지 못하였다.

넷째, 총 24개의 문항은 <문법> 13개 문항과 <독서> 11개 문항으로 출제되었다. 문법 영역은 '상, 중상, 중하, 하'의 난이도 위계를 고루 포함하나 8개의 문항은 언어의 본질에 관련한 지식의 이해력과 추론을 요구하는 독해력을 요구하는 문제였으며, 나머지 5개의 문법 문항은 비교적 어려운 문항으로서 답지의 설명을 읽어가면서 풀어가는 지식 능력을 측정하고 있는 것으로 확인하였다. 반면에 독서영역은 '상'에 해당하는 문항이 없으며 '중하'와 '하'에만 집중되어 대체로 쉬운 문제였다. 독서 영역의 문제가 쉬웠던 것은 평가에 사용된 지문이 100% 교과서에서 출제된 것으로, 대부분 제시문에 드러나는 정보들 간의 관계를 확인하는 내용 확인의 문항으로 이루어졌기 때문으로 해석하였다.

교사는 자신이 출제한 문항, 특히 선택형문항에 대해 문항의 신뢰도, 문항의 적합도, 문항의 적절성, 문항난이도, 타당도 등을 판단할 줄 알아야 한다. 본 연구는 Rasch를 통해 이러한 정보를 얻고자 하였으며 Rasch의 분석이 어느 정도 설명해 주었다고 생각한다. 다만 교사가 일일이 문항분석을 할 수 없으며, 문항에 대한 반응 분석을 통해 알 수 있다는 점에서 결과론적이라는 한계가 있지만 이러한 프로그램의 정교화와 대중화를 통해 평가에 대한 기초 지식을 익힌다면 교사의 문항 제작에 크게 기여할 수 있을 것이라고 생각한다. 이에 대해서는 또 다른 논의가 필요할 것이다.

덧붙여 이 연구는 학교의 시험 문항 자체에 대한 개발과 질적 수준을 담보하는 시험 문제의 제작에 관련한 논의라기보다는, 객관식 선택형 지필평가가 여전히 학교에서 유효하고 또한 신뢰성을 갖는 평가라는 점을 Rasch 모형을 통해 얼마나 입증해 보일 수 있는가를 제시하려고 하였다. 이것을 바탕으로 우리는 객관적이고 발전적인 평가 방향을 모색할 수 있을 것이다.

참고문헌

구본관, 「문법 능력과 문법 평가 문항 개발의 방향」, 『국어교육학연구』 7, 국어교육학회, 2010, 185-216면.

구본관·조용기, 「문법영역 출제의 개선 방향 - 2014학년도 대학수학능력시험 국어 영역을 중심으로」, 『문법교육』 18, 한국문법교육학회, 2013, 1-44면.

김광해, 「국어교육 평가의 이상과 현실」, 『선청어문』 22, 서울대학교 국어교육과, 1994, 39-58면.

김영란·김인숙, 「국어과 국가수준 학업성취도 평가의 문항 개선 방안 연구」, 『국어교육학연구』 44, 국어교육학회, 2012, 67-110면.

김은애·주세형, 「독서 평가 문항의 내용 타당도 분석 : 고등학교 독서 수행 평가를 중심으로」, 『우리말교육현장연구』 4-1, 우리말교육현장학회, 2010, 177-249면.

김중신, 「국가수준 "국어/언어능력 검사"의 비판적 검토」, 『국어교육학연구』 31, 국어교육학회, 2008, 69-108면.

김종윤, 「국어과 교사의 평가 문식성에 대한 연구 - 읽기 영역 평가를 중심으로-」, 고려대학교 석사학위 논문, 2006.

김진주, 「읽기 평가 문항 분석을 통한 개선 방향 연구」, 이화여자대학교 교육대학원 석사학위 논문, 2006.

김창원, 「국어교육 평가의 구조와 원리」, 『한국초등국어교육』 15, 한국초등국어교육학회, 1999, 165-192면.

김혜정, 「고등학교 국어과 평가의 문제점과 개선 방안 - 평가 사례를 중심으로」, 『국어교육학연구』 32, 국어교육학회, 2008, 97-127면.

김혜정, 「국가 수준 학업성취도 평가 개선을 위한 '읽기' 평가 이론의 비판적 검토」, 『국어교육』 134, 한국어교육학회, 2011, 61-86면.

노명완, 「국어과 교육에서의 평가의 문제점과 개선의 방향」, 『국어교육론』, 한샘, 1988.

노명완, 「국어과 교육에서의 평가」, 『한국국어교육연구회논문집』 40, 한국어교육학회, 1990, 69-107면.
노명완, 「국어 시험의 현황과 개선 방안」, 『국어교육학연구』 20, 국어교육학회, 2004, 5-30면.
류수열, 「중등학교 국어시험의 현황과 개선 방향 - 고등학교 중간 고사 및 기말 고사를 중심으로」, 『국어교육학연구』 20, 국어교육학회, 2004, 59-83면.
박영민·최숙기, 「Rasch 모형을 활용한 국어교사의 쓰기 평가 특성 분석」, 『국어교육학연구』 37, 국어교육학회, 2010, 367-391면.
박인기, 「국어교육 평가의 패러다임 변화와 실천」, 『국어교육』 102, 한국국어교육연구회, 2000, 87-111면.
박인기, 「국어과 평가의 반성과 전망」, 『국어교육학연구』 32, 국어교육학회, 2008, 5-31면.
박준용, 「국어과 평가 전문가의 문항 개발 전문성 탐색 -국어과 문항 개발 협의 과정을 중심으로-」, 『한국어문교육』 13, 고려대학교 한국어문교육연구소, 2013, 57-89면.
성태제, 『문항반응이론의 이해와 적용』, 교육과학사, 2001.
성태제, 『문항제작 및 분석의 이론과 실제』, 학지사, 2004[1996].
유재영. 「고등학교 읽기 평가 문항 분석 연구」, 고려대학교 석사학위 논문, 2004.
이삼형, 「국어교육 평가의 관점」, 『국어교육학연구』 9, 국어교육학회, 1999, 257-280면.
이영범, 「중학교 시 단원의 지필평가 문항 실태 분석」, 전남대학교 교육대학원 석사학위 논문, 2011.
이은희, 「문법 평가의 현황과 과제 - 국가 수준 학업 성취도 평가를 중심으로-」, 『문법교육』 15, 한국문법교육학회, 2011, 1-29면.
이현숙, 「다국면 라쉬모형을 적용한 논술 채점 상황에서 채점 설계 및 채점자 특성이 채점의 정확성에 미치는 효과」, 『교육평가연구』

21(4), 한국교육평가학회, 2008, 129-152면.
임천택, 「초등교사의 국어과 평가 전문성 실태분석: 평가문항 개발을 중심으로」, 『학습자중심교과교육연구』 10, 학습자중심교과교육학회, 2010, 347-346면.
정혜승, 「교사의 읽기 평가 전문성 실태: 지필 평가 문항 분석을 중심으로」, 『독서연구』 19, 한국독서학회, 2008, 307-346면.
주세형, 「평가 문식성 신장을 위한 국어과 교사 교육」, 『문법교육』 15, 한국문법교육학회, 2010, 31-50면.
주세형, 「국어과 문법 평가 이론의 발전 방향」, 『국어교육』 135, 한국어교육학회, 2011, 39-66면.
주세형·정지은(2011), 「서울 K고 국어과 정기고사에서의 문법 평가 실태조사」, 『언어와 정보사회』 12, 서강대학교 언어정보연구소, 2011, 37-70면.
조재윤, 「국가 수준 국어과 학업성취도 평가의 적합성 검토 연구 : 국어과 평가의 적합성 검토 VRUT 모형을 사용하여」, 『한국초등국어교육』 46, 한국초등국어교육학회, 2011, 349-377면.
지은림, 「many-facet Rasch 모형을 적용한 대입 논술고사 채점의 객관성 연구」, 『교육평가연구』 9(2). 한국교육평가학회, 1996, 5-22면.
지은림, 「논술형 수행평가를 위한 채점방법들의 비교」, 『경희대학교 교육문제연구소 논문집』 16, 경희대학교 부설 교육연구소, 2000, 235-246면.
지은림·채선희, 『Rasch모형의 이론과 실제』, 교육과학사. 2000.
최미숙, 「국어교육에서의 평가-'수행평가'를 중심으로」, 『국어교육연구』 5(1), 서울대학교 국어교육연구소, 1998, 275-298면.
최미숙, 「국어교육 평가의 원리와 실제-'통합'의 원리를 중심으로」, 『국어국문학』 126, 국어국문학회, 2000, 101-121면.
최미숙, 「국가 수준 국어과 교육성취도평가의 실제와 발전 방안 -평가

문항을 중심으로」, 『국어교육학연구』 20, 국어교육학회, 2004, 237-265면.
최미숙, 「국어과 평가의 반성과 탐색」, 『국어교육』 121, 한국어교육학회, 2006, 79-105면.
최숙기, 「Rasch 평정척도 모형을 이용한 쓰기불안 척도 분석」, 『새국어교육』 87, 한국국어교육학회, 2011ㄱ, 273-300면.
최숙기, 「Rasch 모형을 활용한 요약문 평가 준거 개발 및 타당도 분석 : 고등학생 설명문 텍스트 요약하기를 중심으로」, 『독서연구』 25, 한국독서학회, 2011ㄴ, 415-451면.
최숙기, 「Rasch 모형을 활용한 국어 교사 평가 전문성 척도 개발」, 『청람어문교육』 56, 청람어문교육학회, 2015, 313-340면.
최지현, 「선택형 지필 평가의 한계와 가능성」, 『국어교육』 103, 한국국어교육연구회, 2000, 107-131면.
최지현, 「초, 중, 고등학교에서의 독서평가와 작문평가」, 『독서연구』 28, 한국독서학회, 2012, 142-174면.
최호성, 「선택형 평가와 수행적 평가의 상호관계에 관한 연구」, 『교육이론과 실천』 6, 경남대학교 교육문제연구소, 1996, 155-170면.
하상수, 「고등학교 국어과 총괄평가용 교사제작검사지 문항양호도 분석」, 경남대학교 교육대학원 석사학위 논문, 2009.
한국교육과정평가원, 『국어과 교사의 학생 평가 전문성 신장 모형과 기준』, 한국교육과정평가원, 2004.
Bond, T. & Fox, C. Applying the Rasch Model. New York: Routledge, 2007.
Boon, W. J, Staver, J, R, & Yale, M, S. Rasch Analysis in the Human Sciences. New York: Springer, 2013.
Reckase, M. Unifactor latent trait model applied to multifactor tests: results and implications. Journal of Educational and Behavioral Statistics, 4, 1979, 207-230면.

Smith, R. M. IPARM : Item and Persons Analysis with the Rasch Mod el. Chicago : MESA Press, 1991.
Wright, B. D. & Linacre, J. M.. BIGSTEPS : Rasch Analysis Computer Program. Chicago : MESA Press, 1992
Wright, B. D. & Mead, R. J. BICAL : Calibrating Rating Scale with th e Rasch Model. Research Memorandum No. 23, Statistical Lab., Department of Education, University of Chicago, 1976.
Wright, B. D., Rossner, D., Schulz, M. & Congdon, R. T. MSCALE: A Computer Program. Chicago : MESA Press, 1988.

【Abstracts】

A Rasch analysis of a high school Korean achievement test

Lee Jeong-ae • Shim Jae-woo

The main objectives of this study were to investigate item reliability, subject ability, and item fit indexes based on Rasch modeling. In particular, this study analyzed a Korean mid-term test results collected from 179 second year high school students, who attended a local high school in a mid-sized Korean city. The following were observed by the application of Winsteps 3.68, which is a software program for Rasch analyses: firstly, the person reliability as well as separation index was low, suggesting that the test itself was not successful in separating students along the ability continuum; secondly, all the test items met Rasch expectations. However, five test items had low levels of discrimination and failed to discriminate high ability students from low ability students. Next, personal ability estimates tended to be higher than item difficulty estimates, indicating that items, in general, were too easy for the students. Only one item was rightly targeted against the highest ability students. Finally, the 13 grammar items were evenly distributed in the hierarchy of item difficulties, forming four groups of items that were categorized as 'the most difficult items', 'the second most difficulty items', 'the third most difficult items' and 'the fourth most difficult items'. However, it was found that reading proficiency items were located

only in 'the third most difficult' and 'the fourth most difficult' groups of items. In fact, all of those reading proficiency items were from the text covered in the textbook and most of those reading proficiency items consisted of items that asked students to find simple facts based on the text.

One of the implications of this study is that test designers, including in-service teachers themselves, should reflect on the test-making practices so that valid and reliable items are used in achievement tests.

Key words : Rasch analysis, test reliability, achievement tests, item response theory, person reliability, hierarchy of items.

이 논문은 2016년 9월 30일에 투고되었으며, 2016년 11월 4일에 심사 완료되어 2016년 11월 9일에 게재가 확정되었음.