문항 유형에 따른 과학 능력 추정의 효율성 비교

박 정 · 홍미영
(한국교육과정평가원)

A Relative Effectiveness of Item Types for Estimating Science Ability in TIMSS-R

Park, Chung · Hong, Mi-young
(Korea Institute of Curriculum and Evaluation)

ABSTRACT

Recently, performance assessment that makes growing use of free response items in a large scale assessment has been emphasized. This study is an empirical examination of the effectiveness of free response items in comparison with multiple choice items. Using the information function in Item Response Theory (IRT) framework, item information of free response items and multiple-choice items from the Third International Mathematics and Science Study-Repeat (TIMSS-R) were obtained. Test information of the whole science area as well as each area of science contents was computed. On average, free response items yielded more information than multiple choice items, especially in earth science, physics, chemistry, and life science. This study also showed that free response items were appropriate for students in high science ability. Also, free response items estimated students' science ability more accurately than multiple choice items with smaller number of free response items.

Key Words: item type, performance assessment, free response items, information function, IRT, polytomous IRT model, TIMSS-R, estimates of science ability

I. 서 론

1990년대에 들어서면서부터 우리 나라뿐만 아니라 서구의 여러 나라에서도 교육현장에서 전통적인 평가 방식인 지필식 선다형 검사 문항을 사용하는 것을 지양하고 새로운 형태의 평가 방식을 시도하고 있다. 이는 선다형의 평가 방식이 학생들로 하여금 정답을 찾도록 유도함으로써, 학생들의 창의적인 사고와 적극적인 반응을 억제하고 수동적인 학습자로 만든다는 비판을 받기 때문이다(Kane & Mitchell, 1996; Linn, 1994).

교육 현장에서의 새로운 평가 방식의 강조는 대규모의 평가 연구에도 영향을 미쳐 1995년부터 시작된 제3차 수학 · 과학 성취도 국제비교 연구(the Third International Mathematics and Science Study : TIMSS)나 OECD 주관의 학업성취도 국제비교 연구(Programme for International Assessment: PISA) 같은 국제비교 성취도 평가 연구에서도 수행

*2001.10.10(접수) 2002.2.25(최종 통과)

형 평가 문항이 사용되고 있다. 최근 우리 나라도 수행 평가의 강조로 인하여 각 학교에서는 수행 평가가 시행되고 있으며, 국가 수준의 교육성취도 평가에서도 수행형 평가문항이 사용되고 있다(김명숙외, 1999).

그러나 수행 평가를 강조하고 있음에도 불구하고 이것이 전통적인 선다형 평가 방식에 비해 타당한 평가 방식이며, 기존의 선다형 평가 문항에 비해 어떠한 차이가 있는지를 경험적인 자료에 근거하여 비교한 연구는 그다지 많지 않다. 본 연구에서는 국제교육성취도평가협회(International Association for Evaluation of Education Achievement: IEA)의 주관 아래 1999년에 실시된 제3차 수학 · 과학 성취도 국제비교 반복 연구(Third International Mathematics and Science Study-Repeat : TIMSS-R)에 사용된 선다형과 수행형 평가 문항들을 사용하여, 수행형의 평가 문항과 선다형의 평가 문항이 학생들의 과학 성취도를 측정하는 효율성에 있어서 차이가 있는지를 과학 전체 및 내용 영역별로 나누어 분석하였다.

Ⅱ. 연구방법

1. 분석 자료

본 연구에 사용된 분석 자료는 TIMSS-R에서 사용된 과학 문항에 대한 학생들의 반응이다. TIMSS-R 연구는 연구 참가국 39개국의 8학년 학생들의 수학과 과학 성취도를 비교하기 위하여 구안된 연구로 1999년도에 시행되어, 2000년 12월에 그 결과가 발표되었다. 우리 나라는 전국 150개 학교 6,258명의 중학교 2학년 학생들이 연구에 참여하였다. TIMSS-R의 과학 성취도 평가를 위하여 사용된 문항은 선다형 평가 문항이 104개, 자유 반응형(수행형 평가) 문항이 39개였다(김성숙 외, 1999). TIMSS 연구에서는 선다형 평가문항 이외의 단답형과 서술형의 문항들을 총칭하여 자유 반응형으로 분류하고 있으며, 본 연구에서는 자유 반응형이라는 용어 대신에 수행형 평가 문항이라고 부른다.

각 수행형 평가문항은 하나 이상의 문항으로 이루어진 경우가 있어 실제 자료 분석에 사용된 문항의 개수는 더 많았다. Table 1에 과학 평가에 사용된 영역별 선다형 문항의 수와 수행형 문항의 수를 정리하였다. 수행형 문항에서 괄호 안 문항 수는 실제 자료 분석에 사용된 문항의 수이다.

2. 분석 방법

본 연구에서 선다형 문항과 수행형 문항의 효율성을 비교 분석하기 위하여 문항반응이론(item response theory)[1]의 정보함수(information function)를 사용하였다. 문항반응이론은 학생들의 능력(true ability)

Table 1. Number of items in each science content areas

Number of items	Scientific inquiry	Earth science	Life science	Physics	Chemistry	Environmental issues	Total
Multiple choice	9	16	27	32	13	7	104
Free response	3	5	11(13)*	12	3(5)	5(8)	39(46)
Total	12	21	38(40)	44	16(18)	12(15)	143(150)

* ( ) number of free response items for data analysis

1) 문항반응이론에 관한 자세한 설명은 성태제(2000), 이종성(1990)을 참조

을 표현하기 위하여 개발된 현대의 측정이론으로, 대표적인 장점은 문항의 난이도나 변별도가 검사를 치른 집단에 무관하게 항상 일정한 값을 제공할 수 있다는 점과 학생들이 매번 다른 유형의 검사를 치른다고 해도 자신의 고유한 능력 점수를 받게 된다는 불변성의 특성(invariance characteristic)을 가지고 있다는 것이다. 즉, 고전검사이론에 의한 채점 방식에서는 문항의 특성과 상관없이 문항을 맞으면 1점, 틀리면 0점으로 처리되는 대신, 문항반응이론에 의하여 채점할 때는 '문항의 특성'과 '학생자신의 능력'에 따라 각기 다른 점수를 받게 된다. 따라서 문항반응이론을 사용하여 자료를 분석하고 학생들을 평가할 때는 문항의 난이도를 고려한 학생의 능력 점수를 산출할 수 있다.

문항반응이론의 이러한 측정학적인 장점 때문에 TIMSS, 미국의 국가교육향상평가연구(National Achievement Educational Progress: NAEP), 그리고 PISA와 같은 대규모의 평가 연구에서 문항분석과 결과 처리에 사용되고 있다. 더구나 선다형 문항이 1점 혹은 0점으로 처리되는 데 반하여 수행형의 문항은 0점부터 다양한 점수를 줄 수 있기 때문에 이를 위한 점수 처리 방식이 요구되며, 이것을 해결할 수 있었던 측정 이론의 모형으로 다분문항반응이론을 사용하고 있다. 본 연구에서는 선다형 자료분석을 위해서는 이분(dichotomous) 문항반응이론 모형인 2-모수 로지스틱 모형을 사용하였고, 수행형 문항의 자료분석[2]을 위해서는 다분(polytomous) 문항반응이론 모형[3]의 하나인 일반화부분점수모형(Generalized Partial Credit Model: GPCM)을 사용하였다.

가. 선다형 자료분석에 사용된 위한 이분 문항반응이론 모형

문항반응이론모형은 정규오자이브 함수나 로지스틱 함수를 사용하여, 학생의 능력별로 각 문항에 대한 반응확률을 나타내는 것으로 선다형 문항을 분석할 수 있는 대표적인 이분 반응 모형은 (1)과 같다.

수식 (1)은 능력이 θ인 학생이 난이도가 b_j이고 변별도가 a_j이며 추측도가 c_j인 문항을 맞출($U_j=1$) 확률을 표시한 식이다. 수식 (1)에서 a_j는 문항 변별도, b_j는 문항 난이도 c_j는 문항 추측도라고 부르며, 이 의미는 통상적으로 의미하는 문항 난이도, 변별도, 추측도와 같다. 문항의 변별도인 a_j는 0보다 큰 값으로 크기가 커질수록 변별도가 높은 문항이며, 문항의 변별도 b_j는 음수부터 양수의 범위로 나타나며, 숫자가 커질수록 어려운 문항을 의미한다. 문항의 추측도 c_j는 0보다 큰 값으로 커질수록 추측하여 맞출 확률이 큰 문항임을 의미한다. 수식 (1)의 문항반응이론의 모형은 문항의 특성치인 난이도, 변별도, 추측도를 고려한 능력 모수치 θ의 함수로서, 학생의 능력에 따라서, 문항의 특성에 따라서 문항을 맞을 확률이 달라진다.

본 연구에 사용된 자료의 형태가 행렬표집(matrix sampling)으로 인하여 3-모수 모형을 사용하였을 경우 추정이 가능하지 않은 경우가 발생하기 때문에, 추측도의 영향력을 고려하지 않은 2-모수 모형을 사용하였다.

나. 수행형 자료분석을 위한 다분 문항반응이론 모형

수행형 문항과 같이 0점, 1점, 2점, 3점등으로 채점하는 경우에는 '맞음'과 '틀림'으로 채점하는 선다형 문항에서 사용되는 수식(1)을 확장한, 수식(2)과 같은 다분 문항반응이론 모형을 사용한다.

수식 (2)는 수행형 평가 문항과 같이 부분점수를 주게 되는 경우에 사용할 수 있는 대표적인 다분 문항

1

2) 수행형 평가문항 분석을 위한 문항반응이론의 활용방법은 박정(2001b) 참조
3) 다분 문항반응이론 모형에 대한 설명과 컴퓨터 프로그램의 사용법은 박정(2001a)참조

2

3, 4

반응이론모형인 일반화부분점수모형(GPCM, Muraki, 1992)을 나타낸 것이다. 수식 (2)와 수식(1)의 다른 점은 각 부분 점수(예컨대, 0, 1, 2점)에 해당하는 난이도가 있다는 것이다. 즉, 선다형 문항에서는 맞고 틀림으로만 구분되어 있어, 맞을 경우의 문항의 난이도 b가 1개만 있었던 것에 반하여, 수행형의 경우는 받을 수 있는 점수가 여러 개이므로 각 점수에 대응하는 난이도가 더 있어야 한다. 수식 (2)에서 a_j는 문항 j의 변별도 혹은 기울기 모수치이며, b_{jk}는 문항 j에서 k라는 점수를 받을 때의 어려운 정도를 의미하는 '문항 범주 난이도' 혹은 '문항 단계 난이도'이다.

다. 선다형 문항과 수행형 문항 비교를 위한 정보함수

수행형 문항과 선다형 문항과의 효율성을 비교할 때 문항반응이론에서 유용하게 사용할 수 있는 개념이 정보함수이다. 문항반응이론에서 정보함수는 각 문항이 혹은 각 검사가 얼마나 정확하게 수험자의 능력을 재고 있는가를 알려주는 지수로서, 수험자의 능력에 적절한 문항을 선정하여 좋은 검사를 만드는 준거가 되기도 한다. 따라서 각기 다른 유형의 검사가 어느 정도로 수험자의 능력을 정확하게 잴 수 있으며, 또한 어느 수준의 수험자의 능력을 정확하게 잴 수 있는 지의 효율성을 비교할 수 있는 지수로 사용할 수 있다.

문항반응이론에서 사용하는 정보함수는 문항정보함수와 검사정보함수를 들 수 있는 데, 문항정보함수는 문항이 제공하는 정보를 의미하며, 검사정보함수는 검사가 제공하는 정보를 말한다. 본 연구에서는 선다형 문항의 문항정보를 얻기 위해서 일반적으로 잘 알려진 (Hambleton & Swaminathan, 1985) 수식 (3)을 쓰고, 수행형의 문항의 문항정보를 위해서는 Donoghue(1994)의 수식 (4)을 사용한다.

수식 (3)과 (4)에서 첨자 j는 문항 j를 의미하며, k는 문항 j의 부분 점수를 의미한다. 수식 (3)에서 P_{j0} 은 문항 j를 틀릴 확률을 의미하고, P_{j1} 은 문항 j를 맞을 표시한 것이다. 따라서 수식 (3)에서 P_{j0} 는 1- P_{j1} 이며, 통상적으로 Q_j 로 표시한다. 수식 (3)과 (4)에서 구해진 문항정보 함수를 수식 (5)와 같이 합산하면 검사정보를 얻게된다.

5

수식 (5)에서 제시하는 검사정보 함수는 검사에 포함된 검사영역별로 혹은 검사유형별로 제공하는 정보의 차이를 비교할 수 있게 한다.

또한 검사정보 함수는 능력치 θ의 함수로서 수식 (6)과 같이 능력추정치의 정확도를 알려주는 추정의

표준오차(Standard Error of estimation: SE)로 표시할 수 있다.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (6)$$

따라서 검사 정보함수는 문항이나 검사가 어떤 능력의 피험자들에게 가장 정확한 정보를 제공하고 있는지를 알려주는 지수이다. 즉, 검사정보 함수는 어떤 능력의 학생들에게 그 유형의 검사가 적합한 것인지를 알 수 있는 정보를 제공한다. 따라서 제공되는 정보함수 값이 클수록 능력 추정치의 오차가 적어지고, 제공되는 정보함수 값이 적을수록 능력 추정치의 오차가 커진다. 이 정보함수는 능력 함수이므로 능력수준에 따라 각기 다른 값을 갖게 되므로, 피험자의 능력별로 추정치의 정확도가 어떻게 다른지에 관한 정보를 제공하는 지수로 사용한다. 본 연구의 자료 분석을 위해서 수행형의 평가 문항을 일반화 부분 점수 모형을 적용하여 분석할 수 있는 컴퓨터 프로그램 PARSCALE(Muraki & Bock, 1998)을 사용하였다.

Ⅲ. 분석 결과

선다형 문항과 수행형 문항에 따른 학생들의 과학 능력 추정의 효율성을 비교하기 위하여 우선 선다형 문항과 수행형 문항 전체 결과를 비교하고 난 후, 각 내용 영역별로 두 문항 유형의 결과를 비교 분석하였다.

1. 과학 전체 영역에서의 선다형 문항과 수행형 문항의 비교 분석

Fig. 1은 수식 (3)과 (4)를 사용하여 과학검사에 포함된 선다형 문항과 수행형 문항이 제공하는 정보의 양을 도표화한 것이다. 검사정보 함수값을 구하기 위하여 수식 (5)를 사용하되, 선다형 문항과 수행형 문항의 수가 다르므로 문항의 수가 주는 영향력을 교정하기 위하여 검사정보 함수의 평균값을 사용하였다.

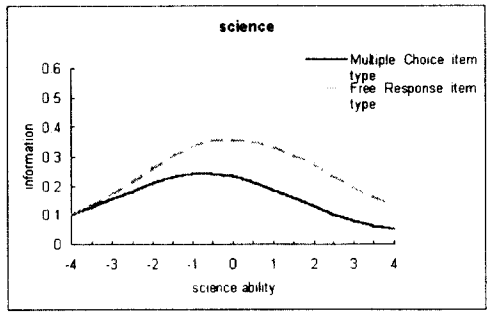

Fig. 1. Average IRT information of science by item type

Fig. 1은 TIMSS-R에 사용된 선다형 문항 104개와 수행형의 문항 46개가 제공하는 정보의 양을 비교한 것이다. Fig. 1의 가로축은 과학 능력을 의미하는 것으로, 문항반응이론의 능력치 θ값이다. 능력치 θ값은 표준정규 분포 상에서 표준화한 수치로 표현되어 -4부터 +4에서 도표화하였다. θ값은 편의상 평균이 50이고 표준편차가 10인 표준점수($T=50+10\theta$)로 선형 변환하여 해석하기로 한다. -4부터 +4까지의 능력 점수를 표준점수로 전환하면 10점부터 90점의 점수 분포로 해석할 수 있다. 이 표준점수는 우리 나라 학생들의 평균을 θ값으로는 0으로, 표준점수로는 50으로 척도화 한다는 의미이다. Fig. 1의 세로축은 수식 (5)에 의한 검사정보량이다. 검사 정보함수 도표는 가로축의 과학 능력에 따라서 제공하는 정보의 양이 다른 것을 알 수 있게 하고, 중간 지점에서 제공하는 정보의 양이 많고, 양극단으로 갈수록, 즉 능력이 낮거나 높은 학생들에게 제공하는 정보의 양이 상대적으로 적다는 것을 보여준다. 정보함수에서의 더 많은 정보를 제공한다는 의미는 학생들의 능력을 추정할 때 추정 오차가 적다는 의미로 정확하게 능력을 잰다는 의미이다.

Fig. 1에서 실선은 선다형 문항들의 정보함수를, 점선은 수행형 문항들의 정보함수를 나타낸다. 따라서 Fig. 1은 TIMSS-R의 수행형 문항들이 선다형 문항들보다 많은 정보를 제공하고 있는 것을 보여주고 있다. 이 말은 수행형의 문항이 선다형의 문항에 비하여 학생들의 과학 능력을 정확하게 추정할 수 있음을 의미한다. 또한 Fig. 1에서 정보의 양이 가장 많은 지

점은 선다형과 수행형이 거의 비슷하면서도 수행형의 경우에 0의 지점에서 선다형의 경우는 -1의 지점에서 가장 많은 정보를 제공함으로서 수행형의 문항이 과학 능력이 높은 학생들에게 좀 더 유용하다는 것을 보여주고 있다. 표준점수로 환산하여 보면, 수행형의 문항들은 우리 나라 학생들의 평균인 50점의 수준의 학생들을 정확하게 추정할 수 있고, 선다형의 문항들은 평균에 못 미치는 40점 수준의 학생들을 좀 더 정확하게 추정할 수 있는 문항들이라고 해석할 수 있다. 따라서 능력이 높은 학생들일 경우 수행형 문항이 선다형의 문항에 비하여 정확하게 능력을 추정할 수 있어, 능력이 높은 학생들에게는 수행형의 문항이 더 적합함을 보여주고 있다.

2. 과학 영역별 선다형 문항과 수행형 문항의 비교 분석

Table 2는 과학 내용 영역별로 선다형 문항과 수행형 문항의 난이도와 변별도의 평균을 정리한 것이다. 문항의 변별도는 문항 유형별로 거의 차이를 보이지 않았지만, 물리가 0.125의 차이를 보였고, 그 다음으로 지구과학이 0.109, 생물이 0.102, 화학과 환경 영역이 평균의 차이인 0.05정도로 수행형 문항의 변별도가 선다형에 비해 높았다. 과학의 본성 영역은 유일하게 수행형 문항보다 선다형 문항이 변별도가 높은 것으로 나타났다. 과학의 본성 영역의 경우 이와 같이 다른 결과를 가져오는 것은 내용 영역별 분석에서 후술하겠지만, 한 문항이 특이한 현상을 보임으로써 낮은 변별도의 문항으로 분류되고 있기 때문이다. 그러나 전반적으로 수행형 문항들이 선다형 문항들에 비하여 학생들의 능력을 조금 더 정확하게 측정할 수 있다는 앞의 결과와 같이 해석할 수 있다.

문항의 난이도는 수행형의 문항이 선다형 문항들에 비하여 0.47 만큼 어렵다. 문항반응이론에서의 난이도는 수식 (1)에서 b에 해당하는 지수로서 능력치 θ와 같은 척도 상에서 해석한다. b의 수치가 양수로 갈수록 어렵다고 해석한다. 따라서 0.47만큼 어렵다는 것은 본 연구에서의 표준점수 척도 상에서는 같은 학생들이 시험을 보았을 경우 수행형 문항의 평균이 선다형 문항의 평균보다 약 4.7점 정도 낮다고 해석할 수 있다. 내용 영역별로 살펴보면, 화학을 제외한 모든 영역에서 수행형 문항이 선다형 문항보다 어렵다는 것을 알 수 있다. 특히 물리, 과학의 본성, 생물 및 환경 영역의 수행형 문항이 선다형에 비해 어렵다는 것을 알 수 있다. 지구과학은 선다형과 수행형의 문항이 비슷한 난이도를 보이고 있고, 화학은 수행형 문항이 선다형에 비하여 약간 더 쉬움을 알 수 있다.

이는 다른 영역의 자유 반응형 문항의 경우에는 국제 비교 연구시 언어의 차이로 인하여 채점 기준이 매우 까다롭게 설정된 경우가 많아 정답율이 낮았으나, 화학 영역의 문항에서는 과학적 지식 이외의 다른 요소들을 채점 기준에 포함시켜 채점 기준이 느슨

Table 2. Average item difficulty and discrimination index of science content areas by item type

	Item type	Scientific inquiry	Earth science	Life science	Physics	Chemistry	Environmental issues	Average
Item discrimination	Multiple choice	0.589	0.586	0.568	0.525	0.653	0.598	0.58
	Free response	0.442	0.695	0.670	0.650	0.704	0.629	0.63
Item difficulty	Multiple choice	-0.488	-0.848	-1.051	-0.936	-0.317	-0.627	-0.71
	Free response	0.615	-0.728	-0.264	0.102	-0.898	-0.290	-0.24

하여 정답율이 높은 경우가 있었다. 예를 들면, 철근의 표면을 보호해야 하는 이유를 쓰는 것과 도금에 필요한 작업 시간이 단축된 새로 개발된 도금 처리 방법을 사용할 때 예상되는 결과를 2가지 쓰는 것으로 구성된 자유 반응형 문항의 경우, 과학적 지식 측면뿐만 노동자들의 고용 문제, 철강 회사의 생산성 등 예상 가능한 모든 응답에 대해서 점수를 부여하였으므로, 학생들의 정답율이 60~63%로 매우 높았다 (박정 외, 2000).

Fig. 2는 모든 내용 영역에서도 수행형 문항이 선다형 문항에 비해 제공하는 정보의 양이 많음을 보여주고 있다. 특히 지구과학과 생물, 물리, 화학 영역에서 수행형 문항이 선다형 문항들보다 훨씬 많은 정보를 제공하고 있음을 보여주고 있다. 수행형 문항이 선다형 문항들보다 많은 정보를 제공하고 있다는 의미는 이 영역들에서는 수행형의 문항들이 선다형의 문항들보다 학생들의 과학 능력을 좀 더 정확하게 측정하고 있다는 것을 의미하므로, 이러한 영역에서의 학생의 능력을 재려고 할 때는 수행형의 문항이 선다형의 문항보다 학생의 능력을 좀 더 정확하게 측정하는 데 적합하다고 볼 수 있으며, 이는 Table 2의 결과와도 일관된다.

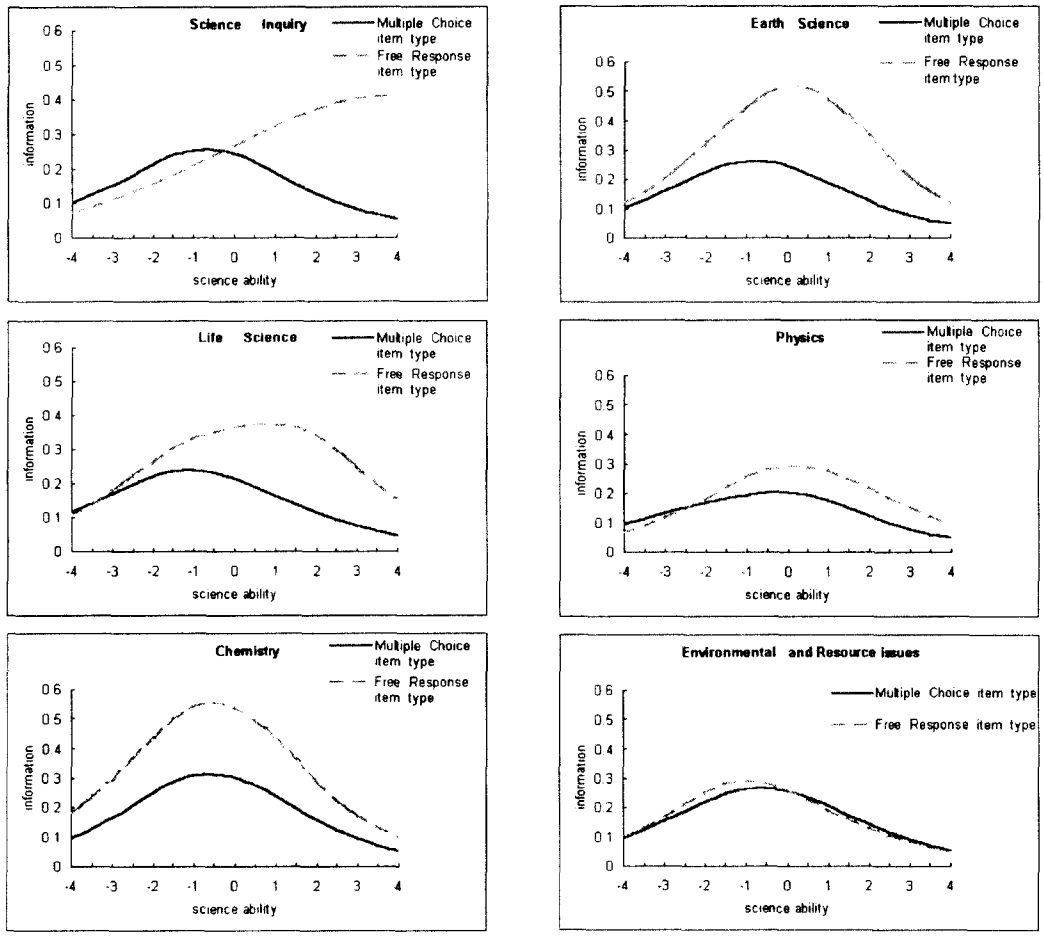

Fig. 2. Average IRT information of science content areas by item type

또한 Fig. 2는 수행형 문항의 정보함수가 선다형의 도표보다 오른쪽으로 치우치고 있음으로써, 수행형 문항이 능력이 높은 학생들에게 적합함을 시사하고 있다. 특히 지구과학, 생물, 물리, 화학 영역에서 수행형이 선다형 문항에 비하여 능력이 높은 학생들에게 훨씬 더 적합한 문항임을 나타내고 있다.

그러나 Fig. 2에서 과학의 본성 영역의 도표는 특이한 형태를 보이고 있다. 과학의 본성 영역의 경우는 수행형 문항들이 제공하는 정보의 양을 표현하는 도표가 다른 영역에 비해 다른 형태를 보이고 있다. 이는 과학의 본성 영역에 포함되어 있는 문항(문항 X03)의 영향이다. 이 문항은 2점 만점으로 부분 점수(1점)를 허용한 경우로서, 문항의 변별도는 0.337로서 수행형 문항의 평균에 미치지 못하며, 이로 인하여 과학의 본성 영역의 전체 평균 변별도가 낮아졌다. 또한, 2점 단계의 난이도는 3.44로서 우리 나라 학생들에게 상당히 어렵다고 해석할 수 있으며, 이로 인하여 Fig. 2에서 난이도의 수준이 오른쪽으로 크게 올라가는 양태를 보이고 있다. 좀 더 구체적으로 문항을 분석하기로 한다.

문항 X03의 분석 결과[4)]

과학의 본성 영역의 수행형 문항에서 특이한 형태를 나타내는 X03의 문항은 심장박동 실험 문항으로 아래와 같다.

> 운동이 끝난 후 심장박동이 정상으로 돌아오는데 걸리는 시간을 조사하려고 한다. 필요한 기구와 재료는 무엇이며, 어떤 과정을 거쳐야 하는지에 대하여 쓰시오.

이 문항의 채점 기준표는 다음과 같았다.

> 2점: 기준으로 설정한 세 가지의 준거(운동 전 맥박 측정, 운동 실시, 운동 후 맥박이 정상으로 될 때까지의 시간 측정)를 모두 기술
> 1점: 한 두 가지의 준거를 결여한 경우
> 0점: 무응답 또는 준거와 무관한 반응

이 문항에 응답한 1535명의 학생 중 29.7%가 2점을 받았고, 16.6%가 1점을 받았다. 0점을 받은 학생은 53.8%이며, 이 중 무응답은 19.2%, 기준으로 제시한 준거와 무관한 반응을 하여 0점을 받은 학생은 34.6%이었다. 다분 문항반응이론 모형을 적용하여 분석한 결과, 1점을 받을 난이도는 1.55이고, 2점을 받을 난이도는 3.44로 상당히 어려운 문항이지만, 문항의 기울기는 0.337로서 상당히 낮은 것으로 나타났다. 문항의 기울기는 문항의 변별도와 비슷한 개념으로 이 문항의 변별도는 낮아서 학생들의 능력수준을 제대로 변별하기 어려운 문항이라고 할 수 있다.

0점, 1점, 2점을 받을 확률을 표시한 문항범주특성곡선을 그리면 Fig. 3과 같다.

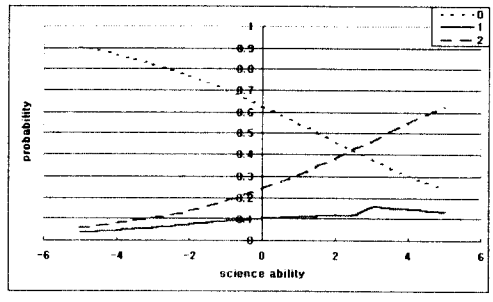

Fig. 3. X03 Item category characteristic curves

Fig. 3의 왼쪽의 상단에 나타나는 점선(---)은 0점, 그림의 하단에 나타나는 실선은 1점, 오른쪽으로 상승하는 곡선은 2점을 받을 확률을 나타낸다. 일반적으로 능력이 낮은 학생이 0점, 능력이 중간 정도인 학생들이 1점, 그리고 능력이 높은 학생들이 2점을 받아야 한다는 것이 기본적인 가정인데, 이 문항은 능력이 낮으면 0점을 받고, 능력이 높은 학생은 2점을 받지만, 중간 수준의 학생들은 1점을 받을 가능성이 거의 없다. 이 문항은 중간 수준의 능력을 가진 학생들을 제대로 변별할 수 없는 문항으로 대부분의 학생이 0점이나 2점을 받으므로 부분 점수인 1점을 주는 것이 별 의미가 없다. 부분점수 1점을 받은 학생들이 중간 수준의 능력을 보이는 것으로 나타나고 있질 않

4) 본 논문에 기술된 X03 문항 분석은 박정(2001b)의 내용을 빌려왔다.

다. 이와 같은 일이 발생하는 것은 문항이 지나치게 어려워서 무응답이나 잘못된 반응을 한 학생들이 많기 때문일 수도 있고, 부분 점수를 주는 1점의 채점 기준표가 적절하지 않을 수도 있으며, 채점의 과정에서 1점을 받을 학생이 0점으로 처리되었을 수도 있어, 채점 기준표를 점검하거나 채점 과정을 점검하는 일이 필요하다.

이 문항의 경우 국제적으로도 2점을 받은 학생들의 비율이 12.0%로 매우 낮으며, 우리 나라 학생들의 평균 정답율이 영국과 싱가포르에 이어 세 번째로 높은 편이다(박정 외, 2001). 따라서, 이 문항의 경우에는 우리 나라 학생들에게만 특별히 어렵다고 간주하기보다는 지필식 검사로 학생들이 자신의 생각을 충분히 측정하기에 적절하지 못한 문항인 것으로 볼 수 있다. 이 문항의 내용은 우리 나라에서는 초등학교 6학년에서 다룬 내용이므로, 학생들이 내용 지식에 있어서는 친숙하지만, 체계적으로 답을 직접 구성하는 것에 어려움을 겪은 것으로 간주된다(박정 외, 2000). 예를 들면 1점을 받은 우리 나라 학생들 중에서 '실험 대상인 사람에게 5분 동안 자전거를 타게 한 다음, 멈추게 한다. 그리고 맥박이 정상으로 돌아올 때까지 걸리는 시간을 측정한다' 라고 답한 비율이 10%이며, 0점을 받은 학생들 중 '운동을 시킨 다음 그 사람의 맥박을 측정한다'라고 답한 비율이 15%에 이르는 등 2점에 해당하는 준거 중 일부만을 쓴 학생들의 비율이 높았는데, 이 중에는 자신이 알고 있는 것을 문장으로 표현하는 과정에서 일부를 누락하였을 수도 있다. 이러한 현상은 문항 X03과 같이 긴 응답을 요하는 문제 해결 과정에서 흔히 발생하는데, 발성 사고법(think -aloud method)이나 회상적 면담법(retrospective interview)을 통하여 학생들의 실제 사고과정을 보다 정확하게 알아보는 것이 필요하다(홍미영, 1995).

Ⅳ. 결론 및 제언

본 연구에서는 수행형 평가문항이 선다형 평가문항에 비하여 학생들의 능력을 추정하는 데 효율적인가를 살펴보기 위하여 양질의 문항이라고 할 수 있는 제 3차 수학 과학 성취도 국제비교 반복 연구에 사용된 선다형과 수행형 평가문항들을 사용하여, 수행형의 문항이 선다형의 문항에 비해 어떤 능력의 학생들에게 어느 정도나 유용한가를 분석하였다. 또한 수행형 평가문항이 선다형 평가문항에 비하여 학생들의 능력을 어떻게 다르게 잴 수 있는지를 내용영역별로 분석하여, 수행형 평가문항의 효율성을 살펴보았다.

연구 결과, 전반적으로 수행형 평가문항이 선다형 평가문항에 비하여 학생들의 능력을 정확하게 추정할 수 있음을 보여주었다. 특히 환경이나 과학의 본성 영역에 비하여 지구과학, 생물, 물리와 화학 영역에서 수행형의 문항이 선다형의 문항에 비하여 학생들의 능력을 정확하게 측정하였다. 또한 수행형의 문항들이 능력이 높은 학생들의 능력을 추정할 때 적은 추정오차를 유발하여, 능력이 높은 학생들에게는 수행형의 문항을 사용하는 것이 더 적절한 방식임을 보여주었다. 또한 수행형의 문항이 선다형의 문항에 비하여 우리 나라 학생들에게는 조금 더 어려운 문항 형태임을 알 수 있었다.

결과적으로, 수행형의 문항이 선다형의 문항에 비하여 학생들의 능력을 더 정확하게 추정함으로서 효율성이 높은 것으로 나타났다. 또한 선다형의 문항에 비하여 적은 수의 문항으로 학생들의 능력을 추정함에도 불구하고, 더 정확하게 능력을 추정함으로서 수행형 평가문항의 효율성을 입증하고 있다. 그러나 본 연구에서는 수행형의 평가문항에 포함된 단답형의 문항들과 서술형의 응답형을 분리하여 분석하지는 못하였다. 이는 본 연구에 포함된 수행형의 문항이 선다형의 문항에 비해 상대적으로 적은 수로 수행형의 문항을 다시 단답형과 서술형으로 구분하여 분석하기 어려웠기 때문이다. 추후 연구에서는 수행형의 문항 형태에서 단답형과 서술형의 문항군으로 분리하여 분석하여 수행형의 문항에서도 단답형의 문항형태와 서술형의 문항 형태가 어떻게 다른 결과를 보이는지를 분석하는 것이 필요하다.

적 요

본 연구는 수행 평가 방식의 중요성이 대두됨에 따

라 실지로 수행형 평가 문항이 전통적인 선다형 평가 문항에 비해 어떤 학생들에게 어떻게 유용한지를 분석한 연구이다. 이를 위하여 제 3차 수학 과학 성취도 국제비교 반복 연구에 사용된 선다형과 수행형 평가 문항들을 내용영역별로 분석하여, 수행형의 평가문항과 선다형의 평가문항의 효율성을 비교 분석하였다. 자료분석 결과 수행형의 문항이 선다형의 문항에 비하여 학생들의 능력을 더 정확하게 추정함으로써 효율성이 높은 것으로 나타났다. 특히 환경이나 과학의 본성 영역에 비하여 지구과학, 생물, 물리와 화학 영역에서 수행형의 문항이 선다형의 문항에 비하여 학생들의 능력을 더 정확하게 추정하고 있음을 보여주었다. 또한 선다형의 문항에 비하여 적은 수의 문항으로 학생들의 능력을 추정함에도 불구하고 추정오차가 적어, 적은 수의 수행형 문항으로도 학생들의 능력을 정확하게 추정할 수 있음을 시사하고 있다.

감사의 글

본 논문은 한국교육과정평가원 2000년도의 TIMSS-R 연구를 위하여 수행한 내용을 수정 보완한 것이다. 본 연구에 사용된 TIMSS-R 자료를 수집하고 정리해 주신 한국교육과정평가원의 연구진들과 자료 수집에 도움을 주신 관계자들께 감사를 드린다.

참 고 문 헌

김명숙, 황혜정, 최승현, 이명희, 강운선, 박선미, 김재춘, 박정, 설현수(1999). 국가수준 교육성취도 평가 연구 II: 사회 · 수학 영역 예비검사 문항개발 및 현장 적용 연구. 한국교육과정평가원. 연구보고 RRE 99-9-1.

김성숙, 임찬빈, 이춘식, 유준희, 서동엽(1999). 제 3차 수학 · 과학 성취도 국제비교 연구 (TIMSS-R) 국내평가 결과 분석 연구. 한국교육과정평가원. 연구보고 RRE 99-7-1.

박 정(2001a). 다분문항반응이론모형. 서울: 교육과학사

박 정(2001b). 문항반응이론을 활용한 수행형 평가문항 분석 방법. 교육학 연구, 39(2), 215-232.

박 정, 홍미영, 김성숙, 전현정(2000). 제3차 수학 · 과학 성취도 국제비교 연구 (TIMSS-R) 국내평가 결과 분석 연구 II. 한국교육과정평가원. 연구보고 RRE 2000-7.

박 정, 홍미영, 나귀수, 김성숙(2001). 제3차 수학 · 과학 성취도 국제비교 연구 (TIMSS-R) 공개문항 분석 자료집. 한국교육과정평가원. 연구자료 ORM 2001-9

성태제(2000). 문항반응이론의 이해와 적용. 교육과학사: 서울.

이종성(1990, 번역). 문항반응이론과 적용. 대광문화사: 서울.

홍미영(1995). 문제와 문제해결자의 특성이 화학 문제해결에 미치는 영향, 서울대학교 박사학위논문.

Donoghue, J. R.(1994). An emprical examination of the IRT information of polytomously scored reading items under the GPCM. *Journal of Educational Measurement*, 31, 295-311.

Hambleton, R. K., & Swaminathan, H.(1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

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

Linn, R. L.(1994). Performance assessment: policy promises and technical measurement standards. *Educational Researchers*, 23(9).

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

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