韓國數學敎育學會誌 시리즈 A <數學敎育>
2004. 5. 제 43권, 제 2호, 151-162.

J. Korea Soc. Math. Ed. Ser. A: The Mathematical Education
May. 2004, Vol. 43, No. 2, 151-162.

수행형 문항과 선다형 문항의 수학적 능력 추정 효율성 비교[1)]

박 정 (한국교육과정평가원)
박 경 미 (홍익대학교)

I. 서 론

최근 몇 년 동안 수학교육 평가에서 일종의 '화두' 역할을 해 온 것은 수행평가이다. 수행평가는 기존의 평가가 지니고 있는 여러 문제점을 일소할 수 있는 대안적인 평가로 주목받기도 하였고, 시간과 노력면에서 고비용 평가이면서도 결과의 객관성 결여라는 취약점 때문에 비판받기도 하였다. 수행평가는 정책적인 의무 조항 혹은 권장 사항으로 학교 현장에 급속도로 보급되었으나, 현재는 대부분 자율적인 실시로 후퇴한 상태이다. 이에 따라 수행평가에 대한 논의도 소강 국면에 접어들었지만, 수행평가는 여전히 많은 연구자와 교사의 관심사가 되고 있다.

수행평가에 대한 관심은 비단 우리나라 만의 현상은 아니다. 서구의 여러 국가 역시 전통적인 평가 도구인 지필식 선택형 검사를 지양하고 새로운 형태의 평가 방식을 시도하고 있다. 기존의 평가 도구로는 최근 수학교육에서 강조하고 있는 창의적 문제해결력이나 의사소통 능력과 같은 고차적인 수학적 사고력을 적절하게 측정하기 어려우며, 이러한 한계를 극복할 수 있는 대안으로 수행평가가 부상하게 된 것이다(Kane & Mitchell, 1996; Linn, 1994).

수행평가는 각급 학교나 교사 차원의 평가 뿐 아니라 대규모의 국제비교 연구에도 영향을 미쳐, 1995년과 1999년에 실시된 제3차 수학·과학 성취도 국제비교 연구(the Third International Mathematics and Science Study, 이후 TIMSS로 지칭)와 그 반복연구(TIMSS-Repeat, 이후 TIMSS-R로 지칭)나 OECD 주관의 학업성취도 국제비교 연구(Programme for International Student Assessment, 이후 PISA로 지칭)의 검사 도구도 수행형 문항을 다수 포함하고 있다. 뿐만 아니라 우리나라의 국가수준 학업성취도 평가에서도 수행형 문항을 사용하고 있다.

이와 같이 수행평가가 강조되고 각급 학교의 평가나 국제비교 연구에서 적극적으로 사용되고 있음에도 불구하고 실험 결과를 통계적으로 분석하여 수행형 문항과 선택형 문항의 여러 가지 특성을 비교한 연구는 그다지 많지 않다(권오남 외, 1999; 성태제 외, 1999; 유현주, 1998 등). 선택형 문항과 수행형 문항에 대한 통계적인 분석을 토대로 특성을 치밀하게 분석하기보다는 대개 상식에 의거한 통념에 따라 문항의 특성을 이해하고 있다. 예컨대 대부분의 사람들은 수행평가가 학생들의 능력을 보다 정확하게 변별할 수 있는 강점을 지니고 있지만 객관성이 결여된 평가라는 인식을 가지고 있으나, 체계적인 연구 분석에 의한 결론이라고 보기는 어렵다.

이에 본 연구는 국제교육성취도평가협회(International Association for Evaluation of Education Achievement: IEA)의 주관 하에 실시된 제3차 수학·과학 성취도 국제비교 반복연구(TIMSS-R)의 결과를 대상으로 선다형과 수행형 문항이 우리나라 학생들의 수학 성취도를 측정하는 효율성에 있어서 차이가 있는지 비교, 분석하고자 한다. 대규모로 시행된 국제적인 비교 연구의 실제적인 자료를 근거로 문항 유형에 따른 특성을 비교함으로써, 수행평가에 대해 사람들이 가지고 있는 인식의 적합성을 점검하는 기회가 될 수 있을 것이다.

---

1) 본 논문은 한국교육과정평가원이 TIMSS-R 연구를 위하여 수행한 내용을 수정·보완한 것이다.

* 2003년 7월 투고, 2004년 4월 심사 완료.
* ZDM 분류 : C83
* MSC 2000 분류 : 97C40
* 주제어 : 문항 유형, 수행평가, 수행형 문항, 선다형 문항, 검사정보함수, 문항반응이론, 수학적 능력 추정

II. 연구 방법

1. 분석 자료

본 연구는 선다형 문항과 수행형 문항의 효율성을 비교하기 위해 TIMSS-R의 검사결과를 분석하였다. TIMSS-R은 전 세계 38개국이 참여한 대규모의 국제비교 연구로, 8학년 학생들의 수학과 과학 성취도를 비교하기 위하여 구안되었다. 우리나라의 경우 전국적으로 표집된 150개 학교의 중학교 2학년 학생 6,258명이 1999년 2월에 실시된 TIMSS-R 검사에 응하였다(김성숙 외, 1999).

문항의 유형별로 볼 때 TIMSS-R의 수학 검사는 선다형[2] 문항 126개, 자유반응형 문항 37개로 구성되어 있다. TIMSS-R에서는 선다형 문항 이외의 단답형과 서술형의 문항을 포괄하여 자유반응형으로 총칭하고 있으며, 본 연구에서는 자유반응형이라는 용어 대신에 수행형[3] 문항이라고 명명하였다 (TIMSS-R의 선다형 문항과 수행형 문항의 예시는 <부록> 참고). <표 1>은 TIMSS-R의 수학 검사에 포함된 선다형 문항과 수행형 문항의 개수를 영역별로 제시한 것이다.

<표 1> 수학 내용 영역에 따른 TIMSS-R 문항의 유형별 개수

	분수와 수감각	대수	측정	비례	기하	자료표현 및 해석, 확률	합계
선다형 문항	44	21	14	7	22	18	126
수행형 문항	11	10	9	4	1	2	37
합계	55	31	23	11	23	20	163

2) 선택형에는 진위형, 연결형, 선다형 등 여러 가지 유형이 있으므로, 선택형은 선다형에 비해 광범위한 개념이다. TIMSS-R의 선택형은 모두 5개의 답지 중 하나를 고르는 방식이므로 본 고에서는 선다형이라고 명명하였다.

3) 수행형 문항은 광의로 해석할 수도 있고, 협의로 해석할 수도 있다. 광의로 해석할 경우 수행형은 피험자의 구성적 반응을 요구하는 문항이므로, 단답형과 서술형을 모두 포괄한다. 그러나 반응의 자유도가 높아야 한다는 조건을 첨부하는 협의의 관점에는 수행형이 서술형만 포함하기도 한다. 본 연구에서는 광의의 관점에서 단답형과 서술형을 모두 수행형으로 보았다.

2. 분석 방법

본 연구에서는 선다형 문항과 수행형 문항에 의한 능력 추정의 효율성을 비교 분석하기 위하여 문항반응이론(item response theory)[4]의 정보함수(information function)를 사용하였다. 문항반응이론은 학생들이 보유하고 있는 능력을 보다 정확하게 알아내기 위하여 개발된 현대의 측정이론이다. 문항반응이론을 고전적인 검사이론과 비교할 때의 장점은 문항의 난이도나 변별도가 검사를 치른 집단에 무관하게 항상 일정한 값을 제공할 수 있고, 학생들이 매번 다른 유형의 검사를 치른다고 해도 자신의 고유한 능력 점수를 받게 된다는 불변성이다. 문항반응이론의 이러한 측정학적인 장점 때문에 TIMSS, PISA, 미국의 국가교육향상평가연구(National Achievement Educational Progress, NAEP)와 같은 대규모의 평가 연구에서 문항 분석과 결과 분석에 사용되고 있다.

선다형 문항은 정답과 오답 여부에 따라 1점 혹은 0점으로 처리되기 때문에 그 결과 분석을 위해서 이분(dichotomous)문항반응이론 모형인 2-모수 로지스틱 모형을 사용하였다. 수행형 문항은 0점과 만점 사이에 부분점수를 부여하기 때문에 그 결과 분석을 위해서는 다분(polytomous)문항반응이론 모형을 사용해야 한다. 본 연구에서는 다분문항반응이론 모형의 일종인 일반화부분점수모형(Generalized Partial Credit Model: GPCM)을 사용하였다[5].

가. 선다형 문항 분석을 위한 이분문항반응이론 모형

문항반응이론 모형은 학생의 능력에 따라 각 문항에 대한 반응 확률을 나타내는 것으로 본 연구의 선다형 문항을 분석할 때 사용한 이분문항반응 모형은 (1)의 2-모수 로지스틱 모형이다.

$$P_j(U_j=1 \mid \theta) = \frac{e^{1.7a_j(\theta_i-b_j)}}{1+e^{1.7a_j(\theta_i-b_j)}} \quad (1)$$

수식 (1)에서 구한 값은 능력이 θ인 학생이 주어진

4) 문항반응이론에 관한 자세한 설명은 성태제(2000), 이종성(1990) 참조.

5) 수행형 문항 분석을 위한 문항반응이론의 활용 방법이나 컴퓨터 프로그램의 사용에 대한 자세한 논의는 박정(2001a, 2001b) 참조.

변별도와 난이도의 문항에 옳게 답할 ($U_j=1$) 확률 P 를 구하는 식이다. 수식 (1)에서 a_j는 문항 변별도, b_j는 문항 난이도라고 지칭하며, 이는 통상적인 의미의 문항의 난이도, 변별도와 동일하다. 문항의 변별도인 a_j는 양수로 값이 커질수록 변별도가 높은 문항이며, 문항의 난이도 b_j는 음수값과 양수값을 모두 가질 수 있는데 값이 커질수록 어려운 문항임을 의미한다.

나. 수행형 문항 분석을 위한 다분문항반응이론 모형

수행형 문항과 같이 0점, 1점, 2점, 3점 등으로 채점하는 경우에는 '맞음'과 '틀림'으로 채점하는 선다형 문항에서 사용되는 수식 (1)을 적용할 수 없다. 수식 (2)는 일반화 부분점수 모형(GPCM, Muraki, 1992)을 나타낸 것인데, 수행형 문항과 같이 부분점수가 설정되어 있는 경우에 사용할 수 있는 다분문항반응이론의 대표적인 모형이다.

$$P_{jk}(U_{jk}=k \mid \theta)=\frac{e^{\sum_{v=1}^{k} a_j(\theta-b_{jv})}}{\sum_{c=1}^{m_j} e^{\sum_{v=1}^{c} a_j(\theta-b_{jv})}}, \quad k=1,2,\{2\},m_j \qquad (2)$$

수식 (2)와 수식 (1)의 차이점은 각 부분점수에 해당하는 난이도가 존재한다는 점이다. 즉, 선다형 문항에서는 정답과 오답으로 구분되어 있어, 맞는 경우에 대한 문항의 난이도 b가 하나만 있었던 것에 반하여, 수행형의 경우는 피험자가 받을 수 있는 점수가 여러 개이므로 각 점수에 대응하는 난이도가 존재하게 된다. 수식 (2)에서 a_j는 문항 j의 변별도이며, b_{jk}는 문항 j에서 k라는 점수를 받을 때의 어려운 정도를 의미하는 '문항 범주 난이도' 혹은 '문항 단계 난이도'를 말한다.

다. 선다형 문항과 수행형 문항의 비교를 위한 정보함수

수행형 문항과 선다형 문항의 효율성을 비교할 때 유용하게 사용할 수 있는 개념이 정보함수(information function)이다. 문항반응이론에서의 정보함수는 고전검사이론에서의 신뢰도에 대응되는 개념으로, 피험자의 능력에 적절한 문항을 선정하여 좋은 검사 도구를 만드는 준거가 된다. 정보함수는 각 문항이나 검사가 얼마나 정확하게 피험자의 능력을 측정하는지를 알려주는 지수이다. 따라서 문항 유형에 따라 정보함수값을 구한 후 비교하면 어떤 유형의 문항이 피험자의 능력을 보다 정확하게 측정할 수 있으며, 어느 수준의 피험자의 능력을 보다 정확하게 잴 수 있는지 그 효율성을 비교할 수 있다.

문항반응이론에서 사용하는 정보함수에는 문항이 제공하는 정보를 의미하는 '문항정보함수'와 검사가 제공하는 정보를 말하는 '검사정보함수'가 있다. 대개 하나의 문항으로부터 계산된 정보의 양은 충분하지 않으며, 단일한 문항으로 피험자의 능력을 추정하는 경우는 드물기 때문에, 여러 문항으로 이루어진 검사로부터 피험자의 능력을 추정하는 경우가 많다.

본 연구에서는 선다형 문항의 정보를 얻기 위해서 일반적으로 많이 사용되는 Hambleton와 Swaminathan (1985)의 수식 (3)을 쓰고, 수행형 문항의 정보를 얻기 위해서는 Donoghue(1994)의 수식 (4)를 사용하였다.

$$I_j(\theta)=(1.7a_j)^2 P_{j0}(\theta)\,P_{j1}(\theta) \qquad (3)$$

$$I_j(\theta)=(1.7a_j)^2\left[\sum_{k=0}^{m_j-1} k^2 P_{jk}(\theta)-\left(\sum_{k=0}^{m_j-1} kP_{jk}(\theta)\right)^2\right] \qquad (4)$$

수식 (3)과 (4)에서 a_j는 앞에서와 마찬가지로 문항의 변별도를 의미한다. 또한 첨자 j는 문항 j를 의미하며, k는 문항 j의 부분점수를 의미한다. 수식 (3)에서 P_{j0}는 문항 j에 대해 틀린 반응을 보일 확률을, P_{j1}은 문항 j에 대해 맞는 반응을 나타낼 확률을 표시한 것이다. 따라서 P_{j0}는 $1-P_j$이며, 통상적으로 Q_j로 표시하기도 한다. 수식 (3)과 (4)에서 구해진 문항정보함수를 합산하면 검사정보함수를 얻게 된다. 다음 수식 (5)에 제시된 검사정보함수는 문항정보함수의 합으로, 검사 문항의 내용영역이나 문항 유형별로 제공하는 정보의 차이를 비교할 수 있게 한다.

$$I(\theta)=\sum_{j=1}^{n} I_j(\theta) \qquad (5)$$

한편 검사정보함수는 능력값 θ의 함수로서 수식 (6)과 같이 능력추정값의 정확도를 알려주는 추정 표준오차(Standard Error of estimation: SE)로 표시할 수 있다.

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

표준오차와 검사정보함수의 제곱근은 서로 역수의 관

계이다. 제공되는 정보함수의 값이 클수록 능력을 정확하게 추정할 수 있으므로 추정값의 표준오차가 작아지고, 제공되는 정보함수값이 작을수록 능력추정값의 표준오차가 커진다. 또 역으로 표준오차가 작으면 피험자의 능력을 보다 정확하게 추정할 수 있기 때문에 정보함수값이 커진다.

검사정보함수는 능력값 θ의 함수로 능력수준에 따라 각기 다른 값을 갖게 되므로, 피험자의 능력별로 추정값의 정확도가 어떻게 다른지를 비교할 수 있다. 따라서 검사정보함수는 검사가 어떤 능력의 피험자들에게 가장 정확한 정보를 제공하는지를 알려주는 지표가 될 수 있다.

III. 분석 결과

문항의 유형의 따라 피험자의 능력을 추정하는 정도를 파악하기 위하여 우선 선다형 문항과 수행형 문항의 분석 결과를 전체적으로 비교한 후, 각 내용 영역에 대하여 문항유형별 결과를 분석하였다. 자료 분석을 위해서 수행형 문항을 일반화 부분점수 모형을 적용하여 분석할 수 있는 컴퓨터 프로그램 PARSCALE(Muraki & Bock, 1998)을 사용하였다.

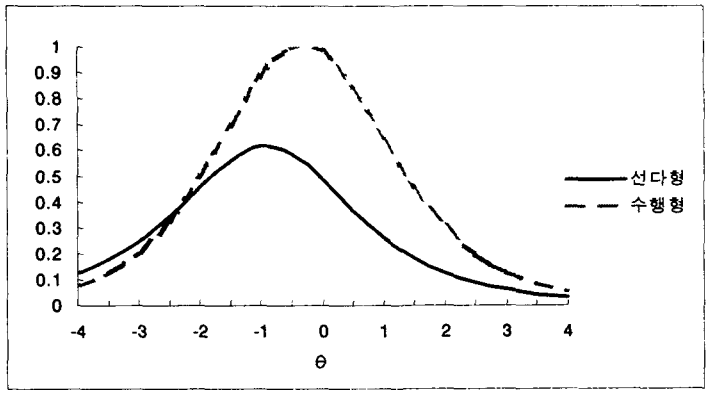
1. 수학 전체 영역에 대한 선다형 문항과 수행형 문항의 비교 분석

TIMSS-R 수학 검사에 포함된 선다형 문항과 수행형 문항이 제공하는 정보의 양을 수식 (3)과 (4)로 계산하였다. 검사정보함수값을 구하기 위하여 수식 (5)를 사용하였는데, 선다형 문항과 수행형 문항의 수가 다르므로 문항의 개수에 따른 영향력을 교정하기 위하여 검사정보함수값의 평균을 사용하였다(박정 외, 2000).

아래 <그림 1>은 TIMSS-R에 포함된 선다형 문항 126개와 수행형의 문항 37개가 제공하는 정보의 양을 그래프화한 것이다. 그래프에서 가로축은 문항반응이론의 능력값 θ로 수학적 능력을 의미한다. 능력값 θ는 표준정규분포로 표준화한 수치로 -4부터 +4까지의 값을 갖는다. θ는 편의상 평균이 50이고 표준편차가 10인 표준점수($T=50+10\theta$로 선형변환하여 척도화하였다. 즉 -4부터 +4까지의 능력값 θ를 표준점수로 전환하여 10점부터 90점의 분포가 되도록 하였다. 그래프의 세로축은 식 (5)에 의해 구해진 검사정보량이다.

위의 그래프에서 실선은 선다형 문항의 검사정보함수 곡선을, 점선은 수행형 문항의 검사정보함수 곡선을 나타낸다. 전반적으로 볼 때, 점선 그래프가 실선 그래프보다 위에 위치하고 있으므로 수행형 문항이 선다형 문항보다 더 많은 정보를 제공한다는 것을 알 수 있다. 피험자에 대해 더 많은 정보를 제공한다는 것은 피험자의 능력을 추정할 때 오차가 적다는 의미이므로, 수행형 문항이 선다형 문항에 비하여 학생들의 수학적 능력을 정확하게 추정한다고 볼 수 있다.

검사정보함수 곡선은 종 모양이므로 수학적 능력이 중간 정도인 피험자에 대해 제공하는 정보의 양이 많고, 수학적 능력이 낮거나 높은 양극단으로 갈수록 피험자에



<그림 1> 수학 전체 영역에 대한 선다형과 수행형 문항의 검사정보함수 곡선

대해 제공하는 정보의 양이 감소한다. 수행형 문항의 검사정보함수 곡선에서 최대값을 갖는 지점은 θ값이 0일 때, 즉 표준점수로 50점이므로 수행형 문항은 평균 정도 학생들의 능력을 보다 정확하게 추정할 수 있다. 또 선다형 문항의 검사정보함수 곡선은 θ값이 -1인 지점에서 최대값을 갖는다. 이 지점은 표준점수로 40점이므로, 선다형 문항은 평균보다 약간 낮은 수준의 학생들의 능력을 보다 정확하게 추정할 수 있다.

한편 능력값 θ가 -2.3 이하에서는 선다형 문항의 정보함수값이 수행형 문항의 정보함수값보다 크다. 수학적 능력이 아주 낮은 학생들은 거의 대부분 수행형 문항을 해결하지 못하므로 변별력이 낮아지게 되고 따라서 선다형이 학생들의 능력 추정에 더 적절한 방식이 될 수 있다. θ가 -2.3을 기준으로 반전이 일어나 이보다 큰 모든 능력값에서 수행형 문항의 정보함수값이 더 크므로 전반적으로 선다형 문항보다 수행형 문항으로 피험자의 능력을 추정하는 것이 더 정확하다고 볼 수 있다.

2. 내용영역별 선다형 문항과 수행형 문항의 비교 분석

가. 변별도와 난이도

TIMSS-R의 내용영역별로 선다형 문항과 수행형 문항의 난이도와 변별도의 평균을 구하였다. <표 2>에 제시된 문항의 변별도와 난이도는 수식 (1)과 (2)의 a_j와 b_j에 해당한다. 변별도는 값이 클수록 변별력이 높으며, 변별도가 0.3보다 낮은 경우는 변별력이 낮아 적절하지 않은 문항이라고 할 수 있다. 난이도는 값이 낮을수록 쉬운 문항인데, TIMSS-R 문항의 경우 대체적으로 난이도가 낮아 모두 음수값을 보였다.

전체적으로 수행형 문항의 변별도가 선다형 문항의 변별도보다 0.2 정도 높다. 내용영역별로 보면, 대수 영역을 제외하고는 수행형 문항의 변별도가 선다형 문항의 변별도보다 높다. 수행형 문항과 선다형 문항의 변별도의 차이가 큰 영역부터 열거하면 측정, 자료 표현 및 해석과 확률, 분수와 수감각, 기하, 비례의 순서이다. 대수는 다른 영역과 달리 선다형 문항의 변별도가 수행형 문항의 변별도보다 약간 높지만 차이가 그리 큰 것은 아니다.

문항의 난이도 평균은 수행형 문항이 선다형 문항에 비하여 0.51 정도 높다. 문항반응이론에서의 난이도는 수식 (1)의 b에 해당하는 지수로서 능력값 θ와 같은 척도 상에서 해석될 수 있다. 수행형 문항의 난이도가 선다형 문항보다 0.51 높으므로, 표준점수로 환산하면 수행형 문항의 평균이 선다형 문항의 평균보다 5.1점 낮게 된다. 내용영역별로 살펴보면, 모든 영역에서 수행형의 문항이 선다형 문항보다 어렵다는 것을 알 수 있다. 특히 비례 영역에서의 수행형 문항과 선다형 문항의 난이도 차이가 가장 크고, 분수와 수감각, 기하, 측정, 자료 표현 및 해석과 확률 영역에서는 평균 정도의 차이를 보이며, 대수 영역에서는 두 가지 문항 유형의 난이도가 비슷한 것으로 나타났다.

변별도란 능력이 높은 학생이 옳은 답을 하고 능력이 낮은 학생이 오답을 하여 문항에 대한 반응을 통해 학생의 능력을 판별해 내는 정도를 말한다. 따라서 대개의 경우 문항의 난이도가 높아질수록 변별도가 높아진다. 난이도와 변별도의 관계는 '높은 난이도-높은 변별도', '낮은 난이도-낮은 변별도'의 경우가 대부분이지만, 반드시 그러한 것은 아니다. <표 2>를 보면 변별도 측면과 난이도 측면에서 수행형과 선다형의 차이가 일치하지 않는다. 비례 영역은 문항 유형에 따른 난이도의 차이가

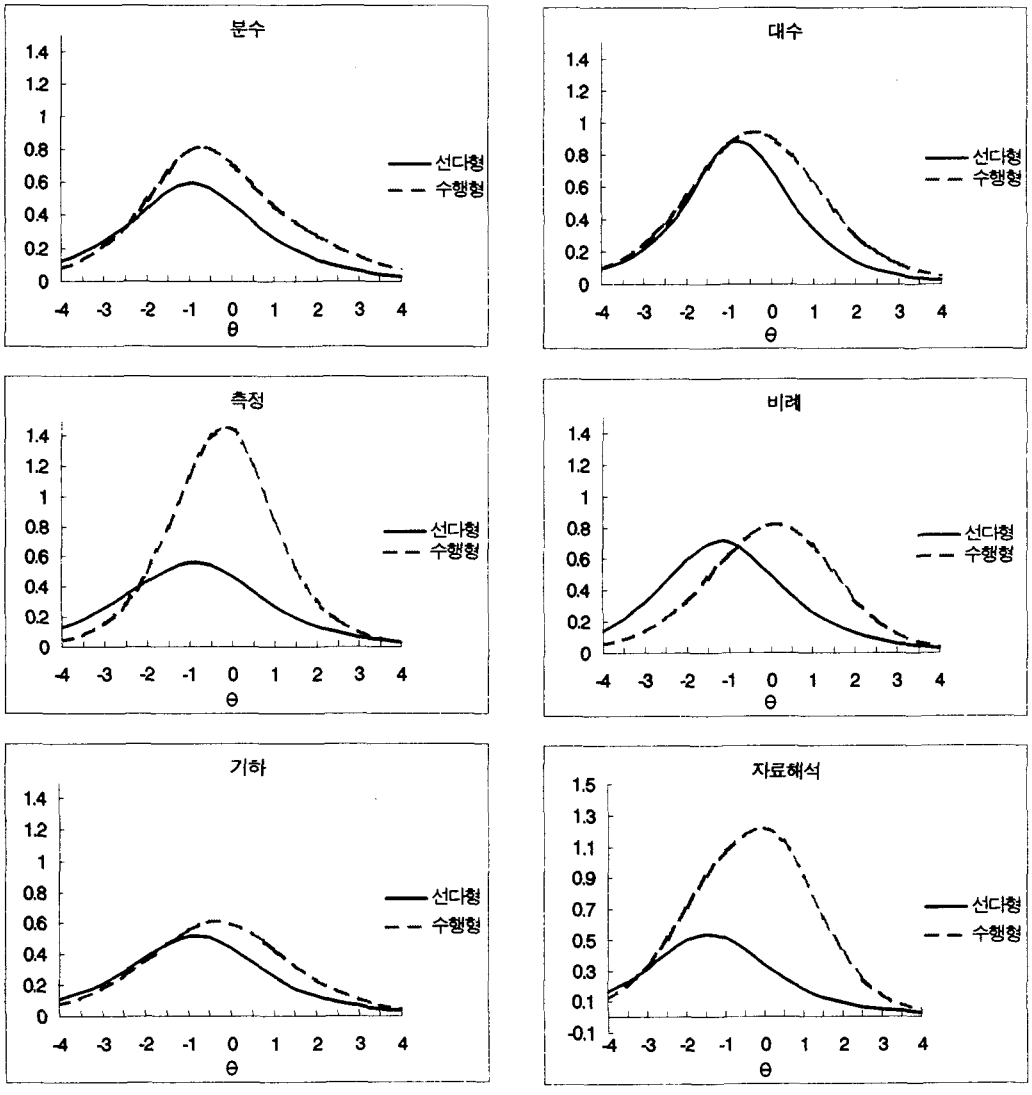
<표 2> 내용영역별 변별도와 난이도

문항특성	문항유형	분수와 수감각	대수	측정	비례	기하	자료 표현 및 해석과 확률	평균
변별도	선다형	0.875	1.061	0.877	0.992	0.785	0.827	0.90
	수행형	1.048	1.022	1.389	1.113	0.920	1.163	1.11
난이도	선다형	-0.991	-0.826	-0.838	-0.953	-0.843	-1.440	-0.98
	수행형	-0.449	-0.618	-0.352	-0.065	-0.317	-0.989	-0.47

크지만 변별도의 차이는 크지 않으며, 역으로 측정 영역은 문항에 유형에 따라 변별도의 차이는 크지만 난이도의 차이는 크지 않다. 예컨대 난이도가 아주 높아 능력이 높은 학생이나 낮은 학생이나 모두 틀리게 되면 변별력이 떨어지므로 '높은 난이도-낮은 변별도'가 나타날 수 있다. 또한 쉬운 문항임에도 불구하고 학생들의 능력을 변별하는 기능을 충실히 수행하는 경우도 있으므로 '낮은 난이도-높은 변별도'도 가능하다.

나. 검사정보함수 곡선

TIMSS-R 수학 검사에 포함된 선다형 문항들과 수행형 문항이 내용영역별로 제공하는 정보함수값을 그래프로 나타내면 다음과 같다.



<그림 2> 내용영역별 선다형 문항과 수행형 문항의 정보함수곡선

<그림 2>에 따르면 모든 내용 영역에서 수행형 문항은 선다형 문항에 비해 피험자에 대해 많은 정보를 제공한다. 특히 측정 영역과 자료해석 영역에서는 수행형 문항이 선다형 문항보다 훨씬 많은 정보를 제공하고, 대수 영역의 경우 문항 유형에 따라 제공하는 정보의 양에 큰 차이가 없음을 알 수 있다.

대수 영역에 포함된 수행형 문항은 다음과 같이 교과서에서 흔히 접할 수 있는 전형적인 문항인 경우가 많다.

어느 특활반의 학생 수는 86명이고, 여학생이 남학생보다 14명 더 많다. 이 특활반의 남학생과 여학생은 각각 몇 명인가? 풀이과정까지 써 보아라.

$x=3$일 때, $\frac{5x+3}{4x-3}$의 값을 구하여라.

위의 문항은 능력 수준과 크게 상관없이 연습의 효과에 의해 정확하게 답할 가능성이 높으므로, 능력 수준에 따라 문제를 해결하는 정도에서 큰 차이가 나타나지 않을 수 있다. 대수 영역에서 선다형과 수행형 문항이 제공하는 정보의 양에 큰 차이가 없다는 것은 두 문항 유형의 변별도가 거의 비슷하다는 의미이다. 대수 영역에 교과서적인 수행형 문항이 다수 포함되어 있다는 것은 이에 대한 부분적인 설명이 될 수 있다.

이와 반대의 논리가 측정 영역이나 자료의 표현 및 해석과 확률 영역에 적용될 수 있다. 이 두 영역의 수행형 문항 중에는 난이도가 그리 높지는 않지만 우리나라 학생들이 익히 연습해 오던 유형이 아닌 경우가 있다. 예를 들어 측정 영역의 수행형 문항 중에는 다섯 개의 정사각형을 십자 모양으로 배열한 복합 도형과 그 전체 넓이를 제시하고, 정사각형의 한 변의 길이와 복합 도형의 둘레를 구하는 문제가 포함되어 있다.[6] 마찬가지로 자료의 표현 및 해석과 확률 영역에도 학생들에게 아주 친숙하지는 않은 수행형 문항이 포함되어 있다. <부록>에 소개한 첫 번째 수행형 문항은 두 잡지의 구독료에 대한 정보를 제공하고 24개월분의 구독료가 더 저렴한 잡지를 선택하는 문항이며, 두 번째의 수행형 문항은 주택의 수와 이를 기호화한 정보를 제시하고, 집 모양의 기호가 주택 몇 채를 나타나는지 구하는 문항이다. 두 문항 모두 평이하고 쉬운 문항이나, 문제의 맥락을 정확하게 이해해야만 해결할 수 있다.

결론적으로 볼 때, 교과서에서 충분히 다루는 전형적인 내용이 수행형 문항으로 출제될 경우 능력이 높은 학생 뿐 아니라 능력이 낮은 학생도 반복적인 연습의 효과로 인해 정확하게 문제를 해결하는 경우가 많다. 따라서 수행형 문항의 변별력이 그리 높지 않으며, 결과적으로 수행형과 선다형의 검사정보함수값도 비슷하게 된다. 이에 반해 익숙하지 않은 내용을 다루는 수행형 문항에 대해서 능력이 높은 학생은 숙고 끝에 문제 해결하는 반면, 능력이 낮은 학생들은 제대로 시도를 하지 못하기 때문에 변별력이 높아질 수 있다. 이로 인해 수행형 문항의 검사정보함수값과 선다형 문항의 검사정보 함수값이 큰 차이를 보이는 것으로 해석된다.

IV. 요약 및 논의

최근 들어 수학교육 평가에서 수행평가에 대한 많은 논의가 이루어져 왔다. 수행형 평가는 학생들의 능력을 제대로 평가한다는 의미에서 '참평가'나 '대안적 평가'라는 별칭을 가지고 있다. 수행형 문항은 기존 평가 방식이 지니고 있는 여러 가지 단점들을 일소할 수 있는 평가도구로 간주되어 왔지만, 수행형 문항과 선다형 문항이 측정학적으로 어떠한 차이를 갖는지 충분하게 탐구되어 왔다고 보기는 어렵다.

본 연구는 수행형 문항이 학생들의 능력을 보다 정확하게 측정하는지를 파악하기 위하여 제3차 수학·과학 성취도 국제비교 반복연구인 TIMSS-R의 자료를 분석하였다. 본 연구의 일차적인 관심사는 수행형의 문항과 선다형의 문항의 능력 추정의 효율성을 비교하는 것이다. 이를 위하여 이분문항반응이론과 다분문항반응이론 모형과 정보함수를 이용하여, 일차적으로 문항 유형에 따라 검사정보함수 곡선이 어떻게 다른지 살펴보았다. 더불어 내용영역별 분석을 통해 선다형과 수행형 문항의 효율성을 비교하였다.

분석 결과, 수행형 문항은 선다형 문항에 비하여 피험자에 대해 많은 정보를 제공한다. 이는 수행형 문항의

6) 비공개 문항이므로 문제의 개요만 설명함.

변별도가 선다형 문항에 비해 높고 피험자 능력을 보다 정확하게 추정할 수 있기 때문에 추정 표준오차가 적음을 의미한다. 내용영역별 비교에서는 측정과 자료표현 및 해석과 확률 영역의 수행형 문항이 선다형 문항에 비하여 학생들의 능력을 추정하는 더 많은 정보를 제공한다. 내용영역별로 검사정보함수값이 다른 이유는 대수영역 수행형 문항의 경우 피험자가 문제해결의 알고리즘을 거의 자동화하여 수행하기 때문이다. 이에 반해 측정과 자료표현 및 해석과 확률 영역의 수행형 문항은 정형화된 내용을 다루지 않기 때문에 문제에 대한 충분한 이해의 토대 위에 사고를 진전시켜 나가야 하며, 따라서 능력에 따라 문제 해결 정도의 차별화가 이루어졌을 것으로 추측된다.

정리하면 수행형 문항은 선다형 문항에 비하여 피험자의 능력을 좀 더 정확하게 추정할 뿐 아니라 상대적으로 적은 수의 문항으로도 피험자의 능력을 추정하는 것이 가능하므로, 보다 더 효율적인 평가 방식이라고 할 수 있다. 선다형 문항과 수행형 문항에 대한 통계적인 분석을 토대로 성취 특성을 분석한 본 연구는 피험자의 능력을 정확하게 변별할 수 있는 문항 유형에 관한 정보를 제공할 수 있다는 점에서 의의를 찾아볼 수 있다.

참 고 문 헌

권오남 · 황숙균 · 권기순 (1999). 중학교 수학과 수행평가의 개발과 적용 효과에 관한 분석, 수학교육학연구 9(1), pp.333-350.

김성숙 · 임찬빈 · 이춘식 · 유준희 · 서동엽 (1999). 제3차 수학 · 과학 성취도 국제비교연구 (TIMSS-R) 국내평가 결과 분석 연구, 한국교육과정평가원 연구보고 RRE 99-7-1.

박정 (2001a). 다분문항반응이론모형, 서울: 교육과학사.

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

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

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

성태제 · 최연희 · 권오남 (1999). 중학교 영어 · 수학 교과에서의 열린 교육을 위한 수행평가 적용 및 효과 분석 연구. 교육부 초등교육정책과 열린교육 연구 과제 보고서.

성태제 (2000). 문항반응이론의 이해와 적용. 서울: 교육과학사

유현주 (1998). 수행평가 과제 제작의 모형 및 준거에 관한 연구. 대한수학교육학회 논문집 8(1), pp. 163-182.

이종성 (1990, 번역). 문항반응이론과 적용. 서울: 대광문화사.

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

A Comparison of Free Response Items and Multiple Choice Items in Terms of Effectiveness of Estimating Mathematical Ability

Park, Chung

Korea Institute of Curriculum and Evaluation

Park, Kyung-Mee

Hongik University

For the past several years, performance assessment has been widely used by mathematics teachers. The superiority of performance assessment items compare to multiple choice items has been discussed by many researchers, however these discussions tend to be lack of empirical data. Thus, this study aims to examine the effectiveness of free response items in comparison with multiple choice items. Using the information function in Item Response Theory(IRT), item information of free response items and multiple choice items from the Third International Mathematics and Science Study-Repeat(TIMSS-R) were obtained and compared. Test informations of the whole mathematics area as well as each content area of mathematics were computed. On average, free response items yielded more information than multiple choice items, especially in measurement and data interpretation. This study also revealed that free response items estimated students' mathematics ability more accurately than multiple choice items with smaller number of items.

* ZDM Classification : C83
* 2000 Mathematics Subject Classification : 97C40
* Key Words : item type, performance assessment, free response item, multiple choice item, information function, IRT, TIMSS-R, estimates of mathematics ability.

<부록> TIMSS-R의 선다형 문항과 수행형 문항의 예

다음 선다형 문항은 '대수' 영역에 해당되며, 우리나라의 정답율은 55.3%이며, 국제 평균 정답율은 23.5%이다 (박정 외, 2001).

다음 표는 x와 y 사이의 비례 관계를 보여준다.

x			
y			

P와 Q의 값은 다음 중 어느 것인가?

① P=40, Q=13 ② P=18, Q=17 ③ P=20, Q=18
④ P=40, Q=18 ⑤ P=18, Q=20

다음 수행형 문항은 '자료의 표현 및 해석과 확률' 영역에 해당되며, 우리 나라의 정답율은 52.1%이며, 국제 평균 정답율은 23.3%이다.

크리스는 다음 월간지 중 하나를 골라 24개월분을 모두 주문하려고 한다. 크리스는 두 종류의 월간지에 대한 다음과 같은 광고를 읽었다. 광고에 나오는 ceds는 크리스의 나라에서 쓰는 화폐의 단위이다.

잡지명 : 청소년 생활	잡지명 : 청소년 소식
24개월분 처음 4개월은 무료 그 후부터는 권당 3ceds	24개월분 처음 6개월은 무료 그 후부터는 권당 3.5ceds

24개월분에 대한 가격은 어느 월간지가 얼마만큼 더 싼가? 풀이과정을 적어라.

TIMSS-R의 문항에 대한 채점기준은 두 자리 수 코드로 분류되어 있다. 분류 코드에서 십의 자리는 정답과 오답이나 부분정답을 나타내는데, 10점대의 코드에 해당하는 반응보다는 20점대의 코드에 해당하는 반응이 수학적으로 더 정답에 가까우며, 더 높은 점수를 받을 가치를 지닌다. 이에 반해 일의 자리는 동일한 점수에 해당하는 반응을 유형

화하여 코드를 부여한 것이다. 각 코드를 기준으로 채점하게 되면 피험자의 다양한 반응이나 오류의 유형을 효과적으로 파악할 수 있다.

코드	응답 유형
	정답 (2점)
20	'청소년 생활'. 두 개의 잡지에 대한 24호분의 가격을('청소년 생활'은 60ceds, '청소년 소식'은 63ceds) 정확하게 계산하고 3ceds가 더 싸다는 결론을 내림
29	이외의 정답 (예, 한 잡지의 금액을 풀이 과정과 함께 올바르게 계산하고 다른 잡지의 금액에 대한 계산을 제시하지 않았으나 '청소년 생활'이 3ceds 싸다는 결론을 올바르게 내린 것)
	부분정답 (1점)
10	('청소년 생활'은 60ceds, '청소년 소식'은 63ceds라고) 정확하게 계산하였으나, 어느 잡지가 더 싼지 결론을 내리지 않았거나 틀린 결론을 내림.
11	(63ceds라는) '청소년 소식'의 계산을 정확하게 하였으나 '청소년 생활'에 대한 가격을 계산하는데 오류를 범함.
12	(60ceds라는) '청소년 생활'의 계산을 정확하게 하였으나 '청소년 소식'에 대한 가격을 계산하는데 오류를 범함.
13	'청소년 생활', 3ceds. 풀이 과정을 제시하지 않음
19	이외의 부분정답 (예, 계산은 정확히 하였으나 차이가 틀린 것)
	오답 (0점)
79	기타
	무응답(0점)
99	공란

다음 수행형 문항은 '자료의 표현 및 해석과 확률' 영역에 해당되며, 우리 나라의 정답율은 89.4%이며, 국제 평균 정답율은 69.1%이다.

다음 표는 어느 도시의 들국화길과 코스모스길에 있는 주택의 수를 나타내고 있다.

도로명	주택의 수
들국화길	30
코스모스길	21

위의 표에 나타난 주택의 수를 그림으로 나타내면 다음과 같다.

들국화길	
코스모스길	

한 개의  이 나타내는 주택의 수는 얼마인가?

코드	응답 유형
	정답 (1점)
10	6채
	오답 (0점)
70	1채
71	5채
79	기타 (지웠거나, X표시를 했거나, 방황한 흔적, 읽기 어렵거나, 제대로 답을 하지 못한 경우)
	무응답(0점)
99	공란