중학생 과학탐구활동 수행평가 시 채점 방식 및 척도의 수에 따른 신뢰도 분석

김형준 · 유준희[1]*
기산중학교 · [1]서울대학교

An Analysis on Reliabilities of Scoring Methods and Rubric Ratings Number for Performance Assessments of Middle School Students' Science Investigation Activities

Hyung Jun Kim · Junehee Yoo[1]*
Gisan Middle School · [1]Seoul National University

Abstract: In this study, reliabilities of holistic scoring method and analytic scoring method were analyzed in performance assessments of middle school students' science investigation activity. Reliabilities of 2, 3, and 4~7-level rubric ratings for analytic scoring methods were compared to figure out optimized numbers of rubric ratings. Two trained raters rated four activity sheets of 60 students by two rating methods and three kinds of rubric ratings. Internal consistency reliabilities of holistic scoring methods were higher than those of analytic scoring methods, while intra-rater reliabilities of analytic scoring were higher than those of holistic scoring methods. Internal consistency reliabilities and intra-rater reliabilities of 3-level rubric rating showed similar patterns of 4~7-level rubric ratings. But students' discriminations, item difficulties and item-response curves showed that the 3-level rubric ratings was reliable. These results suggest that holistic scoring method could be adapted to increase internal consistency reliabilities with improvement in intra-rater reliabilities by rater's conferences. Also, the 3-level rubric rating would be enough for good reliability in case of adapting analytic scoring methods.

Key words: middle school student, science investigation activity, performance assessment, analytic scoring, holistic scoring, rubric ratings, reliability

Ⅰ. 서 론

'맞음'과 '틀림'으로 구분되는 선다형 문항은 지식 위주의 단편적인 사고능력만을 측정한다는 비판에 직면하게 되었다(박정, 2001). 그래서 학습의 과정을 평가하는 수행형의 문항 출제의 비중이 높아지고 있다. 선다형 문항보다 수행형 문항이 학생들의 능력을 더 정확하게 추정함으로서 효율성이 높은 것으로 보고되었다(박정, 홍미영, 2002). 그러나 이런 긍정적인 측면에도 불구하고 수행형 평가는 채점의 신뢰성을 확보하기가 어려운 문제가 발생한다. 수행형의 문항은 선다형 문항보다 완성하는데 시간이 소요되고 각 평가에 포함되는 과제의 수가 적기 때문에 신뢰도를 확보하기가 어렵하고 한다(Linn *et al.*, 1991). 수행평가가 도입된 이래로 많은 학교에서 어려움을 느끼고 있는 부분 중의 하나는 수행평가 결과의 신뢰도라는 지적이 있다(유준희, 박승재, 1999). 그래서 수행형 문항의 신뢰도 확보를 위한 연구 및 채점 방식을 얼마나 분석적으로 하면 신뢰도를 확보할 수 있는지에 대한 연구의 필요성이 제기되었다(김석우, 2007).

측정평가분야의 선행연구에 의하면, 총체적 채점은 수행과제의 전체적인 인상에 의해 단일 점수를 부여하는 방법이고 분석적 채점은 몇 가지 하위요소로 구성된 평가틀에 따라 채점하여 하위요소의 점수를 총합하는 방법으로 서술하고 있다(Klein *et al.*, 1998; 김명숙, 1999; 지은림, 1999). 그래서 총체적 채점은 수행과제의 부분적인 고려보다는 전체적인 과제의 특성에 더 큰 의미를 부여하며(이규민, 2007) 과제 전체

*교신저자: 유준희(yoo@snu.ac.kr)
**2009.12.22(접수) 2010.01.18(1심통과) 2010.02.11(2심통과) 2010.02.16(최종통과)

의 의미가 부분적인 평가요소들의 합보다 더 큰 과제인 경우에 사용된다(Klein *et al.*, 1998). 또한 총체적 채점을 할 때는 수준별로 학생들의 특징을 가장 잘 반영하는 채점 기준을 설정하는 것이 중요하다(Waltman *et al.*, 1998). 분석적 채점은 과제나 질문의 여러 측면에 대하여 얼마나 잘 응답하였는지를 원자화된 평가틀에 근거하여 채점하고 필적이나 선입견 등의 영향이 적기 때문에 총체적 방법보다 객관적인 평가 결과를 제공한다고 주장한다(Klein *et al.*, 1998). 하지만 분석적 채점 시, 평가요소 및 평가요소 간의 관계를 정확하게 규정하지 않으면 어느 평가요소를 적용할지 결정하는데 어려움을 겪을 수 있고 기능적이고 단편적인 면에 치중될 수 있다(이규민, 2007). 채점 시간과 비용의 측면에서는 총체적 채점이 분석적 채점보다 경제적이라는 주장도 있다(Klein *et al.*, 1998).

총체적 채점과 분석적 채점의 결과를 비교한 많은 연구에서 어떤 채점 방식이 신뢰도가 높은지에 대한 일관된 결과가 나타나지 않았다(Black, 1998). 실험 실습이 포함된 과학수행평가를 채점한 결과 분석적 채점이 약간 높다고 보고되었으며(Klein *et al.*, 1998), 문제 해결 형식의 과학수행평가를 채점한 연구에서는 총체적 채점이 약간 높다고 보고되었다(Waltman *et al.*, 1998). 우리나라의 초등학생을 대상으로 한 과학 실험 중심형 수행평가에서는 채점 방식에 상관없이 모두 신뢰로운 측정 결과가 나타났다(이규민, 2007). 과학의 논술형 수행과제를 채점한 결과에서는 총체적 채점과 분석적 채점의 신뢰도 사이에는 거의 차이가 없는 것으로 나타났으며, 학생들을 변별하는데 있어서는 분석적 채점이 다소 유리한 것으로 나타났다(지은림, 2000). 영문편지를 쓰는 과제를 채점한 결과는 분석적 및 총체적 채점 방식 모두 영어작문능력을 신뢰롭게 측정할 수 있으나 채점자 간 신뢰도는 분석적 채점이 총체적 방식보다 낮게 나타났다(김명숙, 1999). 즉 수행평가의 과제 내용이나 특징에 따라 채점 방식 간 신뢰도의 차이가 달라짐을 알 수 있다.

과학탐구활동에서 달성되어야 할 목표를 명확하게 서술한 분석적인 평가틀(APU, GCSE등)이 개발되면서 평가요소에 대한 불명확성은 사라졌지만, 전체가 부분의 합 이상의 의미가 없다는 가정 속에 작은 부분들의 합으로 과학 활동을 축소하여 탐구 과정의 전체적인 의미를 살릴 수 없다는 비판과 평가틀이 구체적이어서 탐구활동을 제한하는 문제가 제기되었다(Woolnough, 1989). 즉, 분석적인 탐구 기능의 평가틀이 실제의 복합적 상황에서 이루어지는 총체적인 탐구활동을 평가하는데 타당도를 담보할 수 없으며, 제한적인 가치를 가진다는 것이다(Black, 1990). 우리나라에서는 2000년도부터 교육적 제반 여건이 미비한 상태에서 교과별로 수행평가 및 서술형 평가를 총점의 30~40% 이상 반영하는 규정이 시행되었다. 그 결과 교사들은 과학 수행평가가 원래 추구하고자 하는 바와는 거리가 멀게 피동적으로 수행평가를 시행하고 있으며(이기영, 안희수, 2005), 우리나라의 학생을 대상으로 중학교 과학탐구활동의 특성을 고려한 수행평가에서 채점 방식에 따른 실증적인 신뢰도 연구는 거의 보고되지 않았다.

또한 과학탐구활동 평가에서 적절한 채점척도의 수를 어떻게 정할 수 있는지에 대한 연구도 부족한 편이다. 채점척도의 수가 작으면 각 점수의 경계에 놓인 학생들이 많아져서 변별력이 떨어지며, 채점척도의 수가 많으면 판단의 기준이 다양해지기 때문에 채점자들의 일관성이 떨어질 수 있어서 적절한 채점척도의 수가 필요하다는 일반적인 주장이 있다(Herman & Winters, 1992). 우리나라 초등학교 수학과 수행평가의 실제적인 평가 상황에서 채점척도의 수가 많을수록 신뢰도 확보에 도움이 된다고 연구결과가 보고되고 있다(김경희, 송미영, 2001). 그러나 현실적으로는 채점척도의 수를 많게 하는 것은 평가의 경제성 면에서 어려움이 있다. 수행평가 실시 과정에서 교사들은 수행평가 계획, 평가, 확인, 자료입력 등의 업무로 과중함을 느끼고 있기 때문에(김석우, 2007), 채점척도의 수를 증가시키는 것은 교사의 부담감을 더 크게 할 가능성이 있다. 또한 선행연구들은 주로 총체적 채점 방식에서 채점척도의 수에 관한 연구이기 때문에 분석적 채점 방식 시 적절한 채점척도의 수에 대한 연구가 요구된다.

이에 본 연구는 중학생의 과학탐구활동 수행평가에서 총체적 채점 방식과 분석적 채점 방식의 신뢰도를 비교분석하고자 하였다. 또한 과학탐구활동에서 분석적 채점 방식을 선택하는 경우, 신뢰도 확보에 적절한 채점척도의 수를 실증적으로 분석하고자 하였다. 본 연구의 결과는 실제 학교 현장에서 수행평가의 채점 방식 및 척도의 수를 결정하는데 필요한 시사점을 제시할 것이다. 본 연구에서 설정한 연구문제는 아래와 같다.

첫째, 중학생의 과학탐구활동 수행평가 시 총체적 채점 방식과 분석적 채점 방식의 신뢰도는 유의미하게 다른가?

둘째, 중학생의 과학탐구활동에 대한 분석적 채점 시 신뢰도 확보에 적절한 채점척도의 수준은 어느 정도로 분석적이어야 하는가?

Ⅱ. 연구 방법 및 절차

1. 연구 대상 및 평가도구

본 연구를 수행하기 위해 서울 소재 영재원에서 중학생 1학년 20명, 2학년 22명, 3학년 18명 등 총 60명을 대상으로 소리의 전달과정을 공기입자모형으로 설명하는 탐구활동을 수행평가로 진행하였다. 해당 영재원에서는 학생들을 학교장의 추천과 면접을 통해서 선발하며, 학생들에게 참 과학탐구활동을 경험하게 하는 것을 주요 목표로 하고 있다. 본 연구에서는 요리책식의 탐구보다 참 과학탐구활동 평가에 관심이 있었으므로 해당 영재원에서 제공하는 과학탐구활동에 참여하는 학생들이 연구대상이 되었다. 탐구활동은 과제1, 과제2, 과제3, 과제4로 이루어져 있고 각각의 과제들은 P-O-E(예상-관찰-설명)의 단계로 구성되었다. 세부적인 과제 내용은 Table 1에 있다. 과제1은 빨대피리를 불어서 큰 소리와 작은 소리, 높은 소리와 낮은 소리가 전달될 때 공기의 움직임을 예상

Table 1
수행 과제의 내용

		수행 과제의 내용			
		과제1	과제2	과제3	과제4
제목		-빨대피리로 소리내기	-용수철을 사용하여 신호 보내기	-소리가 전달되는 동안 공기입자는 어떻게 움직일까?(시뮬레이션)	-소리 만들기
목적		-빨대피리에서 소리가 날 때 공기입자의 움직임을 설명할 수 있다.	-용수철을 사용하여 신호를 보낼 때 용수철의 각 부분은 어떻게 움직이며 신호를 전달하는지, 전체적으로는 어떻게 보이는지 설명할 수 있다.	-시뮬레이션을 사용하여 소리가 날 때 공기입자의 움직임을 설명할 수 있다.	-소리 만들기 활동을 통해 공기 중에서 소리가 어떻게 전달되는지 설명할 수 있다.
단계	예상하기	-빨대피리에서 난 큰 소리, 작은 소리와 높은 소리, 낮은 소리가 전달될 때 공기의 움직임은 어떻게 다를까?	-용수철 한쪽 끝에서 다른 쪽 끝으로 신호를 보내는 동안 용수철의 움직임을 예상해보자.	-다음 그림과 같이 교실 안에 공기입자가 있다면, 소리가 전달될 때 공기 입자는 어떻게 움직일까?	-진동수가 100㎐ 일 때의 소리와 200㎐ 일 때의 소리는 우리 귀에 어떻게 다르게 들릴까? -진폭을 두 배로 하면 우리 귀에 어떻게 다르게 들릴까?
	관찰하기	-Adobe Audition 으로 측정하여 파형을 그려보자.	-용수철 한쪽 끝에서 다른 쪽 끝으로 신호를 보내는 동안 용수철은 어떻게 움직였는가?	-전체적인 공기입자들의 움직임과 빨간색으로 표시된 공기입자 하나의 움직임을 관찰하고 그 결과를 나타내보자. -빨간색으로 표시된 공기입자의 위치를 시간에 따른 그래프로 나타내보자.	-공기입자가 1초에 100번과 200번 진동하는 경우 위치-시간 그래프는 어떻게 될까? -공기입자가 진동하는 폭을 두 배로 하면 위치-시간 그래프는 어떻게 될까?
	설명하기	-빨대피리에서 난 큰 소리와 작은 소리, 높은 소리와 낮은 소리가 전달될 때 파형은 어떻게 다른가?	-용수철의 한쪽 끝에서 다른 쪽 끝으로 신호를 보내는 동안 용수철의 각 부분은 어떻게 움직이며 신호를 전달하는지, 전체적으로는 어떻게 보이는지를 설명해보자.	-공기를 통해 소리가 전달될 때, 공기 입자가 어떻게 움직여 소리가 전달되는지 설명해보자. -큰소리와 작은 소리, 높은 소리와 낮은 소리가 전달될 때 공기입자의 움직임은 어떻게 다를까?	-공기 중에서 소리가 어떻게 전달되는지 공기전체의 움직임과 공기입자 하나의 움직임으로 설명해보자. 이때 소리의 특성에 따라 공기입자의 움직임이 어떻게 달라지는지 설명하라.

하고 녹음한 소리의 파형을 관찰한 후 설명하는 것이다. 과제2는 파동 용수철을 사용하여 진폭이 큰 신호와 작은 신호, 1초 동안 신호를 많거나 적게 만들어 보내는 활동에서 용수철의 각 부분의 움직임을 예상, 관찰, 설명하게 하였다. 과제3은 시뮬레이션을 작동시켜서 큰 소리와 작은 소리, 높은 소리와 낮은 소리에 대한 공기 입자의 움직임을 예상, 관찰, 설명하게 하였다. 마지막으로 과제4는 excel을 활용한 프로그램으로 직접 큰 소리와 작은 소리, 높은 소리와 낮은 소리를 만들어 듣고 최종적으로 공기입자의 움직임으로 소리의 특성을 설명하게 하였다. 학생들은 조별로 과제를 수행하고 개인별로 활동지를 작성하였으며, 학생들이 작성한 활동지를 수합하여 채점하였다.

2. 평가요소 및 평가틀

과학탐구는 증거와 논리적 추론 사이의 상호작용을 통하여 이루어진다는 주장과(AAAS, 2001), 과학탐구의 목표는 자연 현상에 대한 설명을 제공함으로써 자연 세계의 중요한 법칙을 이해하는 것이라는 주장(Mullis *et al.*, 2005)은 모두 과학 탐구에서 증거의 수집과 추론의 상보적 관계를 인식한 것이다. 세부적인 과학 탐구의 요소는 탐구 문제 및 가설 설정, 탐구의 설계, 자료의 수집, 자료의 분석과 해석, 결론 도출 및 설명(Mullis *et al.*, 2005), 또는 탐구 설계하기, 자료수집하기, 정확하고 안전하게 활동하기, 자료 수집 방법 및 자료의 타당도와 신뢰도 평가하기(QCA, 2007)로 약간씩 다르게 강조되기도 한다. 그러나 신뢰로운 자료의 수집, 수집된 자료와 학생의 이해를 바탕으로 설명 모형을 추론하기 등은 공통적으로 포함되는 요소로 해석할 수 있다.

과학탐구에 대한 위의 이론적 배경을 근거로 본 연구에서는 중학생 과학탐구활동의 핵심을 학생이 가지고 있는 이론적 모형과 관찰한 증거 사이의 상호작용에 두었으며, 이를 P-O-E(예상, 관찰, 설명) 단계를 통하여 구현하고자 하였다. 각 단계에서 평가하고자 하는 탐구능력은 TIMSS 2007의 평가틀(Mullis *et al.*, 2005) 중에서 본 연구에서 강조하는 요소를 발췌하여 사용하였으며, 본 연구에서 설정한 평가요소를 Table 2에 나타냈다. 예상하기와 설명하기 단계에서 평가요소는 '적용 능력'이며, 평가 하위요소는 모형 사용 및 관련짓기 능력으로 설정하였다. 예상하기 단계에서의 적용 능력은 '주어진 조건이나 정보들로부터 과학적 현상을 개념, 모형 등을 적용하여 설명할 수 있다'로, 설명하기 단계에서의 적용 능력은 '관찰이나 측정된 자료를 바탕으로 과학적 현상을 모형, 개념 등을 적용하여 설명할 수 있다'로 정의할 수 있다. 관찰하기 단계에서는 '자료 수집과 변환능력'이 평가요소이며, 평가하위요소로는 '관찰 및 측정 능력'과 '자료변환 능력'을 설정하였다. '자료 수집과 변환 능력'은 '탐구 활동 수행 시 관찰이나 측정을 통해 수집된 자료를 그래프와 도형을 사용하여 해석될 수 있다'로 정의하였다. 예를 들어 과제1 예상하기 단계에서 과제내용은 "큰 소리와 작은 소리가 전달될 때 공기의

Table 2
과제별 평가요소 및 평가 하위요소

평가요소 (단계)	평가하위 요소	평가내용	과제1	과제2	과제3	과제4
적용능력 (예상하기)	모형사용 능력	주어진 조건이나 정보들로부터 모형을 사용하여 과학적 현상을 설명할 수 있다.	○	○	○	
	관련짓기 능력	주어진 조건이나 정보들로부터 과학적 현상과 현상사이를 관련지어 설명할 수 있다.		○		○
자료 수집 및 변환 능력 (관찰하기)	관찰 및 측정 능력	적절한 언어로 관찰 및 측정된 자료를 기술할 수 있다.	○	○	○	○
	자료변환 능력	관찰이나 측정을 통해 수집된 자료를 그래프나 도형으로 변환할 수 있다.	○	○	○	○
적용능력 (설명하기)	모형사용 능력	관찰이나 측정된 자료를 바탕으로 모형을 사용하여 과학적 현상을 설명할 수 있다.		○	○	○
	관련짓기 능력	관찰이나 측정된 자료를 바탕으로 과학적 현상과 현상사이를 관련지어 설명할 수 있다.	○	○	○	○

움직임은 어떻게 다를까?"로, 적용능력 중 모형사용능력이 평가 하위요소로 포함되었다.

3. 채점 방식

연구문제 1에 답하기 위하여 채점 방식을 총체적 채점 방식 1가지와 분석적 채점 방식 2가지로 나누어 비교 분석하였다. 총체적 채점 방식에서는 과학 현상을 관찰과 측정된 증거들을 사용하여 설명하는 능력을 총체적으로 채점하였다. 총체적 채점 방식의 평가 수준에 관련한 선행연구 중 Etkina *et al.*(2006)이 과학 수행평가의 수준을 공란, 부적절함, 향상이 필요함, 적절함으로 4수준으로 나누었고 Halonen *et al.*(2003)은 과학탐구 평가에 사용하는 5수준의 채점척도를, Hafner *et al.*(2003)는 과학 탐구과제의 구두발표 평가 시 5수준의 채점척도를 사용하였다. 본 연구에서는 선행 연구 결과 및 학교에서의 활용을 고려하여 총체적 채점 방식의 채점척도를 전문가적 설명, 진보한 학생설명, 보통학생설명, 초보적인 설명, 설명하지 못함 등의 5수준으로 설정하였다.

분석적 채점의 경우는 평가 영역을 대범주로 한 경우와 소범주로 한 경우를 구분하여 신뢰도 분석을 하였다. 분석적 채점 방식1은 평가영역을 대범주로, 분석적 채점 방식2는 평가영역을 소범주로 환원하여 채점한 것이다. 분석적 채점 방식1은 적용능력(예상), 자료수집 및 변환능력, 적용능력(설명)등 3가지의 평가요소를 대상으로 채점하였고, 분석적 채점 방식2는 평가 하위요소인 모형사용능력(예상), 관련짓기 능력(예상), 관찰 및 측정 능력, 자료변환능력 ,모형사용능력(설명), 관련짓기 능력(설명) 중 과제별로 각각에 해당되는 4~6가지의 평가 하위요소를 설정하여 채점하였다. 국가 수준의 성취기준과 평가기준이 분석적인 평가요소에 대하여 3수준으로 기술되어 있기 때문에(한국교육과정평가원, 2001), 본 연구의 분석적 채점 방식1과 2는 평가요소별로 채점척도를 3수준으로 나누어서 채점하였다. 학생들이 충실히 수행하여 도달하기를 기대하는 수준을 '중'으로, 중 수준 이상의 성취를 '상'으로, 최소한의 필수 수준만을 성취한 정도를 '하'수준으로 설정하였다. 채점 방식에 따라 설정된 활동 4의 채점 기준을 Table 3에 제시하였다.

4. 채점척도의 수

연구문제 2에 답하기 위하여 분석적 채점 방식에서 채점척도의 수를 2수준, 3수준, 4~7수준으로 채점하여 비교 분석하였다. 채점척도 2수준은 평가요소 별로 상, 하의 수준으로 나누어 채점하며 단답형 과제의 채점과 비슷하다. 단답형 과제보다 부분점수를 줄 수 있는 수행형 과제가 더 학생들의 능력을 효율적으로 추정한다고 보고되었지만(박정, 홍미영, 2002) 보다

Table 3
채점 방식에 따른 활동 4의 채점기준 예시

총체적 채점			분석적 채점1			분석적 채점2			
평가요소	채점척도	배점	평가요소	채점척도	배점	평가요소	평가하위요소	채점척도	배점
과학 현상을 모형과 관련지어 설명할 수 있는 능력	전문가적	20	적용능력 (예상)	상	4	적용능력 (예상)	관련짓기 능력	상	4
				중	2			중	2
				하	0			하	0
	진보	15	자료수집 및 변환능력	상	6	자료수집 및 변환능력	관찰 및 측정 능력	상	2
				중	3			중	1
				하	0			하	0
	보통	10					자료변환능력	상	4
								중	2
								하	0
	초보	5	적용능력 (설명)	상	10	적용능력 (설명)	모형사용능력	상	6
				중	5			중	3
								하	0
	설명못함	0		하	0		관련짓기 능력	3	4
								2	2
								1	0

다양한 채점척도와 비교하는 실증적인 연구가 필요하기 때문에 채점척도 2수준을 비교 대상으로 포함하였다. 채점척도 3수준은 상, 중, 하의 수준으로 나누어 채점하는 것으로 미국 NAEP의 학업성취도 평가 등에서 수행평가 수준의 수를 3수준으로 기본, 숙련, 진보로 나누고 있음(Plake & Hambleton, 1999)을 참고하였다. 채점척도 4~7수준은 평가요소별로 내용에 따라 4 내지 7수준의 채점척도로 나누어서 가장 분석적으로 채점하였는데 Herman *et al.*(1992)이 제시한 5~7 단계가 채점자들이 잘 구분한다는 연구결과를 참고하였다. Table 4는 과제4 중 "공기 중에서 소리가 어떻게 전달되는지 공기입자전체의 움직임과 공기입자 하나의 움직임으로 설명해보자. 이때 소리의 특성에 따라 공기입자의 움직임이 어떻게 달라지는지 설명하라"라는 과제에서 측정하고자 하는 모형사용능력에 대한 채점기준표이다. 학생의 응답이 "소리가 전달될 때 공기입자가 진동을 전달하면서 조밀하고 넓은 부분이 생겨나고 전진하며 진동을 전달한다. 이때 공기입자 하나는 앞뒤로 진동한다. 큰 소리를 전달할 때는 공기입자가 크게 진동하며, 작은 소리를 전달할 때는 작게 진동한다. 높은 소리를 전달할 때는 공기입자가 빨리 진동하며 낮은 소리를 전달할 때는 느리게 진동한다."이면 6점에 해당한다. Table 4에서 나타난 바와 같이 평가 수준의 경계는 가장 핵심적인 개념인 종진동과 진동에 대한 일반적인 개념의 포함여부를 기준으로 설정하였다.

5. 채점 및 분석 도구

채점자는 모두 4명으로 중등학교에서 교직경력 3년 이상의 수행평가 경험이 있는 현직 과학 교사다. 채점자1과 2는 분석적 채점으로 먼저 채점한 후에 총체적인 채점을 했다. 채점자1과 2는 한 방법이 다른 방법에 영향을 미치지 않도록 하기 위해 두 채점 사이의 시간 간격을 일주일 정도로 유지하였고 학생 순서를 임의로 바꾸었다(이규민, 2007). 분석적 채점이 총체적 채점의 신뢰도에 영향을 주는지 확인하기 위하여 정적 집단 비교를 하였다. 즉, 채점자3과 4는 분석적 채점은 하지 않고 총체적 채점만 하도록 하여 채점자 1, 2와 비교하였다.

채점자1과 2는 분석적 채점 기준과 총체적 채점 기준에 관해서 서로 의견을 조율하면서 각 등급에 해당하는 학생응답을 공유하고 채점 기준표에 숙달되도록 4시간 정도 훈련하였다. 채점자3과 4는 각 등급에 해당하는 채점 기준과 학생응답의 예시안을 채점자1과 협의하여 검토한 후 총체적 채점을 실시하였다. 평가의 신뢰도 분석을 위해서는 채점자 내 신뢰도와 채점자 간 신뢰도를 검토하였다(성태제, 2002). 개발한 평가도구와 채점기준으로 학생들의 능력 수준을 정확히 측정하였는지를 문항반응이론에 기초하여 학생들의 점수를 확인하였다. Wilson *et al.*(1995)등에 의하면 학생들의 과제에서 고전검사이론의 곤란도에서는 의미 있는 차이를 볼 수 없는 경우도, 문항반응이론의 곤

Table 4
공기입자의 움직임에 대한 모형사용능력 채점기준표 예시

<table>
<tr><th colspan="2">채점척도 4~7수준</th><th colspan="2">채점척도 3수준</th><th colspan="2">채점척도 2수준</th></tr>
<tr><th>배점</th><th>채점기준</th><th>배점</th><th>채점기준</th><th>배점</th><th>채점기준</th></tr>
<tr><td>6</td><td>공기입자 모형을 종진동으로 설명하고 전체적인 파동의 전달을 설명함.</td><td rowspan="4">6</td><td rowspan="4">공기입자 모형을 종진동으로 설명하고 전체적인 파동전달을 설명함.</td><td rowspan="4">6</td><td rowspan="4">공기입자 모형을 종진동으로 설명함.</td></tr>
<tr><td>5</td><td>공기입자 모형을 종진동으로 설명하고 전체적인 파동의 전달 모습만 설명함.</td></tr>
<tr><td>4</td><td>공기입자 모형을 종진동으로 설명하고 전체적인 파동의 전달 방향만 설명함.</td></tr>
<tr><td>3</td><td>공기입자 모형을 종진동으로 설명함.</td></tr>
<tr><td>2</td><td>공기입자 모형을 진동으로 설명하고 전체적인 파동의 전달 모습만 설명함.</td><td rowspan="2">3</td><td rowspan="2">공기입자 모형을 진동으로 설명함.</td><td rowspan="3">0</td><td rowspan="3">공기입자 모형을 종진동으로 설명하지 못함.</td></tr>
<tr><td>1</td><td>공기입자 모형을 진동으로 설명함.</td></tr>
<tr><td>0</td><td>공기입자 모형을 움직임으로 설명하거나 설명하지 못함.</td><td>0</td><td>공기입자 모형을 움직임으로 설명하거나 설명하지 못함.</td></tr>
</table>

란도에서는 효과적으로 구분할 수 있다고 주장하였다.

기초통계 분석프로그램으로 spss 12.0을 사용하였고, 문항반응이론에 기반한 분석프로그램으로 미국 버클리 대학 평가 및 측정 센터(Berkely Evaluation and Assessment Research Center, BEAR)에서 개발한 Grade Map을 사용하였다(Wilson & Sloane, 2000). 이 프로그램은 문항반응이론에 근거하여 다양한 그래프와 보고서를 제공하는데 본 연구에서는 학생들의 능력추정치와 평가요소의 곤란도를 볼 수 있는 줄기 · 잎 그래프와 각 평가요소의 수준에 대한 학생들의 반응을 분석한 문항특성곡선을 사용하였다.

Ⅲ. 연구 결과 및 논의

1. 평가요소 및 평가 기준의 타당도 분석

본 연구에서는 과제별 평가요소 및 평가 하위요소를 개발한 뒤 내용과 평가기준의 타당도를 과학교육 전문가 1명과 중학교 근무경력이 3년 이상인 현직 교사 4명으로 구성된 자문단과 함께 검토하고 수정 보완하였다. 총 5차례의 협의회를 거쳐 과학 탐구활동으로 적절한 내용을 구성하고 학습목표를 설정했으며 과제별 평가요소 및 평가 하위요소가 그 목표에 부합하고 중학교 학생의 수준에 적절하다고 합의될 때까지 수정 보완하였다.

구인타당도를 검증하기 위해 평가 하위요소들 간의 상관관계를 분석하였고 결과는 Table 5와 같다. 분석 결과, 평가 하위요소간의 상관계수는 0.38~0.60이고 평가 하위요소와 총점간의 상관계수는 0.69~0.82로 통계적으로 유의미한 정적 상관이 있다고 나타났으며, 본 연구에서 설정한 평가 하위요소들이 과학 탐구 능력의 구인으로서 타당성을 확보했다고 볼 수 있다.

2. 총체적 채점 방식과 분석적 채점 방식의 신뢰도 분석

채점 방식에 따른 신뢰도를 과제 간 내적 일치도, 채점자 간 신뢰도를 통하여 분석하였다. 과제 간 내적 일치도를 구하기 위한 Cronbach′ α의 분석 결과는 Table 6과 같다. 분석 결과 과제 4개 사이의 내적 일치도는 채점 방식에 따라 0.70~0.81 사이에 분포하며, 전체적으로 비교적 높게 나타났다. 채점자 1과 채점자 2 모두에게서 총체적 채점 방식의 경우, 다른 채점 방식에 비해서 과제 간 내적일치도가 높게 나타났다.

채점 방식에 따른 채점자 1과 채점자 2와의 채점자 간 신뢰도를 구하기 위해 Pearson의 상관계수를 추정한 결과는 Table 7과 같다. 채점자 간 신뢰도는 과제1에서 0.67~0.85로 약간 낮게 나타났고 나머지 과제는 0.83~0.92로 대체로 높게 나타났으며 총점에서는 0.92~0.95로 높게 나타났다. 평가요소 4~6가지인 분석적 채점2가 채점자 간 신뢰도는 0.95로 가장 높게 나타났다.

채점 방식별 채점자 간 신뢰도 차이를 검정하기 위해 $Z_{관찰}$[1] 값을 산출하여 Table 8에 제시하였다. 과제1, 3, 4에서 분석적 채점 방식2와 총체적 채점 방식이 통

Table 5
평가 하위요소 간의 상관관계

채점자별	모형사용 능력(예상)	관련짓기 능력(예상)	관찰 및 측정 능력	자료변환 능력	모형사용 능력(설명)	관련짓기 능력(설명)	총점
모형사용능력(예상)	1.00						
관련짓기능력(예상)	.46*	1.00					
관찰 및 측정 능력	.48*	.52*	1.00				
자료변환능력	.47*	.38*	.39*	1.00			
모형사용능력(설명)	.58*	.42*	.47*	.39*	1.00		
관련짓기능력(설명)	.60*	.53*	.50*	.53*	.60*	1.00	
총점	.80*	.70*	.73*	.69*	.79*	.82*	1.00

*p<.01

1) 채점자 간 신뢰도의 차이는 상관계수 r_1, r_2의 표준화된 값 Z_{r_1}, Z_{r_2}를 이용하여 $Z_{관찰} = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$ 을 구한다. 유의도 .05 수준에서 $Z_{관찰}$ 값이 1.96이상 또는 −1.96이하인 경우 두 상관계수 간에 차이가 있다고 할 수 있다(이관용, 김기중, 1993)

Table 6
채점 방식에 따른 평균 점수 및 과제 간 내적 일치도

채점 방식		과제1		과제2		과제3		과제4		총점		과제 간 내적 일치도
		평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차	
총체적 채점	채점자 1	7.17	4.36	11.73	8.26	9.15	5.89	8.83	6.66	36.88	19.88	0.77
	채점자 2	6.47	4.35	13.48	9.04	9.05	4.66	9.58	6.46	38.58	20.50	0.81
분석적 채점1 (평가요소 3가지)	채점자 1	7.12	2.83	12.53	9.15	9.67	5.41	9.17	4.89	38.48	17.77	0.73
	채점자 2	8.27	4.21	13.68	9.10	10.60	4.91	8.78	4.48	41.33	18.49	0.77
분석적 채점2 (평가요소 4~6가지)	채점자 1	6.85	2.86	12.75	6.93	9.77	4.41	11.12	3.91	40.48	13.84	0.70
	채점자 2	6.80	2.88	14.08	6.53	8.87	4.26	10.82	3.92	40.57	13.59	0.72

Table 7
채점 방식에 따른 채점자 1과 채점자2의 채점자 간 신뢰도

채점 방식	과제1	과제2	과제3	과제4	총점
총체적 채점	.71*	.90*	.73*	.83*	.92*
분석적 채점1 (평가요소 3가지)	.67*	.92*	.87*	.77*	.94*
분석적 채점2 (평가요소 4~6가지)	.85*	.89*	.89*	.92*	.95*

*p<.01

Table 8
채점 방식별 채점자 간 신뢰도 차이 검정($Z_{관찰}$)

채점 방식	활동1	활동2	활동3	활동4	총점
분석적 채점2-총체적 채점	1.97*	−0.27	2.63*	2.14*	1.30
분석적 채점1-총체적 채점	−0.41	0.62	2.16*	−0.90	0.82
분석적 채점2-분석적 채점1	2.37*	−0.89	0.48	3.04*	0.50

*p<.05

계적으로 차이가 유의미하게 나타났으며, 과제1, 4에서 분석적 채점 방식2와 분석적 채점 방식1이 통계적으로 차이가 유의미하게 나타났다. 채점 방식에 따른 채점자 간 신뢰도의 차이가 총점에서는 통계적으로 유의미하지 않았지만, 과제별로 본다면 분석적 채점 방식2가 다른 채점 방식에 비해 채점자 간 신뢰도가 높다고 해석할 수 있다.

Table 6에서 과제 간 내적일치도가 총체적 채점 방식에서 분석적 채점 방식보다 높게 나타났다. 이는 총체적 채점 방식의 특징으로 해석할 수 있으나 분석적 채점을 먼저 해서 나타난 후광효과일 수도 있기 때문에, 비교집단인 채점자3과 4를 활용하여 후광효과의 여부를 검증하였다. 즉, 채점자3과 4는 분석적 채점을 하지 않고 총체적 채점을 수행하여 과제 간 내적일치도 및 채점자 간 신뢰도를 분석하였다. 과제 간 내적일치도의 결과는 Table 9와 같다. 비교집단의 과제 간 내적일치도는 0.74와 0.81로 나타났으며, 실험집단은 0.77 및 0.81로 나타났다.

총체적 채점 방식에서 채점자 간 신뢰도의 결과는 Table 10과 같다. 과제별로 실험집단의 채점자 간 신

Table 9
총체적 채점 방식의 채점자별 평균 점수 및 과제 간 내적일치도

채점자		과제1		과제2		과제3		과제4		총점		과제 간 내적 일치도
		평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차	
실험 집단	채점자1	7.17	4.36	11.73	8.26	9.15	5.89	8.83	6.66	36.88	19.88	0.77
	채점자2	6.47	4.35	13.48	9.04	9.05	4.66	9.58	6.46	38.58	20.50	0.81
비교 집단	채점자3	7.75	4.60	13.40	9.38	12.97	6.28	13.08	5.05	47.20	19.83	0.74
	채점자4	6.95	4.06	14.42	8.38	11.37	6.45	12.58	4.56	45.32	19.54	0.81

*p<.01

Table 10
총체적 채점 방식에서 채점자 간 신뢰도

채점자별	과제1	과제2	과제3	과제4	총점
채점자1, 2	.71*	.90*	.73*	.83*	.92*
채점자3, 4	.73*	.74*	.76*	.72*	.89*

*p<.01

뢰도는 0.71~0.90으로 나타났으며, 비교집단은 0.72~0.76으로 나타났다. 과제2에서 실험집단과 비교집단의 채점자 간 신뢰도의 차이 검증 값($Z_{관찰}$)이 −2.79로 통계적으로 유의미하지만 과제1, 3, 4에서는 $Z_{관찰}$ 값이 0.28~1.49로 통계적으로 유의미하지 않았다. 총점에서는 실험집단의 채점자 간 신뢰도가 0.92로 나타났고 비교집단은 0.89로 나타났다. 실험집단과 비교집단의 채점자 간 신뢰도의 차이를 알아보기 위해 z관찰 값을 산출한 결과 0.89로 통계적으로 유의미한 차이가 나타나지 않았다.

따라서 실험집단의 총체적 채점에서 과제 간 내적일치도가 분석적 채점보다 높게 나타난 것을 후광효과로 해석하기는 어렵다. 결론적으로 중학생의 과학탐구활동에서는 총체적 채점 방식이 과제 간 내적일치도가 높게 나타났으며, 분석적 채점 방식이 채점자 간 내적일치도가 높게 나타났다. 과제 당 채점시간은 분석적 채점 방식이 대략 6시간이 걸리고 총체적 채점 방식은 2시간 정도 걸렸다. 평가의 목적이나 현실적 여건을 고려하여 채점 방식을 선택할 수 있으며, 어떤 채점 방식을 선택하더라도 각각의 단점을 보완하기 위한 노력이 필요하다.

2. 분석적 채점 시 채점척도의 수에 따른 신뢰도 분석

본 연구에서는 채점척도의 수에 따른 신뢰도를 판단하기 위해서 과제의 내적일치도, 과제 간 내적일치도, 채점자 간 신뢰도, 문항특성곡선 등을 분석하였다. 내적일치도를 구하기 위한 Cronbach′ α의 값은 Table 11과 같다. 채점척도의 수에 따른 채점자간 신뢰도를 구하기 위해 Pearson의 상관계수를 추정한 결과는 Table 12와 같다. 채점척도 2수준에서 채점자 2의 과제1의 내적일치도가 0.28로 낮지만 전체적으로는 과제의 내적일치도와 총점의 내적일치도가 0.33에서 0.80으로 적절하였다. 채점자 간 신뢰도는 모든 경우에 0.78이상으로 높게 나타났으며 통계적으로 유의미하였다. 따라서 각각의 채점척도의 수에 따라 개발된 채점기준표가 적절하다고 추정할 수 있다.

분석 결과, '맞음'과 '틀림'으로 구분되는 채점척도 2수준의 경우 과제의 내적일치도가 0.28~0.65로 낮게 나타났다. 채점척도 3수준의 경우는 0.35~0.76, 채점척도 4~7수준의 경우는 0.33~0.75 사이에서 과제의 내적일치도를 나타났다. 총점의 내적일치도는 각 과제별 점수 사이의 내적일치도를 분석한 결과인데 채점척도 3수준의 경우는 0.71, 0.72로, 채점척도 4~7수준의 0.70 및 채점척도 2수준의 0.65, 0.69보다 약간 높게 나타났다. 전체적으로 채점척도 3수준의 내적일치도가 다른 채점척도의 수준보다 높다고 해석된다.

채점척도의 수에 따라 채점자 간 신뢰도가 통계적으로 의미 있는 차이를 보이는지 검증하기 위하여 $Z_{관찰}$

Table 11
분석적 채점 시 채점척도의 수에 따른 평균 점수 및 과제의 내적일치도

채점척도		과제1			과제2			과제3			과제4			총점		
		평균	표준편차	내적일치도	평균	표준편차	내적일치도	평균	표준편차	내적일치도	평균	표준편차	내적일치도	평균	표준편차	내적일치도
2수준	채점자1	8.10	3.67	0.35	11.03	7.76	0.59	14.37	5.84	0.46	9.88	5.41	0.39	43.38	16.53	0.69
	채점자2	7.77	3.72	0.28	11.42	8.44	0.65	15.10	5.51	0.45	9.77	5.34	0.38	44.05	16.75	0.65
3수준	채점자1	6.85	2.87	0.38	12.58	7.28	0.76	10.20	4.12	0.63	11.12	3.91	0.53	40.75	14.06	0.71
	채점자2	6.80	2.88	0.35	14.08	6.53	0.74	8.87	4.26	0.73	10.82	3.92	0.42	40.57	13.59	0.72
4~7수준	채점자1	6.67	2.86	0.35	11.97	7.03	0.75	11.73	4.01	0.63	10.87	4.05	0.50	41.23	13.76	0.70
	채점자2	6.62	2.84	0.33	12.77	7.10	0.75	11.70	3.85	0.65	10.45	4.02	0.46	41.53	13.72	0.70

Table 12
분석적 채점 시 채점척도의 수에 따른 채점자 간 신뢰도

채점척도	과제1	과제2	과제3	과제4	총점
2수준	.78*	.84*	.87*	.86*	.92*
3수준	.85*	.93*	.93*	.92*	.95*
4~7수준	.84*	.93*	.92*	.91*	.96*

*p<.01

을 구한 결과, 과제2에서 채점척도 2수준의 채점자 간 신뢰도와 나머지 채점척도의 채점자 간 신뢰도의 차이 검증 값($Z_{관찰}$)이 −2.33으로 통계적으로 유의미한 차이가 나타났다. 즉 채점척도 2수준의 채점자간 신뢰도가 다른 채점척도 수준에 비해 낮다고 해석된다. 나머지 과제와 총점에서는 유의미한 차이가 나타나지 않았다.

채점척도의 적절성을 판단하기 위하여 줄기 · 잎 그래프와 문항특성곡선을 분석하였다. 줄기 · 잎 그래프를 통하여 능력 추정치[2]에 대한 학생 분포와 평가요소별 수준의 곤란도[3]를 확인하였다(Roberts *et al.*, 1997). 줄기 · 잎 그래프의 왼쪽 부분에는 능력 추정치에 대한 학생 분포를 나타내는데 x표시는 학생의 상대적인 도수 분포이고 y축은 능력 추정치를 나타낸다. 줄기 · 잎 그래프의 오른쪽에는 평가요소별 수준의 곤란도를 나타내는데 숫자는 평가요소의 수준들이고 y축은 곤란도의 추정치를 나타낸다. 예를 들어 1.1은 평가요소 1인 모형사용능력(예상)의 1점을 의미하고, 1.2는 평가요소 1인 모형사용능력(예상)의 2점을 의미한다.

과제별로 채점척도의 수에 따른 줄기 · 잎 그래프를 Fig. 1과 Fig. 2에 제시하였다. 모든 과제에서 채점척도 3수준의 경우 능력 추정치에 따른 학생 분포가 가장 정규분포[4]와 유사하였다. 과제1에서 채점척도 2수준의 평가요소별 수준의 곤란도는 −1.25~2.72이고 중간 이상의 곤란도를 대표하는 평가요소가 부족하다고 판단된다. 채점척도 3수준은 −2.58~2.54, 채점척도 4~7수준은 −2.47~2.94이고 그 사이의 곤란도를 대표하는 수준들이 존재함을 볼 수 있다. 과제2에서 채점척도 2수준의 평가요소별 수준의 곤란도는 −1.01~1.33이고 중간 이하의 곤란도를 대표하는 평가요소가 부족하다고 판단된다. 채점척도 3수준의 평가요소별 수준의 곤란도는 −2.53~3.94이고 그 사이의 곤란도를 대표하는 수준들이 존재함을 볼 수 있다. 채점척도 4~7수준의 평가요소별 수준의 곤란도는 −1.23~2.91이고 모형사용 능력(설명)의 2점과 3점, 관련짓

2) 문항반응이론에 의한 능력추정치는 표준점수를 사용하여 평균이 0이고 표준편차가 1인 척도를 사용한다. 능력추정치 ( )에 따라 문항의 답을 맞힌 확률은 $P(\theta)=\frac{1}{1+e^{-1.7(()-b)}}$ 이다(성태제, 2005). 여기서 b는 문항의 곤란도다.

3) 문항의 답을 맞힐 확률이 0.5에 해당하는 능력추정치를 말한다.

4) 정규성 검정을 Shapiro-Wilk으로 한 결과 채점척도의 다른 수준들은 0.81~0.91(유의도 0.01 수준)이지만 채점척도 3수준은 0.86~0.95(유의도 0.01 수준)이었다.

기 능력(설명)의 2점과 3점에 해당하는 수준들이 모두 −0.84~−0.67 주변에 유사한 곤란도로 밀집하여 있다. 과제3에서 채점척도 2수준의 평가요소별 수준의 곤란도는 −0.77~0.80이고, 채점척도 3수준은 −2.59~3.35이고, 채점척도 4 7수준은 −2.45~3.15이다. 채점척도 4~7수준은 모형사용 능력(예상)의 2점과 3점, 관련짓기 능력(설명)의 1점과 2점에 해당하는 수준들이 모두 −0.67 주변에 유사한 곤란도로 밀집하여 있다. 과제 4에서 채점척도 2수준의 평가요소별 수준의 곤란도는 −1.77~0.98이고, 채점척도 3수준은 −2.24~2.85이고, 채점척도 4~7수준은 −1.57~2.21이다. 채점척도 4~7수준의 경우는 관련짓기 능력(설명)의 1점과 2점의 수준과 모형사용 능력(설명)의 2점도 모두 −0.65~−0.83 주변에 유사한 곤란도로 밀집하여 있다.

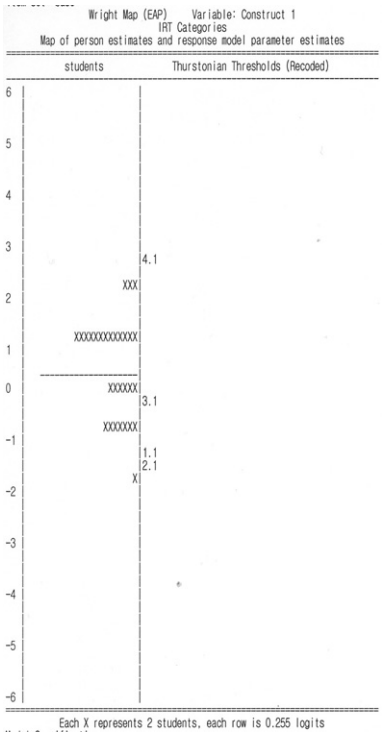
Fig. 1과 Fig. 2에 제시된 줄기 · 잎 그래프를 보면 과제1은 세 가지 채점척도에서 모두 평가요소들의 곤란도가 고루 분포함을 볼 수 있다. 하지만 과제2, 3, 4의 채점척도 4~7수준은 다른 채점척도에 비해 평가요소별 수준의 곤란도가 −1.0~1.0 사이에 모여 있고 특히 평가요소 중 관련짓기 능력의 곤란도가 −0.84~−0.65 사이에 많이 모여 있음을 볼 수 있다. 자세한 특성을 확인하고자 과제2, 3, 4 중에서 관련짓기 능력에 대한 채점척도의 수에 따른 문항특성곡선을 Fig. 3에 제시하였다. 채점척도 3수준의 과제2, 3, 4에서 '관련짓기 능력' 평가요소의 문항특성곡선을 보면 0점은 낮은 능력 추정치를 가진 학생이, 1점은 중간 능력 추정치를 가진 학생이, 2점의 수준은 높은 능력 추정치를 가진 학생이 받을 확률이 높다. 하지만 채점척도 4~7수준의 경우는 0점은 능력 추정치가 낮은 학생이 받을 확률이 높고 1점과 2점의 수준은 능력 추정치가 중간 정도인 학생이 받을 확률이 낮아 수준 사이의 구분이 명확하지 않다.

종합하면 과제 간 내적일치도와 채점자 간 신뢰도는 채점척도 2수준의 경우보다는 채점척도 3수준의 경우나 4~7수준의 경우가 높았다. 줄기 · 잎 그래프의 분석 결과 능력추정치에 따른 학생 분포는 채점척도 3수준이 다른 채점척도의 수준에 비해 정규분포를 나타났으며, 곤란도의 경우는 채점척도 2수준이나 채점척도 4~7수준보다 채점척도 3수준이 골고루 분포하여 각 수준별로 적절하게 구분되었다고 해석할 수 있다. 과제2, 과제3, 과제4의 관련짓기 능력에 대한 문항특성곡선을 분석해보면 채점척도 4~7수준보다 채점척도 3수준이 좀 더 적절한 수준의 곤란도로 나누어졌다고 해석할 수 있다. 또한 관련짓기 능력에 대한 문항특성곡선에 의하면, 채점척도 3수준이 능력 추정치에 따라 학생을 잘 변별하는 것으로 해석된다. 따라서 본 연구의 과학탐구활동의 수행형 과제를 분석적으로 채점하는 경우 채점척도 3수준이 채점척도 2수준이나 4~7수준보다 더 적절하다고 할 수 있다. 현실적으로 가장 분석적인 채점척도 4~7수준을 사용하는 경우 교사의 시간이나 노력이 더 소모되는 점을 감안하면, 분석적 채점 시 채점척도 3수준으로 충분히 신뢰도나 타당도를 확보할 수 있다는 시사한다.

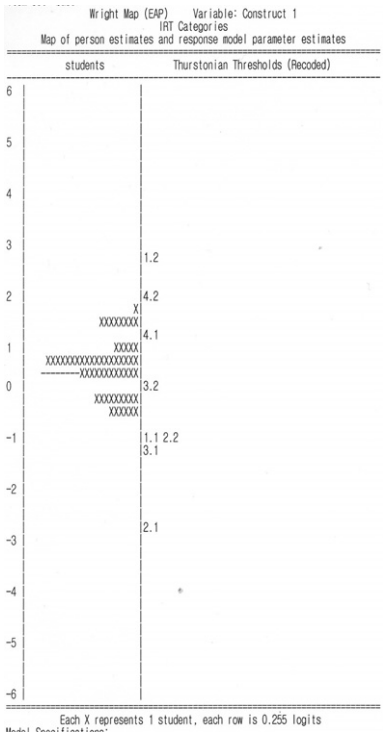
Ⅳ. 결론 및 시사점

본 연구는 중학교 과학탐구활동 수행평가 시 총체적 채점 방식과 분석적 채점 방식의 신뢰도를 비교하고, 분석적 채점 방식의 경우 신뢰도를 확보하기 위한 채점척도의 수를 제안하는 것을 목적으로 하였다. 이를 위해 채점 방식과 채점척도의 수에 따라 중학생의 과학탐구활동지를 두 명의 채점자가 채점한 뒤 신뢰도를 분석하였다. 채점 방식에 따른 신뢰도를 분석해 보면, 총체적 채점 방식의 과제 간 내적일치도가 0.77~0.81로 분석적 채점 방식의 0.72~0.77보다 높게 나타났다. 분석적 채점 방식2의 경우 과제에 따라 채점자 간 신뢰도가 0.85~0.92로 총체적 채점 방식의 0.71~0.90에 비해 높게 나타났다. 정적 집단 비교를 한 결과 분석적 채점 방식은 총체적 채점 방식에 후광효과를 주지 않는 것으로 나타났다. 본 연구 결과에 의하면 분석적 채점 방식은 채점자 간 신뢰도가 높게 나타나고, 총체적 채점 방식은 과제 간 내적일치도가 높게 나타난다고 해석할 수 있다. 이러한 결과는 과학탐구활동을 전체적인 눈으로 평가할 수 있는 총체적 채점을 학교 현장에서 신뢰도 확보가 어렵다는 이유로 채택하지 않을 이유가 없다는 것을 시사한다. 다만 총체적 채점의 평가 방식을 선택하는 경우에는 교사 간협의 등 채점자 간 신뢰도를 확보하기 위한 방안이 요구된다.

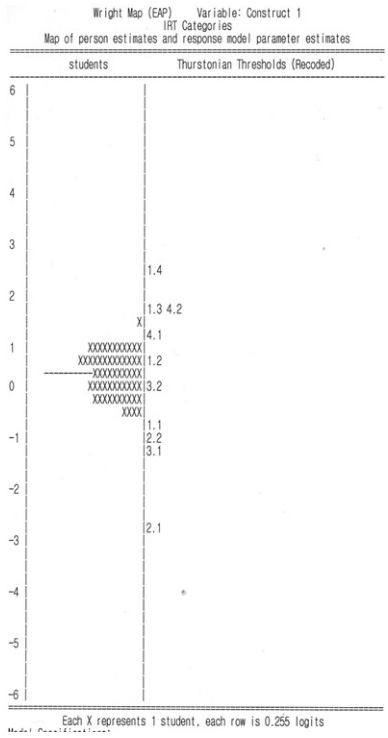
분석적 채점 방식을 선택하는 경우에는 신뢰도과 경제성을 동시에 확보할 수 있는 평가 기준의 수를 설정하는 것이 필요하다. 본 연구 결과에 따르면, 채점척도 3수준과 채점척도 4~7수준의 과제 간 내적일치



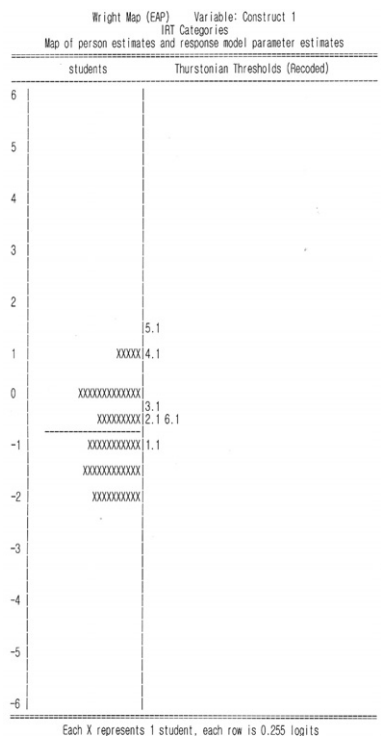
과제1(2수준)



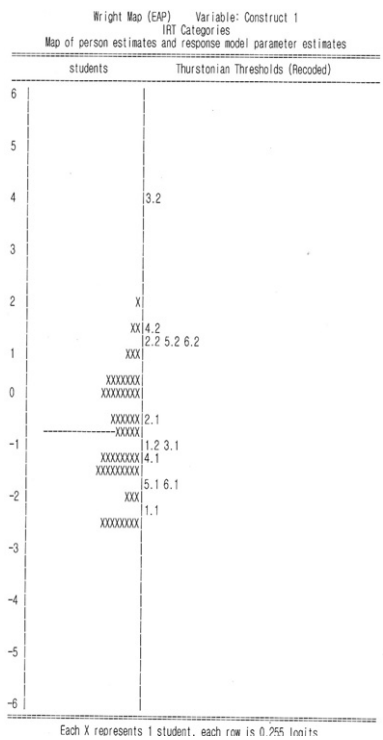
과제1(3수준)



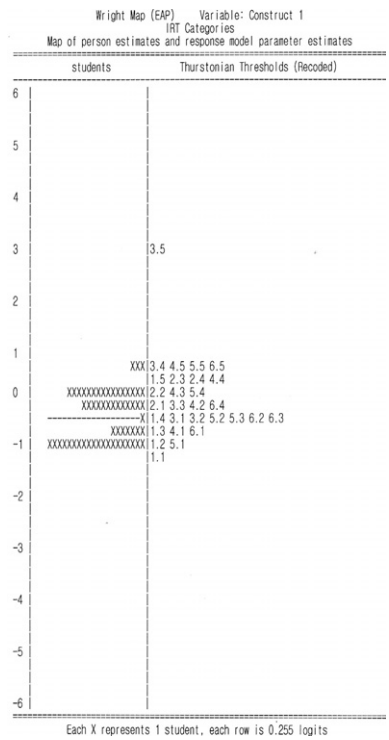
과제1(4~7수준)



과제2(2수준)

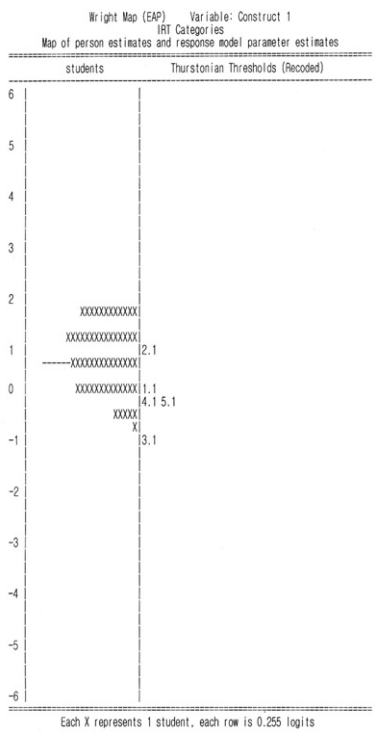


과제2(3수준)

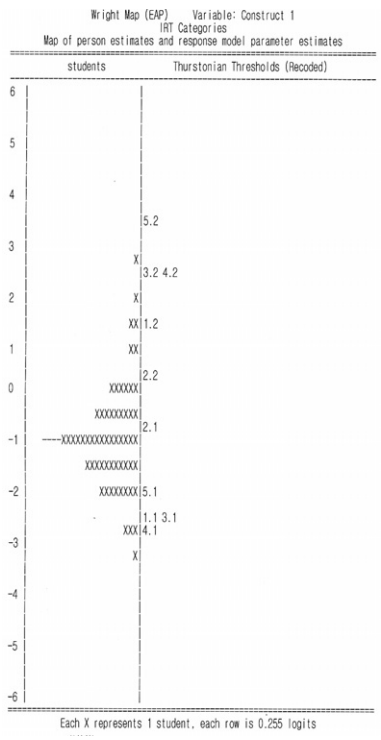


과제2(4~7수준)

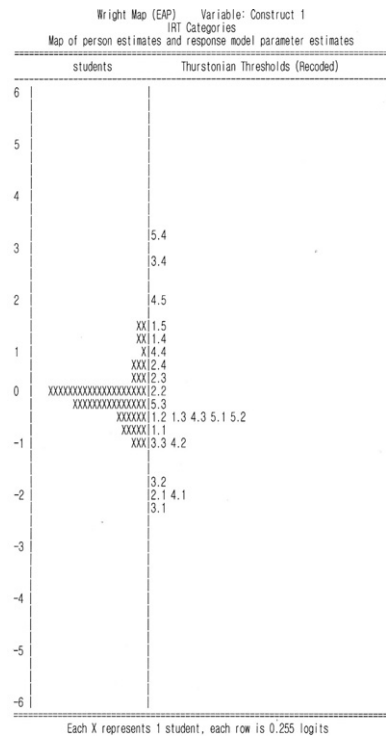
Fig. 1 *평가 수준의 수에 따른 줄기 · 잎 그래프 (과제1, 과제2)*



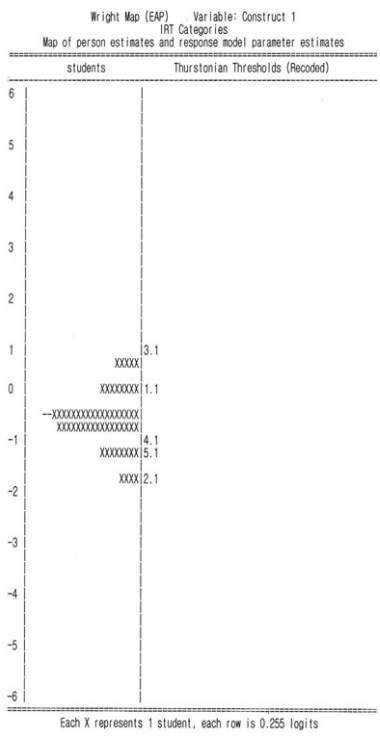
과제3(2수준)



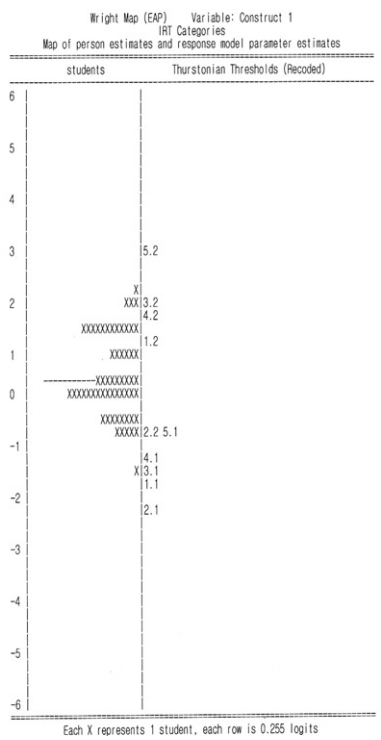
과제3(3수준)



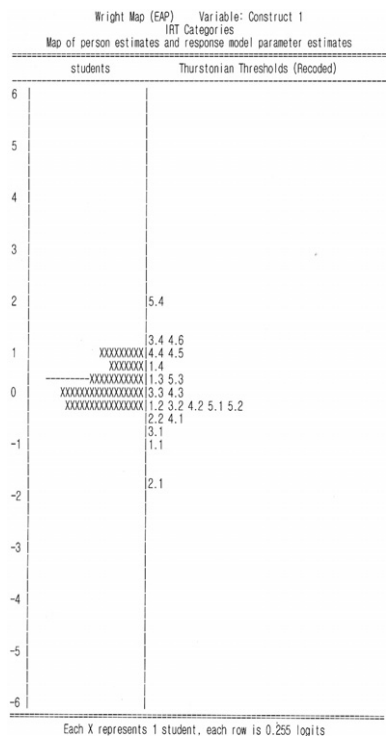
과제3(4~7수준)



과제4(2수준)

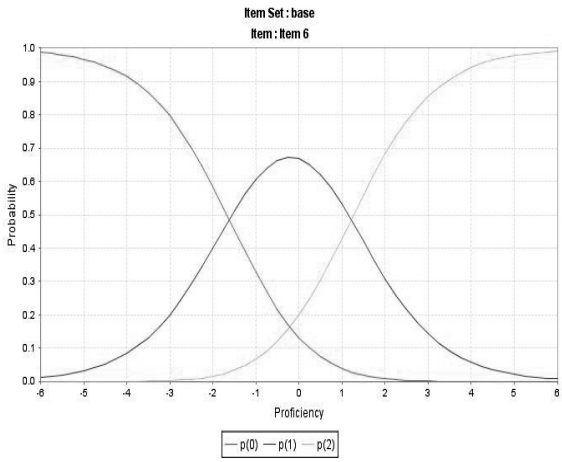


과제4(3수준)

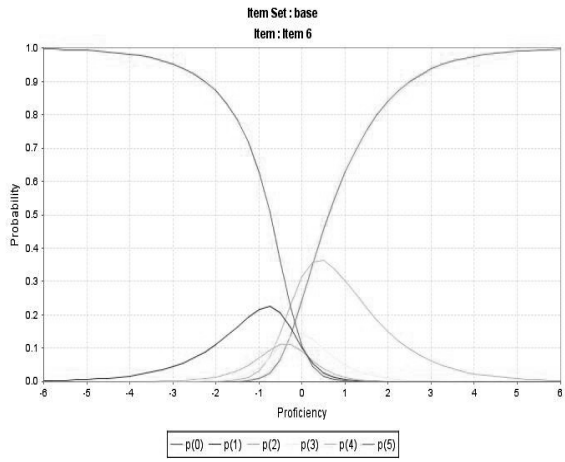


과제4(4~7수준)

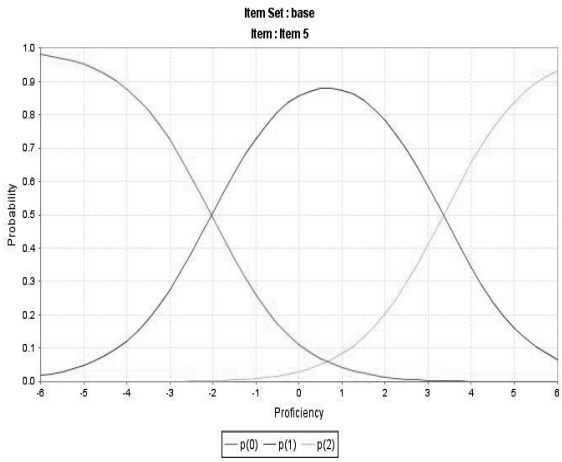
Fig. 2 *평가 수준의 수에 따른 줄기 · 잎 그래프 (활동3, 활동4)*



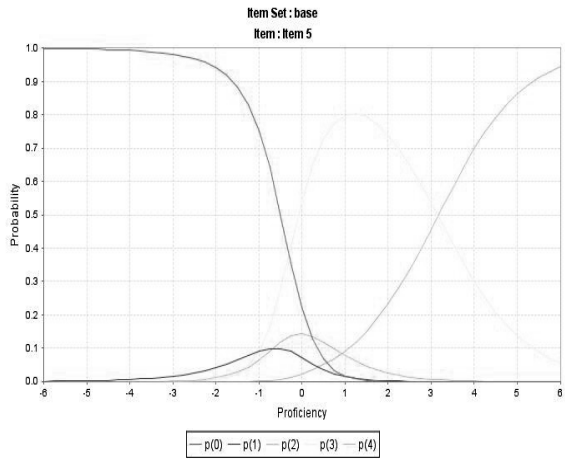
과제 2의 관련짓기 능력(3수준)



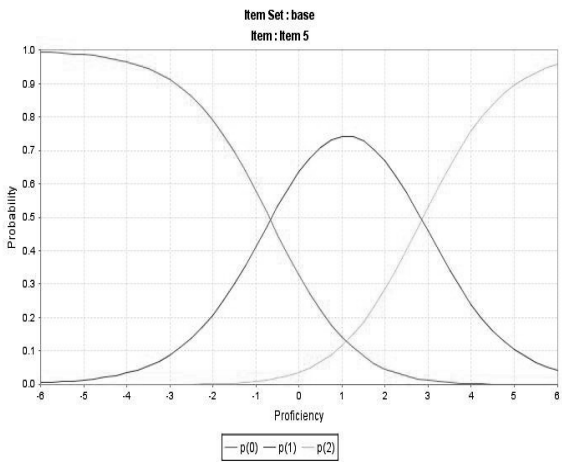
과제 2의 관련짓기 능력(4~7수준)



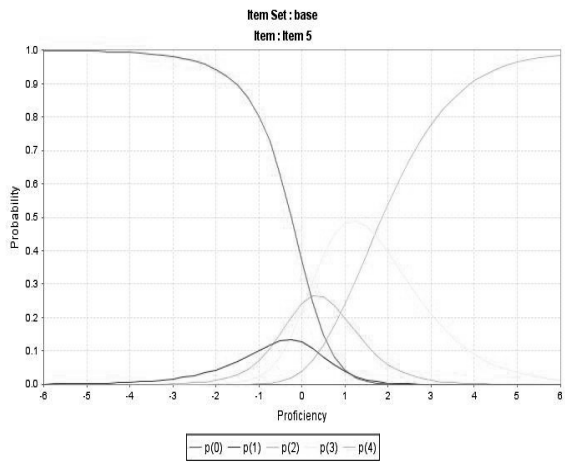
과제 3의 관련짓기 능력(3수준)



과제 3의 관련짓기 능력(4~7수준)



과제 4의 관련짓기 능력(3수준)



과제 4의 관련짓기 능력(4~7수준)

Fig. 3 *평가수준의 수에 따른 문항특성곡선 (과제 2, 과제3, 과제4)*

도가 0.70 및 0.71로 비슷하게 나타났다. 채점척도 3수준에서 과제별 채점자 간 신뢰도가 0.85~0.93으로 다른 경우에 비해 높게 나타났고, 총점에 대한 채점자 간 내적일치도는 0.95로 채점척도 4~7수준의 0.96보다 낮게 나타났지만 그 차이는 통계적으로 유의미하지 않다. 줄기 · 잎 그래프의 분석 결과 능력추정치에 따른 학생 분포는 채점척도 3수준이 다른 채점척도의 수준에 비해 정규분포를 나타났으며, 곤란도의 경우는 채점척도 2수준이나 채점척도 4~7수준보다 채점척도 3수준이 골고루 분포하여 각 수준별로 적절하게 구분되었다고 해석할 수 있다. 따라서 본 연구의 과학탐구활동 과제를 분석적으로 채점하는 경우, 채점척도 3수준이 채점척도 2수준이나 4~7수준보다 더 적절하다고 할 수 있다.

본 연구는 특정한 탐구활동의 수행평가를 제한된 표본을 대상으로 수행되었기 때문에 연구 결과를 일반화하기에는 제한점이 있다. 보다 다양한 대상과 탐구활동에 대한 추가 연구가 필요하다. 또한 총체적인 채점 방식에 대한 신뢰도를 높일 수 있도록 채점기준이나 점수분포에 관련한 채점자 간 협의를 어떻게 진행해야 하는지 추후 연구가 필요하다.

국문 요약

중학생의 과학탐구활동 수행평가 시 총체적 채점과 분석적 채점의 신뢰도를 비교 분석하였으며, 분석적 채점을 하는 경우에는 신뢰도 확보를 위하여 채점척도의 수준을 어느 정도로 분석적으로 해야 하는지를 조사하였다. 중학생들이 작성한 4개의 과학탐구과제에 대한 활동지를 두 명의 채점자가 총체적 채점 방식, 분석적 채점 방식, 분석적 채점 중 채점척도를 2, 3, 4~7수준으로 다르게 하여 채점하였다. 총체적 채점 방식은 과제 간 내적 일치도가 높게 나타났으며, 분석적 채점 방식은 채점자간 신뢰도가 높게 나타났다. 또한 채점척도 3수준의 경우는 4~7수준의 경우와 활동간 내적 일치도와 채점자간의 신뢰도가 유사하게 나타났으나, 능력추정치별 학생의 분포, 문항곤란도 및 문항특성곡선의 경우 채점척도 3수준의 경우가 적절한 것으로 나타났다. 이러한 연구 결과는 과학탐구활동 수행평가 시 총체적 채점 방식을 선택하는 경우는 과제 간 내적일치도를 높일 수 있으며 분석적 채점 방식에 비해 낮게 나타나는 채점자 간 일치도를 높이기 위한 채점자간 협의등 방안이 필요하다는 것을 시사한다. 또한 분석적 채점 방식을 선택하는 경우는 채점척도 3수준으로 충분히 신뢰도를 확보할 수 있다는 점을 시사한다.

참고 문헌

김경희, 송미영 (2001). 채점척도에 따른 채점자의 일관성과 피험자 능력 추정의 정확성 비교. 교육평가연구, 14(1), 327-347.

김명숙 (1999). 영어작문 수행평가의 채점행위 분석 연구. 교육평가연구, 12(2), 25-54.

김석우 (2007). 고등학교 과학과 수행평가 실태분석 및 개선방안. 교육평가연구, 20(4), 53-73.

박정 (2001). 문항반응이론을 활용한 수행형 평가 문항 분석방법. 교육학연구, 39(2), 215-232.

박정, 홍미영(2002). 문항 유형에 따른 과학 능력 추정의 효율성 비교. 한국과학교육학회지. 22(1), 122-131.

성태제 (2002). 타당도와 신뢰도. 학지사.

성태제 (2005). 문항반응이론의 이해와 적용. 교육과학사.

유준희, 박승재 (1999). 과학과 수행평가. 열린교육연구, 7(1), 247-262.

이관용, 김기중 (1993). 기초 심리통계학, 법문사.

이규민 (2007). 초등학교 과학과 수행평가의 총체적 채점과 분석적 채점 방식에 대한 일반화가능도분석. 아동교육, 16(4), 169-184.

이기영, 안희수 (2005). 중등학교 과학 수행평가의 평가 유형과 채점 방식 및 신뢰도 분석. 한국과학교육학회지, 25(2), 173-183.

지은림 (1999). 사회과 보고서 수행평가를 위한 총체적 채점과 분석적 채점의 비교. 교육평가연구, 12(2), 11-24.

지은림 (2000). 논술형 수행평가를 위한 채점방법들의 비교. 경희대학교 교육문제연구소 논문집, 16, 235-246.

한국교육과정평가원 (2001). 제7차 교육과정에 따른 성취기준 평가기준.

American Association for the Advancement of Science. (2001). ATLAS of science literacy. Vol. 1. American Association for the

Advancement of Science.

Black, P. J. (1990). APU science – the past and the future. School Science Review, 72(258), 28–43.

Black, P. J. (1998). Testing: friend or foe? : Theory and Practice of Assessment and Testing. Falmer Press.

Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., Rosengrant, D., & Warren, A. (2006). Developing and assessing student scientific abilities. Physical Review Special Topics – Physics Education Research, 2(2), 020103-1-020103-15.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. International Journal of Science Education, 25(12), 1509–1528.

Halonen, J. S., Bosack, T., Clay, S., & McCarthy, M. (2003). A rubric for learning, teaching, and assessing scientific inquiry in psychology. Teaching of Psychology, 30(3), 196–208.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA; ASCD.

Klein, S. P., Stecher, B. M., Shavelson, R., McCaffrey, D., Bell, R. M., Comfort, K., Othman, A. R., & Ormseth, T. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11(2), 121–137.

Linn, R. L., Baker. E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15–21.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O' Sullivan, C. Y., Arora, A., & Erberber, E. (2005). TIMSS 2007 assessment frameworks. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Plake, B. S., & Hambleton, R. K. (1999). A standard-setting method designed for complex performance assessments: categorical assessments of student work. Educational Assessment, 6(3), 197–215.

Qualification and Curriculum Authority (2007). Science: Programme of study for key stage 4 in the national curriculum 2007. London: Qualifications and Curriculum Development Agency.

Roberts, L., Wilson, M., & Draney, K. (1997) The setup assessment system: An overview. BEAR report series, SA-97-1, University of California, Berkeley.

Waltman, K., Kahn, A., & Koency, G. (1998). Alternative approaches to scoring: The effects of using different scoring methods on the validity of scores from a performance assessment. CSE Technical Report, 488.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. Applied Measurement in Education, 13(2), 181–208.

Wilson, M., Sloane, K., Roberts, L., & Henke, R. (1995). Setup course Ⅰ, issues, evidence and you: Achievement evidence from pilot implementation. University of California, Berkeley.

Woolnough, B. E. (1989). Toward holistic view of precesses in science education, in J. Wellington (Ed.) Skills and processes in science education: a critical analysis(pp. 115–134). London: Routledge.