과학 수행 평가 문항의 선정 및 제작을 위한 평가 준거의 개발

김은진* · 박현주** · 강호감*** · 노석구***
(부산교육대학교*) · (조선대학교**) · (인천교육대학교***)

Developing of a Criterion for Selecting and Producing of Performance Assessments in Science Education

Eun-Jin Kim* · Hyun-Ju Park** · Ho-Kam Kang*** · Suk-Gu Noh***
(Pusan National University of Education*) · (Chosun University**) · (Inchon National University of Education***)

ABSTRACT

We have developed of a criterion that would help to select and develop performance assessments in science education. There are six categories of a criterion: Fidelity, satisfaction, content validity, fairness & suitability, reliability & objectivity, and usability. There are the total of 25 sub-categories under the six categories. Five science educators evaluated the validity of the criterion. For reliability of the criterion, Kendall's tau-b was used. Eight science educators and elementary teachers evaluated three performance assessment tasks for the correspondence of the criterion. This study also discuss the implications of this criterion as well.

Key words: performance assessment, criterion of selecting assessment tool fidelity, satisfaction

I. 서 론

세계화, 정보화, 다양화 사회로의 변화는 자율적이며, 다양한 개성과 고등사고 능력을 갖춘 인간의 육성을 필요로 하고 있다. 따라서 과학교육에서는 이러한 능력들을 육성할 수 있도록 그 목표를 정하고 교수하며 그에 따르는 적절한 평가가 이루어져야한다.

1980년대 말까지 평가의 대명사였던 선다형 지필검사는 그 측정의 한계를 극복하지 못하고 90년대에 들어서 수행평가 등의 여러 가지 대안평가에게 자리를 내주었다(Baron, 1991; Gable, 1996; Hart, 1994; Tarmir, 1998). 우리 나라의 학교 현장에서도 대부분 수행평가가 이루어지고 있다. 그러나, 과학 교육 현장에서 사용되는 수행평가도구들은 과학의 본성과 특성을 인식하고 수행평가에 대한 충분한 이해를 통해서 이루어진 수행평가 도구의 개발 및 실행이라기 보다

*2002.9.30(접수) 2002.12.10(1차 통과) 2002.1.14(최종 통과)

는 종래의 지필 평가가 아닌 관찰법, 실험실기, 실험 보고서 작성 등의 과학 실험활동과 관련된 평가로 이해되고 있는 경우가 많고(김영순, 1999; 임영득 등, 1999; 임영득 등, 2001), 심지어 우리 나라 현실에 적합치 않은 상태로 번역되어 혼란을 가중시키는 경우도 있었다(남명호, 1995). 이렇게 수행평가 문항에 대한 문제들이 야기되고 있는 것은 수행평가 문항의 질을 판단할 수 있는 적절한 준거가 존재하지 않기 때문이라 볼 수 있다.

일반적으로 특정 목적을 가지고 평가문항을 제작할 때에는 그 제작 과정에 있어서 평가문항의 질을 판단할 수 있는 준거가 제시되는 것이 통례였으며, 이들 준거를 통해 일선 교사들이나 교육관계자들은 평가문항의 질을 판단하였다. 그러나 수행평가는 측정관에 기초한 전통적인 평가들과는 그 기본적인 전제와 목적이 다르므로 기존의 전통적인 평가문항들의 질을 평가하기 위해 사용했던 준거들을 수행평가에 그대로 차용하는 것은 부적절하다. 따라서 수행평가의 특성과 기본목적을 살린 수행평가문항을 위한 평가준거가 만들어져야한다(Linn *et al.*, 1991).

따라서 본 연구에서는 과학교육에서 수행평가문항을 개발하거나 선별하여 사용하고자 할 때, 문항의 질을 판단하기 위한 지침으로서 과학 수행 평가 문항에 대한 평가 준거를 개발하였다.

Ⅱ. 연구 방법 및 절차

과학 수행 평가 문항의 선정 및 제작을 위한 평가준거는 2000년 6월부터 2002년 4월에 걸쳐 제작되었다. 본 준거의 초안은 본 연구팀인 과학교육연구자 4인에 의해 문헌검토와 국내외 과학수행평가문항에 대한 검토 및 연구자간 논의를 통해 구성되었다. 연구의 구체적인 절차는 다음의 6단계로 볼 수 있다.

1. 과학수행평가문항의 질을 결정하기 위한 평가준거로서 대 범주의 개발 완성

본 연구팀은 수행평가에 대한 이론적인 문헌연구와 국내외에서 개발된 다수의 수행평가문항들을 수집하여 검토하고 이를 통해 과학수행평가문항의 질을 결정하기 위한 평가준거 개발의 필요성을 인식하였고, 이에 일차적으로 과학수행평가문항을 위한 평가준거의 대 범주 추출의 이론적 배경을 이끌어내었고, 이로부터 수행평가의 특성과 기본목적에 대한 고려 및 심리측정학에서의 준거들을 고려하여 대 범주를 개발하였다.

2. 대 범주에 따른 하위항목의 추출 및 논의

개발된 대 범주에 대한 각 하위항목은 대 범주에 따른 보다 실제적인 항목으로서 직접 문항의 질을 판단하는 구체적인 기준이 될 수 있도록 최대한 단순하고 명확한 용어로 기술하였다. 한편, 대 범주의 중요성을 고려하여 각 범주별로 하위항목의 수를 조절하므로써 대 범주의 중요성에 따른 가중치를 따로 계산하지 않아도 되도록 하였다. 즉, 첫째 범주로서 "과학 교과목표에 비추어 본 문항 평가 목표의 충실도"는 8개의 하위항목이 있는 반면, 다섯 번째 범주로서 "문항의 신뢰도와 객관도"는 3개의 하위항목을 가지고 있어서 항목 당 같은 배점을 하여도 범주에 따라 그 총점이 달라지게 되므로 가중치를 준 효과를 가져올 수 있다.

3. 본 준거를 점검표 형식으로 구성

개발된 대 범주와 그에 따른 하위항목을 실질적인 점검표의 형식으로 연구팀이 개발 작성하였다(부록). 준거표는 각 범주별로 소계를 구하여 기록할 수 있는 공간들 두었고, 전체적인 총점도 구하여 기록할 수 있도록 하였다. 그리고 평가자간 의견란을 두어 나중에 다른 사용자가 이미 평가된 도구를 사용하고자 할 때 참고할 수 있도록 하였다. 범주별 소계의 기록공간은 사용자가 특정한 범주의 점수에 관심이 있을 경우 참고할 수 있도록 하기 위한 의도에서 마련하였다. 또한 하위 항목은 총 25개로서 항목 당 2점 만점을 받을 경우 50점이 되고, 이를 2배수로 할 경우 100분위 점수로 변환하기에 용이하므로 여러 과학수행평가도구가 본 준거표로 평가받았을 경우, 사용자

가 원한다면 이들 도구들 간의 질을 쉽게 비교할 수 있도록 하였다.

4. 과학교육 전문가에 의한 평가준거의 타당도 검증 및 검토

개발된 평가준거의 타당도 검증은 과학 수행평가 내용전문가 5인이 내용타당도를 검증하는 방식으로 하였다(정종진, 1999; 황정규, 1998). 타당도 검증에 참여한 과학 수행평가 내용 전문가는 과학교육학 박사 3명과 중등 과학과 교직 경력을 가진 과학교육전공 박사과정생 2명이었다. 아울러 이들은 본 준거표의 형식적인 면과 문장의 어법과 의미 등에 관한 표현들을 검토하였고, 이를 토대로 1차 수정을 하였다

5. 과학교육 전문가 및 초등예비교사에 의한 본 준거의 신뢰도 검증

본 준거의 신뢰도의 검증을 위하여, 과학교육전문가 2인과 초등예비교사 6인에 의해 과학 수행평가도구에 대한 본 준거의 적용이 이루어졌다. 신뢰도 검증을 위해 사용된 문항은 3가지로 각각 H연합에서 개발한 자연과 초등학교 5학년 1학기 수행평가 문항, K교육청 장학자료로 배부된 초등학교 6학년 1학기 수행평가문항, 그리고 I교육대학 과학교육연구소에서 자체 개발한 초등학교 1학년 2학기 자연과 수행평가문항이었다.

신뢰도 검증은 과학수행평가도구에 대한 8명의 준거 점수 상관을 통한 그 합치도로 알아보았으며, 이 때 변인의 특성을 고려하여 순위변수(rank variable)를 위한 상관계수인 Kendall 상관계수(Kendall's tau-b)를 사용하였다.

6. 평가준거의 완성

점검표 형식으로 제작된 과학 수행 평가 도구의 선정 및 제작을 위한 평가준거의 형태를 다듬고, 그에 대한 사용 설명을 덧붙임으로써 완성하였다.

본 연구의 절차는 Fig. 1과 같다.

Ⅲ. 결과 및 논의

1. 과학 수행 평가 도구의 선정 및 제작을 위한 평가준거의 개발

1) 대범주의 추출

수행평가는 심리측정학에 기초한 전통적인 평가들과 비교할 때 기본적인 목적과 전제가 다르다. 즉, 전통적인 평가문항들은 수험자의 차이를 알아내기 위한

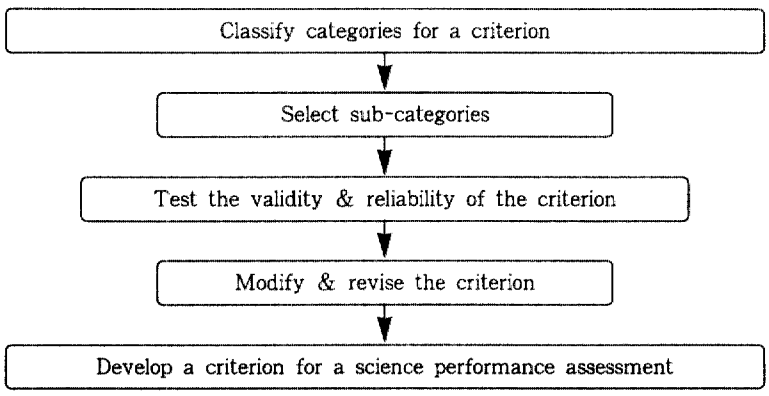

Fig. 1. The procedure of this study

것이 우선적인 목표였으므로 수험자 자신의 학습과정이나 학습결과 보다는 전체 집단 속에서 수험자 개개인의 위치를 변별해내는 것이 더 중요했으나 수행평가는 학습자 개인의 학습과정을 모니터링하고 그 효과를 피드백하는 데에 최우선의 목적이 있다. 따라서 기존의 전통적인 평가 문항들의 질을 평가하기 위해 사용했던 기준이나 준거들을 그대로 차용하는 것은 부적절하다(Linn *et al.*, 1991). 그러나 이 준거들 중 평가의 일반적인 의미를 반영하는 성격의 기준들은 본 평가준거에서도 고려하였다. 그러므로 과학수행평가 문항 평가준거 개발의 첫 단계인 대 범주 추출에 있어서 가장 중요한 조건은 수행평가의 특성과 기본 목적에 대한 고려이며, 심리측정학에서 사용하는 일반적인 준거들도 함께 고려하였다.

아래 Table 1에 과학과 수행평가문항의 평가준거개발을 위한 대 범주를 추출하여 제시하였다.

Table 1. Selected categories for a criterion

	Categories
1	Fidelity
2	Satisfaction
3	Content Validity
4	Suitability & Fairness
5	Reliability & Objectivity
6	Usability

(1) 과학 교과 목표에 대한 충실도(Fidelity)

어떤 교수활동이나 평가활동을 할 때 1차적으로 목표를 명확히 해야하며, 그 목표는 상위 목표와 부합되어야 한다. 따라서 이 범주는 모든 평가 문항의 질을 평가할 때 가장 우선적으로 고려되어야할 사항이며, 가장 중요한 범주이다. 본 연구에서 이 범주는 현행 과학교과목표 뿐 만 아니라 수행평가를 통해 놓쳐서는 안될 과학 평가목표에 대한 영역들도 포함시켰다(김은진, 2000; Baron, 1994; Hart, 1994; Kim et al, 2000; Tamir, 1998).

(2) 수행평가의 특징과 구성 요소에 대한 만족도 (Satisfaction)

이 범주는 과학과 수행평가문항이 수행평가의 기본 목적과 특징을 토대로 그 장점을 얼마나 잘 살리고 있는가를 평가하는 범주이다. 이 범주는 수행평가 문항의 구성요소인 수행과제, 반응양식, 채점체계에 대한 만족도 뿐 만 아니라 각 구성요소를 중심으로 수행평가의 특징과 장점을 살릴 수 있는 요건들이 포함되어있는 지를 점검하고, 평가자의 다양화와 학습자를 위한 선택의 기회 여부와 같은 요건들도 포함하였다(Bransford & Stein, 1984; Brown & Shavelson, 1996; Glaser, 1984; Herman, *et al.*, 1992; Linn *et al.*, 1991; Resnik & Klopfer, 1989).

(3) 문항의 내용 타당도 (Content Validity)

전통적인 평가문항의 내용 타당도는 특정 문항의 정답을 기준으로 보았다. 그러나 수행평가에서 타당도는 정답을 제한할 필요가 없고, 학습자의 모든 반응을 정보로 간주한다는 점에서 전통적인 평가문항의 내용 타당도와 비교할 때 더 융통성이 발휘되어야한다(Bodin, 1993; Linn, *et al.*, 1991). 즉, 수행평가 문항은 학습자의 다양한 반응에 대한 수용과 학생들의 과제 해결과정에서 표출되는 다양한 양상에 대한 인정과 의미를 부여할 수 있도록 제작되어야한다. 그리고 문항에 담고 있는 내용은 정확한 지식으로 진술된 것이어야 한다.

(4) 학습자에 대한 내용의 적합도 및 공정성 (Suitability & Fairness)

문항의 평가목표가 정해졌을 때 평가 문항의 내용과 소재가 평가 대상으로 하는 학습자의 수준에 적합할 뿐 만 아니라 다양한 환경의 학습자들에 대한 공정성을 유지하므로써 환경에 의한 또는 성별이나 특이적 조건에 의한 불이익을 당하지 않도록 배려하고 있어야한다.

(5) 문항의 신뢰도와 객관도(Reliability & Objectivity)

수행평가에서는 평가자가 직접 관찰하고 판단하는 평가과정이 많으므로 이 두 가지는 거의 같은 의미를 지닌다. 규준지향 평가의 경우 문항의 신뢰도는 매우 중요한 항목이었으나 목표 지향 평가에서는 신뢰도는 있으면 좋은 바람직한 부수적인 조건으로 여긴다(황정규, 1991). 수행평가는 문항자체의 신뢰도보다는 평가자 스스로 일관성을 유지하고, 평가자간 신뢰도를 높일 수 있도록 문항 작성시에 채점체계를 명확히 하

고 상세화하므로써 보완될 수 있다(Solano-Flores *et al.*, 1997).

(6) 사용상의 편의성 및 실용성(Usability)

이 범주는 교사가 수업에 직접 문항을 투여할 때 요구되는 교수활동 이외의 노력에 관한 것으로 수업 상황, 교사의 시간요구, 채점의 편리성, 소요 비용 등의 측면을 고려할 때 적절하게 사용할 수 있는가 하는 점에 관한 것이다(Brown & Shavelson, 1996).

2) 대 범주에 따른 하위항목의 추출

위에 제시된 6가지 대 범주에 따른 하위항목은 다음과 같다.

(1) 과학 교과 목표에 대한 충실도에 대한 하위항목

① 문항에서 과학적 탐구 사고력의 요소가 평가되고 있는가? - 이 항목은 과제의 해결과정에서 학습자가 합리적 사고 과정이나 추론, 증거에 기초한 판단 등의 과학적 사고를 통한 과학적 탐구과정을 경험할 수 있도록 문항이 구성되어있는가를 의미한다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

② 과학적 탐구를 위한 수공기능이 평가되고 있는가? - 이 항목은 과제의 수행과정에서 학생들이 과학탐구에 수반되는 여러 가지 수공기능중 한 가지 이상을 체득할 수 있는 기회를 제공하고 있는가를 의미한다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

③ 과학적 태도 영역이 평가되고 있는가? - 이 항목은 과제를 수행하는 과정에서 학습자가 과학과 관련된 가치를 느끼고 과학적 태도, 즉, 호기심, 합리성, 판단의 일시보류, 개방성, 비판성, 객관성, 정직성, 겸손, 과학지식의 한계에 대한 인식과 같은 태도를 고양할 수 있도록 그 내용이 구성되어있는가를 의미한다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

④ 과학적 지식의 적용력이 평가되고 있는가?- 이 항목은 문항의 평가 목표에 있어서 특정 지식을 학생들이 알고 있는가와 같은 단순사고가 아니라 그 지식을 직접 체득하여 적용할 수 있는가와 같은 보다 고등 사고기능을 평가할 수 있는 내용이 갖추어져 있는가를 의미한다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998). 특히 지식의 적용 상황에 있어서 일상적인 상황을 제시함으로써 일상경험 속에서 과학지식을 적용할 수 있는지를 평가하는가 살펴볼 수 있다.

⑤ 창의적 사고영역이 평가되고 있는가? - 이 항목은 과제 내용의 새롭고 기발함이 아니라 학습자로 하여금 특이하고 기발한 생각을 할 수 있도록 유도하는 과정이 포함되어 있는가를 의미한다. 과제 내용의 기발함은 학습자의 흥미를 끌 수는 있겠지만 그렇다고 반드시 학습자의 창의적 사고를 이끌어낸다고는 말할 수 없다. 사전에 정답이 있음을 예고한다든가 또는 정답이 있을 것 같은 복선을 문항 속에 제시하는 것은 학습자의 창의적 사고를 저해한다. 뿐 만 아니라 정해진 시간 안에 답을 찾아낼 것을 요구한다든가 학습자에게 무언의 압력을 받게 하는 문항이라면 학습자는 창의적 사고를 할 수 없다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

⑥ 타인과의 의사소통력이 평가되고 있는가? - 이 항목은 과제의 수행과정에서 타인과 의견교환이 요구되는지를 의미한다. 반드시 문항 속에 타인과 토의해 보라는 언급이 없더라도 내용의 흐름상 수업을 주도하는 교사에 의해서 의견교환의 단계를 넣을만한 여지가 있다면 의사소통과정이 있는 것으로 볼 수 있다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

⑦ 반성적 사고과정이 평가되고 있는가? - 이 항목은 학습자가 과제를 해결해가는 과정에서 생각하고 판단하고 평가하는 등의 반성적 사고를 할 수 있도록 유도하는가를 의미한다. 학습자의 반응양식에서 구체적으로 여러 가지 생각할 수 있는 여유를 둘 수도 있지만 일지 작성(journaling)이 포함된 과제라면 대부분 반성적 사고과정의 기회를 주고 있다고 볼 수 있다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

⑧ ICT를 활용한 정보수집과정이 포함되어있는가? - 이 항목은 과제의 해결과정에서 학습자가 인터넷이나 문헌 정보 등의 각 종 정보를 활용할 수 있도록 정보를 주는가를 의미한다. 즉, 중요한 또는 관련

사이트의 탐색 기능이 있는 인터넷의 URL을 알려준다든지 주요 문헌정보의 제목과 소장처를 알려준다든지 하는 정보를 담고 있다면 정보 수집 면에서좋은 문항으로 평가될 수 있다(김은진, 2000; Baron, 1994; Hart, 1994; Kim *et al.*, 2000; Tamir, 1998).

(2) 수행평가의 특징과 구성 요소에 대한 만족도 범주의 하위항목

① 수행과제가 학습자의 동기유발을 위한 요소를 포함하고 있는가? - 실제적(authentic)이고, 일상적인 상황을 제시하고 있는가? 이 항목은 매우 중요한 항목으로서 학습자의 학습 태세를 유발할 수 있는가를 결정하는 관건이 된다. 이 항목은 과제의 내용이 얼마나 기발하며, 실제적이고, 흥미로워서 학습자를 끌어들일 수 있는가 하는 것이다(Herman, *et al.*, 1992; Larkin, 1989; Resnick & Klopfer, 1989).

② 반응 양식 중 학습자를 위한 학습의 피이드 백 과정이 포함되어있는가? - 이 항목은 학습에 대한 모니터링 결과를 다시 학습에 투입해서 피이드 백 하는 절차가 포함되어 있는지를 의미한다(Brown, *et al.*, 1983; Brown & Shavelson, 1996; Glaser & Pellegrino, 1987).

③ 반응 양식 중 학습자의 학습과정을 모니터링 할 수 있는 과정이 포함되어 있는가?- 이 항목은 학습자 스스로가 그리고 교사나 다른 평가자가 학습자의 학습과정과 상태를 모니터링 할 수 있는 요소들을 갖추고 있는가하는 것이다(Brown, *et al.*, 1983; Brown & Shavelson, 1996; Glaser & Pellegrino, 1987).

④ 자기평가, 교사평가, 학부모 평가 등 다양한 평가자에 의한 평가 절차가 있는가? - 이 항목은 수행평가의 특징 중 하나인 평가자의 다양화에 관한 항목이다. 학습자의 학습에 대한 평가자를 다양화하므로써 학습과정과 결과에 대한 신뢰도를 높이고 과학교육관련자들 모두의 관심을 높일 수 있다.

⑤ 학습자 주도적인 수업이 이루어질 수 있도록 구성되어있는가? - 이 항목은 과제가 학습자의 역할과 행동을 얼마나 제한하고 있는가를 의미한다. 또한 이 항목은 학습자에게 얼마나 다양한 선택의 기회를 제공하고 있는가를 포함한다.

⑥ 수행평가문항이 수행과제, 반응양식, 채점체계를 모두 갖추고 있는가? - 수행평가의 구성요소에는 수행과제, 반응양식, 채점체계가 있다(Brown & Shavelson, 1996). 따라서 이 항목에서는 이 3가지 요소들이 모두 갖추어져 있는지가 중요 기준이 된다.

⑦ 학생의 반응에 대한 평가 체계에 융통성이 있는가?- 학생의 다양한 반응의 수용을 통해 학생과 교사와의 상호작용을 증진하고 학생의 창의적 사고력과 반성적 사고력을 배양하기 위해서 평가체계에 평가자의 융통성을 발휘할 수 있는 여유를 두어야한다.

(3) 문항의 내용타당도 범주의 하위항목

① 평가목표에 제시된 과학개념과 문항의 내용이 적합한가? --이 항목은 문항을 사용할 교사가 직접 판단할 수도 있으나, 문항의 제작사에서 과학 내용 전문가 및 과학교육 전문가에 의해 타당성을 입증 받고 문항에 직접 표시하여 배포한다면 이를 믿고 선택할 수 도 있다.

② 문항에 포함되어있는 과학지식이 정확한가? - 이 항목은 문항의 타당도와는 다르다. 즉, 문항의 내용이 평가 목적에 부합된다 할 지라도 과학 지식 자체가 잘못된 과학 내용을 담고 있다면 더 큰 문제를 야기할 수도 있다. 따라서 이 항목 또한 매우 중요한 항목이다.

③ 문항의 표현 중 모호한 부분이 없고 명확한가? - 내용진술은 명확하여야한다. 모호한 표현은 학습자의 혼란을 유발할 수 있고, 오히려 오개념을 불러일으킬 수도 있다.

(4) 학습자에 대한 내용의 적합도 및 공정성 범주의 하위항목

① 문항의 내용이 학습자의 인지적 수준에 적합한가? - 이 항목은 문항의 내용과 표현 방식이 대상 학습자의 인지 수준에 적절한가를 의미한다. 단 통합학년을 대상으로 하는 문항의 경우에는 문항 속에 대상 학생의 수준에 따른 과제 해석의 정도와 수준이 포함되어 있어야하며, 그것이 없다면, 교사용 지도서에 나타나있는 과제의 적용 수준에 대한 지침의 정도에 따라 평가하도록 한다.

② 문항의 내용이 학습자의 문화적 환경, 생활 환경면에서 그리고 성별이나 지역간 차이 등을 고려할

때 학습자에게 적합하며, 공정한가?- 이 항목은 문항의 사용시 교사가 직접 자신의 학생들을 비추어 판단하게 되지만, 우수한 수행평가문항이라면 특수한 문화환경에 구애받지 않도록 그리고 성별이나 문화적 차이 등에 의해 불이익을 당하지 않도록 제작되어야 한다.

(5) 사용상의 편의성 및 실용성 범주의 하위항목

① 문항 사용을 위한 사전 사후 준비면에서 적절한가?- 이 항목은 교사가 문항을 교실현장에 투입하려 할 때 투입 전후에 요구되는 시간의 정도가 얼마나 되는가하는 것이다. 즉, 교사용 지도서에 문항에 대한 소개와 자세한 안내가 포함되어있는지, 현장 적용을 위한 각종 양식의 예가 있는지, 그리고 정보를 얻을 수 있는 출처를 제시해주고 있는지 등을 점검한다. 이러한 안내가 충분히 되어있다면 교사가 문항의 사용을 위해 소요해야하는 평가외 시간의 요구량은 줄어들 것이다.

② 문항의 현장 투여시 시간면에서 적절한가? - 이 항목은 문항의 투입에 소요되는 시간이 실제 학교에서의 수업시간과 비교해 보았을 때 적절한가하는 것이다.

③ 문항사용의 비용면에서 경제적인가? - 이 항목은 문항을 적용할 때 소요되는 비용에 관한 점검 항목이다. 과학과 수행평가는 흔히 소모품 등의 비용이 소요되므로 그 정도가 실제적인 수준을 넘어선다면 아무리 좋은 문항이라도 현장 적용에는 어려움이 있을 것이다.

(6) 문항의 신뢰도와 객관도 범주의 하위항목

① 상세화된 채점 체계를 가지고 있는가? 채점 체계가 상세화될수록 채점결과의 신뢰도는 높아진다. 그러나 지나친 상세화는 오히려 평가에 대한 융통성을 떨어뜨리므로 수행평가의 장점을 발휘할 수 없게 한다. 따라서 필수적인 내용이나 직접적으로 평가 목표와 관련되는 것만 채점 체계를 만들고, 지나친 상세화는 지양하도록 한다. 각 채점 항목의 등급은 3등급 정도가 적당하다(Shavelson & Ruiz-Primo, 1997).

② 평가자가 채점기준에 대해 충분히 이해할 수 있도록 제시되어 있는가? - 수행평가는 평가자 직접 관찰을 통해서 평가하게되는 경우가 많으므로, 채점의 기준에 대한 충분한 설명을 통해 평가자의 차이에 의한 오차를 줄여준다면 신뢰도를 더 높일 수 있다(박도순과 홍후조, 1999; 유광렬, 1997; 황정규, 1991).

3. 평가 준거표의 개발

위에서 논의된 과학과 수행평가문항을 위한 평가준거들을 실제로 사용하기에 용이하도록 항목별로 점수를 부여하고 표로 제작하였다(부록 1).

점수는 3등급으로 0, 1, 2 로 주며, 0은 대상으로 하는 문항이 평가 항목의 내용을 포함하고 있지 않은 경우, 1은 갖추고는 있으나 충분치 못한 경우, 2는 항목의 내용을 적절히 갖추고 있는 경우이다. 점수란은 대 범주별 합계점수와 총점을 알아보기 쉽게 하기 위하여 각 범주별 합계점수 기록란을 마련하였다. 대 범주별 합계점수의 비율은 수행평가에 있어서 범주의 중요성을 반영한다.

4. 평가준거의 내용 타당도 검증 및 평가준거표의 구성 검토

평가준거의 내용타당도를 검증하기 위하여 과학과 수행평가문항 또는 평가문항의 제작 경험이 있는 과학교육학 전문가 5명에게 전체 항목 25개에 대한 내용타당도를 검증 받았다(정종진, 1999; 황정규, 1998). 검증 결과, 과학교육학 전문가 5명 모두 본 준거의 내용이 과학 수행 평가문항의 질을 판단하기 위한 준거로서 적합하다는데 의견의 일치를 보였다. 그리고 평가준거표의 구성으로서, 안내문, 표의 형태, 문장의 표현 등 형식적인 면을 검토하여 수정함으로써 사용상의 편의를 높였다.

5. 평가준거의 신뢰도 검증

평가준거의 신뢰도의 검증을 위해 동일한 문항에 대한 평가자간 합치도를 구하는 방식을 사용하였다. 이를 위해 평가자간 일치도 계수로서 Kendall 상관계수(Kendall's tau-b)를 구하였다. 신뢰도 검증을 위해 초등학교 과학 수행평가문항 3종류를 활용하였

고, 평가자로 과학 수행평가문항제작의 경력을 가진 과학교육학 박사 2명과 과학 수행평가제작의 경험이 없으나 수행평가에 대한 이론적인 학습을 한, 초등학교 예비교사 6명이 참여하였다.

3가지 과학 수행평가문항에 대한 이들의 평가점수를 통한 Kendall 계수가 아래 Table 2에 제시되어 있다. 이 결과에 따르면 각 문항에 대한 Kendall 계수는 각각 0.27, 0.024, 0.009로서 이들의 통계적 유의치는 각각 69.1%, 76.5%, 98.0% 이다. 따라서 이 세 문항에 대한 평가자간 합치도는 높다고 할 수 있다.

6. 준거표의 완성

개발된 평가준거를 실제로 과학교육 현장에서 사용할 수 있도록 점검표 형식을 빌어 평가준거표로 제작하고 그에 대한 사용 설명을 덧붙였다. 평가자의 의견란은 이후에 또 다른 사용자에게 도움이 될 수 있는 의견이 있다면 적어두도록 하기 위한 기록 공간으로 마련하였다(부록 1).

본 연구에서 개발한 '과학 수행 평가 문항의 선정 및 제작을 위한 평가준거'는 과학수업에서 사용되는 수행평가문항의 이상적인 내용과 형태를 제안한다. 뿐 만 아니라 과학 수행 평가에서 고려해야하는 범주를 제시하고 특히, 중요시 해야하는 범주에 대해서도 제안하고 있다. 따라서 본 준거표의모든 항목을 만족시키지는 못한 문항이라도 특정 범주에서 높은 점수를 받은 문항은 나름대로 바람직한 특성을 가진 문항이라고 판단될 수 있다. 특히 첫 번째 범주와 두 번째 범주는 과학 수행평가에 있어서 가장 중요한 범주라 할 수 있으므로 이 범주에서 높은 점수를 받은 문항은 나머지 항목을 만족시키지 못한다해도 과학수행평가문항으로써 장점을 가진 문항이라고 판단될 수 있다. 그러나 사실상 본 준거표는 중요한 범주에 대한 가중치를 이미 둔 상태로 제작되었으므로 과학수행평가문항으로서 좋은 문항은 준거표의 총점에서도 높은 점수를 받게된다.

이 준거표는 과학 수행평가문항에 대한 이상을 제시하여주므로, 모든 과학 수행평가문항이 본 준거표에 제시된 항목들을 완전히 만족시킬 수는 없을 것이다. 그러나 과학 수업 현장에서 수행평가문항을 선정하거나 제작하여 사용할 때 이를 기준으로 한다면 과학과에서 수행평가의 본질을 구현하고 그 목표를 달성하는데 기여하게 될 것이며, 궁극적으로 과학교육의 질을 향상시키는데 기여할 것으로 기대한다.

Ⅳ. 결론 및 제언

본 연구에서는 과학과 수행평가의 질을 높이기 위한 목적으로 수행평가에 대한 이론적 논의를 통해 과학과 수행평가문항의 선정·제작을 위한 평가 준거를 제작하였다. 본 준거는 충실도, 만족도, 내용타당도, 공정성 및 적합성, 신뢰도와 객관도, 사용상의 편이성

Table 2. Correspondence among the three evaluators

Publication*	Grade-Semester	Content/Task	K's W	Probability
I	1-2	Survival in winter Let's help doner!	0.27	.691
K	5-1	Solution & Solvent /Solution & Solvent	0.024	.765
G	6-1	Molecule /Molecule	0.009	.980

* I : a performance assessment task by institute for science education in Inchon National University of Education
K : a teacher's material by Korea Teachers' Federation
G : a teacher's material by office of education in Gyunggido

및 실용성의 6개의 대 범주로 구성되며, 각 대범주는 몇 개의 하위 항목으로 구성된다.

이 평가 기준은 과학교육에 있어서 다음과 같은 함의를 갖는다 .

첫째, 과학교육에서 수행 평가에 대한 이해를 도울 수 있다.

둘째, 현재까지 보급되었거나 실제 현장에서 사용하고 있는 과학과 수행평가문항들의 질을 조사해 봄으로써 현재 사용되고 있는 과학과 수행평가문항들의 문제점을 파악할 수 있다.

셋째, 현장 교사들이 양질의 과학과 수행평가문항을 선별하여 사용하는데 지침서의 역할을 할 수 있다.

넷째, 과학과 수행평가를 제작하는데 체크리스트의 역할을 함으로써, 과학과 교육목표에 도달하는데 기여할 수 있다.

적 요

본 연구에서는 과학교육에서 수행평가의 올바른 사용과 정착을 위하여 과학 수행평가문항의 선정과 제작을 위한 평가준거를 개발하였다. 이를 위하여 과학과 평가목표 및 수행평가에 대한 이론적 논의를 통해 평가준거의 대범주로 평가 목표의 충실도(Fidelity), 수행평가의 특징과 구성 요소에 대한 만족도(Satisfaction), 문항의 내용 타당도 (Content Validity), 학습자에 대한 내용의 적합도 및 공정성(Suitability & Fairness), 문항의 신뢰도와 객관도(Reliability & Objectivity), 사용상의 편의성 및 실용성(Usability)의 6가지 범주를 개발하였고, 각 대범주별 하위항목을 선정하여, 총 25개의 하위 항목을 가진 평가준거표를 개발하였다. 그리고 개발된 평가준거표의 타당도는 과학교육 연구자 5명의 안면타당도로 보았고, 신뢰도는 3가지 종류의 과학 수행평가문항을 과학교육 전문가 2명과 초등예비교사 6명이 평가하고, 그 결과를 Kendall계수를 통한 평가자간 합치도로 검증하였다. 그 결과 합치도는 통계적으로 유의한 수준으로 나타났다. 이상의 과정을 통하여 최종적인 평가준거표를 완성하여 제시하였다.

참 고 문 헌

강인애(1997), 왜 구성주의인가?, 문음사.

김영순(1998), 초등학교 자연과 수행평가에 대한 문화기술적 연구, 한국교원대학교 석사학위논문.

김영채 역(1993), 학습심리학, 박영사.

김은진(2000). 과학 교과 수행평가틀의 개발. 서울대학교 박사학위 논문.

남명호(1995). 수행평가의 타당성 연구. 고려대학교 박사학위논문.

임영득, 조혜경, 한안진, 박현주, 송민영, 김은진, 홍석인, 강호감, 노석구(1999), 초등학생의 자연과 수행평가실태조사 및 초등학교 자연과 수행평가도구의 개발, 초등과학교육, 18(1), 41-51.

임영득, 조혜경, 한안진, 박현주, 송민영, 김은진, 홍석인, 강호감, 노석구(2001). 초등학생의 자연과 수행평가실태조사 및 초등학교 자연과 수행평가도구의 개발 II. 한국과학교육학회지, 21(2), 459-472.

전성연, 최병역 역(1998), 학습 동기, 학지사.

정종진(1999). 학교학습의 극대화를 위한 교육평가의 이해. 양서원.

황정규(1998). 학교학습과 교육평가. 교육과학사.

Baron, J.(1991). Performance assessment: Blurring the edge of assessment, curriculum and instruction. In G. Kulm & S. Malcom(Eds.), *Science assessment in the service of reform*. American Association for the Advancement of Science: Washington, D. C., 247-266.

Bransford, J. D. & Stein, B. S.(1984). *The ideal problem solver: A guide for improving thinking, learning, and creativity*, W. H. Freeman: NY.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C.(1983). Learning, remem-bering, and understanding, In J.H. Flavell and M. Markman (Eds.), *Cognitive Development*, Wiley: New York, 3, 77-166.

Brown, J. H., & Shavelson, R. J.(1996). *Assessing hands-on science: A teachers' guide to performance assessment*. Sage Publication Company: CA.

Glaser, R.(1984). Educating and thinking: The role of knowledge. *American Psychologist*, 39, 93-104.

Glaser, R. & Pellegrino, J. W.(1987). Aptitudes for learning and cognitive processes. In F. Weinert & R. Kluwe(Eds.), *Metacognition, motivation and understanding*. Lawrence Erlbaum Associates: Hilsdale, NJ, 267-288.

Herman, J. L., Aschbacher, P. R., & Winters, L.(1992). *A practical guide to alternative assessment, Selecting Assessment Tasks*. U.S.A : University of California.

Hart, D.(1994). *Authentic Assessment: A Handbook for Educators*. Addison-Wesley Publishing Company.

Kim, E. J., Park, H. J., Kang, H. K., & Noh, S. G.(2000). Developing a framework for science performance assessment. Paper presented at 2000 NARST conference in New Orleans, LA.

Larkin, J. H.(1989). What knowledge transfers? In L.B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hilsdale. NJ: Eribaum, 283-305.

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

Resnick, L. B. & Klopfer, L. E.(1989). Toward the thinking curriculum: An overview. In Resnick, L. B. & Klopfer, L. E.(Eds.), *Toward the thinking curriculum: Current cognitive research. 1989 yearbook of Association for Supervision and Curriculum Development*, ASCD: Alexandria, VA.

Tarmir, P(1998). Assessment & Evaluation in Science Education: Opportunities to learn & outcomes. In B. J. Fraser & K.G. Tobin(ed), *International Handbook of Science Education* (Part 2), Kluwer Academic Publishers.

Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A.,Schultz, S. E., Wiley, E. W., Brown, J. H.,(1997). On the development and scoring of classification and observation science performance assessments. Paper presented at the annual meeting of the AERA, Chicago, IL, April 24-28, 1997. ED 411 314.

부록 1

〈과학과 수행평가 문항 평가준거〉

문항제목:

※ 점수부여방법: 대상으로 하는 문항이 평가 준거의 내용을 적절히 갖추고 있다고 판단되면: 2점.
평가 준거의 내용을 갖추고는 있으나 충분치 못하다고 판단될 경우: 1점
평가 준거의 내용을 갖추고 있지 않은 경우 : 0점

평가자의 총평란은 이후에 또 다른 사용자에게 도움이 될 수 있는 의견이 있다면 적어주십시오.

대 범 주 / 준 거 항 목 번 호		Score	
		등급	총점
(1) 과학 교과 목표에 대한 충실도 (Fidelity)	(1)-① 과학적 탐구 사고력 요소가 평가되고 있는가? (1)-② 과학적 탐구를 위한 수공 기능요소가 평가되고 있는가? (1)-③ 과학적 태도 영역이 평가되고 있는가? (1)-④ 과학적 지식의 적용력이 평가되고 있는가? (1)-⑤ 창의적 사고과정이 평가되고 있는가? (1)-⑥ 타인과의 의사소통력이 평가되고 있는가? (1)-⑦ 반성적 사고과정이 평가되고 있는가? (1)-⑧ ICT를 활용한 정보수집과정이 포함되어있는가?		
(2) 수행평가의 특징과 구성 요소에 대한 만족도	(2)-① 수행과제의 내용에 있어서 학습자의 동기유발을 위한 요소를 포함하고 있는가? (2)-② 반응양식 중 학습자가 학습에 대한 피드백(feedback)을 받을 수 있는 과정이 포함되어있는가? (2)-③ 반응양식 중 학습자의 학습과정을 모니터링 할 수 있는 절차가 포함되어있는가? (2)-④ 자기평가, 교사평가, 학부모 평가 등 다양한 평가자에 의한 평가를 절차가 있는가? (2)-⑤ 학습자 자기주도적인 학습이 이루어 질 수 있도록 구성되어있는가? (2)-⑥ 수행평가 문항이 수행과제, 반응양식, 채점체계를 모두 갖추고 있는가? (2)-⑦ 학생의 반응에 대한 평가 체계에 융통성이 있는가?		
(3)문항의 내용타당도 (Content Validity)	(3)-① 평가목표에 제시된 과학개념과 문항의 내용이 적합한가? (3)-② 문항에 포함되어있는 과학지식이 정확한가? (3)-③ 문항의 표현 중 모호한 부분이 없고 명확한가?		
(4)학습자에 대한 내용의 적합도 및 공정성(Suitability)	(4)-① 문항의 내용이 학습자의 인지적 수준에 적합한가? (4)-② 문항의 내용이 학습자의 문화적 환경과 지역 환경 및 성별에 있어서 공정한가?		
(5)사용상의 편의성 및 실용성 (Usability)	(5)-① 문항의 사용을 위한 사전 사후 조치가 용이한가? (5)-② 문항의 현장 투입 시 해결을 위한 시간은 적절한가? (5)-③ 문항 사용의 비용면에서 경제적인가?		
(6)문항의 신뢰도와 객관도 (Reliability & Objectivity):	(6)-① 상세화된 채점 체계를 가지고 있는가? (6)-② 평가자가 채점기준에 대해 충분히 이해할 수 있도록 제시되어 있는가?		
총 점			
평가자의 총평			