과학 실험 평가 도구 개발을 통한 탐구 능력 평가의 타당화에 관한 연구

우종옥 · 이항로 · 김승훈*
(한국교원대학교) · (경기 심원고등학교)*

(1996년 11월 14일 받음)

I. 서 론

과학 탐구 능력을 학생들에게 습득시키기 위한 가장 효과적인 방법들 중의 하나로 실험의 역할을 크게 강조하고 있다(Leonard, 1983; Shymansk, Kyle and Alport, 1983; Tobin, 1982). 이의 구체적인 실현을 위해 탐구 중심 교육 방법의 현장 적용을 위한 실험 연구가 진행되어 왔으며, 탐구 학습 결과를 평가하기 위한 탐구 능력 평가 도구의 개발 연구도 함께 추진되어 왔다.

이러한 과학 교육 개혁 운동의 결과로 PSSC, CHEM, BSCS, ESCP, SAPA, SCIS등 수많은 혁신적 교육과정이 개발되었고, 이를 평가하기 위한 연구로서 영국의 APU와 TAPS, 미국의 NAEP, Hur(1984)가 개발한 과학 탐구 평가틀(SIEI) 등이 만들어졌다. 이들 과학 탐구 능력 측정 도구는 주로 지필 평가 형태로 개발되었지만, APU의 실험 기능 평가나 SISS(The Second IEA Science Study)와 같이 표준화된 실험 기능 탐구 능력 평가 도구의 개발은 최근에 이루어지고 있다(권재술 외, 1994).

제 6차 교육과정에서는 탐구 중심의 과학교육과 평가가 과거 어느 때보다도 중요시되고 있다. 과학 탐구 능력 신장이 과학교육의 중요한 목표로 설정되어 있으며, 과학 교과의 내용 체계에서 탐구 영역을 분리 독립시켜 이에 대한 교육을 한층 강조하고 있다. 또한 교육 과정상에 공통 과학이 탐구 중심의 필수 과목으로 설정되어 있으며(교육부, 1992), 대학수학능력 시험도 과학 탐구 능력의 평가에 일차적인 목표를 두고 있다(국립교육평가원, 1992).

그러나 우리 나라 과학교육에서 탐구 학습과 평가의 중요성을 강조하여 왔음에도 불구하고 탐구 학습이 학교 현장에서 제대로 이루어지지 않는 중요한 원인 중의 하나는 탐구적 실험 지도 및 이에 알맞은 평가 방법이 제대로 개발되어 있지 않다는 것(허 명, 1991)과 탐구 능력에 관한 평가의 준거, 평가 방법, 문항 개발의 기법에 대한 기초 연구가 많이 이루어지지 못했고, 평가 자료가 풍부하게 개발되지 못한 것 등을 지적하고 있다(권치순, 1988).

특히 실험 평가는 평가자 주관에 의해서 평가되기 때문에 평가자의 신뢰도와 평가의 타당도에 많은 문제점이 제기되고 있다(권재술 외, 1994). 평가의 타당도는 무엇보다도 학문의 본질을 제대로 반영하고 있는지에 달려 있다고 볼 수 있고 실험에 대한 평가가 제대로 이루어져야만 타당도 높은 평가가 이루어졌다고 말할 수 있다. 이와 같이 실험 평가의 타당도를 높이기 위해서는 정확한 평가의 관점 및 척도 개발을 필수 조건으로 하고 있다.

실험 탐구 능력의 평가는 실제 실험을 통하여 이루어지는 것이 바람직하지만, 이를 수행하는데 있어서의 현실적인 어려움 때문에 대부분의 평가가 지필 평가 방법에 의해서 이루어지고 있는 실정이다.

그러나 이러한 평가는 지필 평가에 의한 탐구 능력 측정 결과가 실험 과정에서의 탐구 능력 측정 결과와 같을 것이라는 전제하에서 이루어지는 것인데, 최근의 연구 결과에 의하면 이 두 조건에서 학생들의 성취도 사이의 상관 관계가 탐구 능력 요소에 따라 차이가 있는 것으로 밝혀졌다(김동찬, 1991). 이와 같이 지필 평가 결과와 실험 평가 결과가 다른 것은 지필 평가 문항은 문제 상황이 단순하고 구체적으로 제시되지만 실제 실험은 모든 것을 종합한 전체적인 상황이기 때문인 것으로 해석할 수 있다(최병순 외, 1994). 또한 지필 평가에서는 탐구 능력 중 상위의 심체적 영역 평가는 할 수

없으므로 모든 탐구 능력 요소를 종합적으로 측정할 수 있는 실험 평가 도구의 개발이 필요하다. 따라서 정확한 탐구 능력의 측정을 위해서는 탐구 능력의 모든 요소를 포함하고 종합적인 상황이 제시되는 실험 과정에서의 평가가 더 바람직하다고 볼 수 있다. 또한 이제까지 개발된 탐구 능력 측정 도구는 대부분이 지필 평가 유형이며, 때로는 실험 평가 도구라는 이름으로 개발되기도 하였으나 이것들은 실제 실험을 수행하면서 학생들이 활용하는 탐구 능력을 측정하고자 하는 평가 도구가 아니라 모의 실험 상황에서 학생들의 탐구 능력을 측정하기 위한 지필 평가가 주종을 이루어 왔다(최병순, 1994).

외국의 경우는 실제 실험 과정에서 탐구 능력을 측정하고 평가하기 위한 실험 프로그램과 평가의 관점 및 척도 개발 연구가 비교적 많은 편이다. 그러나 국내의 경우는 지필 평가에 의해서 탐구 능력을 측정하기 위한 도구 개발에 관한 연구는 다수 있었던 반면 직접 실험을 하는 과정 속에서의 탐구 능력을 측정하기 위한 평가 도구의 개발은 미흡한 실정이다. 6차 과학과 교육과정의 공통 과학 교과에서 강조하는 일차적인 목표가 직접 실험 수행(hand on)을 통한 탐구 능력 신장에 있다고 본다면 타당도와 신뢰도가 높은 실험 탐구 능력 평가 도구의 개발에 관한 연구는 더더욱 절실히 필요하다고 볼 수 있다.

따라서 본 연구에서는 일선 학교 현장에서 적용 가능하고, 타당도와 신뢰도가 높으며 실제 실험을 통하여 탐구 능력을 측정할 수 있도록 고등학교 지구과학 교과의 실험 주제를 중심으로 한 평가 도구를 개발함으로써, 실험을 통한 탐구 능력 평가 방법과 이를 통한 과학 탐구 능력 평가의 타당화 방안에 관한 단서를 찾고자 하였다.

Ⅱ. 연구 방법 및 절차

본 연구는 상기한 연구의 필요성에 의해 과학 탐구 능력 평가에 관한 국내외 선행 연구를 토대로 실험을 통한 탐구 능력 평가 도구 개발과 현장 적용을 실시하였다. 또한 지필 평가를 실시한 결과와 상호 비교, 분석하여 과학 탐구 능력 평가의 타당화 방안을 제시하는 연구로서 구체적인 연구 절차와 방법을 제시하면 다음과 같다.

1. 실험 탐구 능력 평가틀의 선정

1) 과학 탐구 과정 모형

대학수학능력 시험에서 요구하는 과학 탐구 능력 신장 및 평가를 위해 우종옥 등(1992)이 번안 제시한 5 단계 과학 탐구 과정 모형을 본 연구의 과학 탐구 과정 모형으로 선정하였다(표 1).

〈표 1〉 탐구 단계 및 세부 탐구 과정 요소

탐 구 단 계	하 위 탐 구 과 정 요 소
1. 문제 인식 및 가설 설정	1-1. 문제 인식 1-2. 가설 설정
2. 탐구의 설계	2-1. 실험 설계 2-2. 변인 통제 2-3. 실험 기구 선정
3. 탐구의 수행	3-1. 관찰, 측정, 실험 3-2. 실험 기구 조작 3-3. 자료 수집
4. 자료의 해석	4-1. 추리 4-2. 예상 4-3. 상관 관계/인과 관계 4-4. 자료의 변환
5. 결론 도출 및 평가	5-1. 결론 도출 5-2. 적용 및 평가

〈표 1〉에 나타낸 바와 같이 본 연구에서 선정한 과학 탐구 과정 모형은 탐구 단계가 5단계로 이루어져 있으며, 이 모형을 근거로 14가지의 하위 탐구 과정 요소를 선정하였다. 선정한 탐구 과정 요소의 평가 목표를 상세화하여(우종옥 등, 1992) 평가 문항 개발의 준거로 활용하였다.

2) 평가의 관점 및 척도

논술형 문항이나 실험 실기 능력을 측정하는데 있어서 평가자의 주관적인 판단이 배제되고 객관성을 유지하기 위해서는 평가의 준거와 척도가 상세히 제시되어야만 한다.

본 평가 척도는 평가자의 지나친 주관이 개입되거나 하나하나의 척도에 극단적인 평정값이 주어지는 사례를 방지하기 위하여 평가 척도에 흔히 사용되는 5단계의 평정을 지양하고 있으나, 본 연구의 제한성 때문에 3 단계의 평정 척도를 설정하였다(표 2). 그러나 문항의 특성에 따라 3단계로 평가하기 어려운 경우에는 여러 단계로 세분화하여 평가하도록 하였

〈표 2〉 탐구 실험 평정 척도

평 가 관 점	평가척도
문제의 의도를 정확하게 파악하고 완벽하게 설명을 하였거나, 또는 독특한 아이디어나 실험 방법을 도입하여 기대 수준 이상의 활동을 한 경우	상
설명이 다소 부족하거나, 또는 대체로 수행할 수 있을 것으로 기대되는 수준의 활동을 한 경우	중
문제의 의도대로 설명하지 못했거나 틀리게 진술, 또는 잘못된 실험 방법을 도입하거나 기대 이하의 활동을 한 경우	하

다.

문항의 점수는 보통 한 문항당 (상)에는 4점을, (중)에는 2점을, (하)에는 0점을 주되 문항의 난이도와 비중에 따라 차등 배점할 수 있도록 하였고, 한 실험 주제당 20점 만점으로 하였다.

2. 실험 탐구 능력 평가 도구 개발의 준거 및 개발

1) 실험 주제의 선정

제 6차 고등학교 지구과학 교과서(Ⅰ, Ⅱ)에 실린 필수 실험 내용 체계 중에서 4주제를 선정하였고, 또한 교과서에서 다루지는 않지만 지구과학적 소재로 실험 탐구적 접근이 가능한 실험 주제 1개를 선정하여 문항을 개발하였다(표 3).

〈표 3〉에 제시한 바와 같이 '토양의 공극율과 투수율 측정', '태양의 방위각과 고도의 측정', '수온 약층', '단열 팽창과 응결', '해양 지각의 이동 속도'의 5주제이며, 각 주제당 최소한 5단계 중 3단의 탐구 단계를 측정할 수 있도록 하였다.

2) 실험 평가 도구의 형태

(1) 실험 평가 문제지(학생용)

① 실험 제목 : 실험을 통해 탐구해야 할 내용의 특징을 나타내는 제목을 붙인다.
② 준비물 : 실험시 필요한 기구 및 재료를 기록한다.
③ 문제 : 실험 과정 및 실험 결과시 나타날 수 있는 내용을 실험을 실시하면서 답할 수 있도록 문제를 구성한다.

(2) 실험 평가 지침서(교사용)

① 실험 제목 : 실험을 통해 탐구해야 할 내용의 특징을 나타내는 제목을 붙인다.
② 평가 목표 : 평가 목표는 가급적 행동 용어로 상세히 기술한다.
③ 준비물 : 실험 준비시 필요한 기구 및 재료를 기록한다.
④ 지도상 유의점 : 실험 지도상의 유의점을 상세하게 기록한다.
⑤ 평가의 준거 :

가. 교사용 실험 지도서에 학생용 평가 문항의 내용을 모두 제시하여 지도 교사가 편리하게 활용할 수 있도록 한다.
나. 각 문항별로 정답, 탐구 요소, 평가 목표, 평가 관점, 평가 준거 및 평가 척도 등을 제시한다.
다. 평가 준거는 가능한 한 단순화하고 상세화하여 채점시 논란이 없도록 한다.

⑥ 배점 : 보통 한 문항당 4점씩 배점하였으며, 문항의 특성에 따라 6점, 3점으로 차등 배점한 것도 있다. 그러나 한 실험 주제당 모두 20점 만점으로 하였다.

(3) 평가 문항의 내용 타당도 검증

본 연구에서는 개발한 실험 탐구능력 평가 문항이 선정된 탐구 능력 요소와 평가 목표가 일치하는지 여부를 과학교육 전문가와 지구과학 교육 전공 박사 과정 대학원생들에게 의뢰하여 일치하는 정도를 퍼센트(%)로 나타내었다.

3. 실험 탐구 능력 평가 문항의 현장 검증

1) 연구 대상의 표집 및 현장 투입

1차 현장 검증은 중소도시 인문계 2학년인 5명을 대상으로

〈표 3〉 실험 주제별 문항수

탐구요소 / 실험주제	Ⅰ		Ⅱ			Ⅲ			Ⅳ				Ⅴ		계
	문제 인식	가설 설정	실험 설계	변인 통제	실험 기구 선정	관찰, 측정, 실험	실험 기구 조작	자료 수집	추리	예상	상관 관계	자료 변환	결론 도출	적용 및 평가	
1. 토양의 공극률과 특수율 측정		1				1						1	1	1	5
2. 태양의 방위각과 고도 측정						1				1		1	1	1	5
3. 수온 약층			1			1						1	1	1	5
4. 단열 팽창과 응결	1					1				1		2	1		6
5. 해양 지각의 이동 속도		1							1		1	3	1		7
계	1	2	1			4			1	2	1	8	5	3	28

(주. Ⅰ : 문제 인식 및 가설 설정, Ⅱ : 탐구의 설계, Ⅲ : 탐구의 수행, Ⅳ : 자료의 해석, Ⅴ : 결론 도축 및 평가)

하였으며, 2차 현장 검증은 중소도시 인문계 2학년인 25명을 대상으로 하였다.

본 연구 대상에게 1차로 개발된 실험 탐구 능력 평가 도구를 5일 동안 정과 수업이 종료된 후에 1차 현장 검증을 실시하였고, 2차 현장 검증은 매주 금요일 특별활동 시간을 이용하여 5주간 실시하였다.

1차 현장 검증은 개발하고자 하는 평가 문항의 문제점 및 전체 문항을 수행하는데 소요되는 시간이 적절한지를 점검하기 위해 실시하였고, 2차 현장 검증은 1차 현장 검증에 나타난 문제점을 수정·보완하여 평가 문항의 신뢰도 및 타당도를 높이기 위해 실시하였다.

2) TESIS의 투입

본 실험을 실시한 집단에 실험을 통한 탐구 능력과 지필 검사를 통한 탐구 능력간의 상관 관계를 알아보기 위해 이항로(1991)가 개발한 TESIS를 투입하였다. 평가는 실험을 실시한 다음 주에 1시간 동안 본 연구자의 감독 하에 실시하였다.

3) 평가 도구의 신뢰도 검증

본 연구는 실험 주제별 5~8 문항밖에 되지 않아 적은 수의 문항으로 신뢰도를 추정하면 신뢰도 계수가 너무 작고 측정의 오차가 너무 커서 신뢰도에 문제가 생기기 때문에 5개의 실험 주제를 하나로 묶어서 신뢰도를 추정하였다

4) 평가자간 신뢰도 및 평가자간 일치도 검증

(1) 평가자간 신뢰도

논술형 고사나 예체능의 심체적 영역의 수행 결과에 대한 채점 혹은 평가 결과는 객관도에 의하여 분석된다(성태제, 1995). 객관도(objectivity)란 평가자의 주관적인 편견을 얼마나 배재하였느냐의 문제이다. 그러므로 객관도란 한 평가자가 다른 평가자와 얼마나 유사하게 평가하였느냐의 문제와 한 평정자가 많은 측정 대상에 대하여 계속적으로 일관성 있게 측정하였느냐의 문제로 구분할 수 있는데, 전자를 평가자간 신뢰도(inter-rater reliability) 혹은 채점자간 신뢰도라 하며, 후자를 평가자내 신뢰도(intra-rater reliability) 혹은 채점자내 신뢰도라 한다. 본 연구에서는 연구의 제한성 때문에 평가자간 신뢰도만을 구하였다.

평가자간 신뢰도를 추정하는 방법에는 상관 계수법, 일치도 통계와 Kappa 계수, 일반화 가능도 이론 등이 있지만 본 연구에서는 평정 점수가 연속 변수인 점수로 부여될 때 두 평정자가 동일한 집단의 피험자에게 얼마나 유사하게 점수를 부여하였느냐를 분석하는 단순 적률 상관 계수법을 이용하였다.

상관 계수가 높으면 두 채점자는 동일한 채점 기준에 의하여 채점한 것으로 분석되며, 상관 계수가 낮으면 채점자가 각기 다른 채점 기준에 의하여 채점하였음을 나타낸다.

(2) 평가자간 일치도

여러 명의 평가자가에 동일한 대상에 대한 평가 결과가 얼마나 상호 일치하는가를 나타내는 지수로 Kendall의 일치도 계수(Kendall's coefficient of concordance)가 있는데 이를 흔히 W계수로 나타내고 있다(임인재, 1991).

모든 평가자가 완전히 일치하게 평가하는 경우에는 각 피험자에게 주어진 점수의 합 사이에 최대의 변산을 갖는 반면, 평가자간에 불일치하면 할수록 피험자에게 주어진 점수의 합의 분산은 적어져서 완전히 불일치하는 경우에는 변산은 전혀 없게 된다. 따라서 모든 평가자가 일치함으로써 생기는 최대의 변산에 대한 현재 얻어진 합의 변산의 비로서 평가자간의 일치의 정도를 나타내는 지표로 Kendall의 W계수이다.

본 연구에서도 평가자간의 일치도를 구하기 위하여 Kendall의 W계수를 이용하였다. 평가자간에 완전히 일치되었을 때는 W=1이 되고 최대의 불일치의 경우는 W⟩0이 된다. 즉, W는 $0 \leq W \leq 1$ 사이의 값을 가지며 음수는 취할 수 없다.

5) 평가 문항의 난이도 및 변별도 분석

논문형과 같이 부분 점수가 있을 때에는 모든 반응들이 정답과 오답으로 양분되기 어렵고, 오히려 부분 점수에 의하여 어느 정도로 정답인가 하는 판단을 할 수 있다. 따라서 본 연구에서는 각 문항마다 부분 점수가 있을 경우에 사용할 수 있는 식(김성훈, 1992)을 이용하여 난이도를 구하였다. 본 연구에서는 전체 학생 수에서 상위 집단 27%, 하위 집단 27%를 나누어서 변별도를 구하는 공식을 이용하였으며, 논문형 문항과 같이 부분 점수를 주는 문항의 변별도를 구하였다(김성훈, 1992).

Ⅲ. 연구 결과 및 논의

1. 내용 타당도

1차로 선정한 5개의 실험 주제에서 한 실험당 5~8개의 실험 탐구 능력 평가 문항을 개발하여 총 28개 문항에 대한 내용 타당도를 과학교육 전문가(과학교육 전공 교수 2명, 지구

과학 교육 박사 과정 대학원생 4명)에게 내용 타당도를 점검하도록 의뢰하였다. 타당도 점검을 의뢰한 결과 대부분 일치하였고, 한 실험 주제에 대한 문항은 전체적으로 문제가 지적되어 다른 주제로 대체하였다.

1차 점검때 지적된 문항과 대체한 문항을 1차 점검때 의뢰한 전문가에게 다시 타당도를 검증 받은 결과, 총 168개의 응답(response : 6명의 평정자×28문항)중 82.7%가 평가 목표와 일치했다.

2. 신뢰도

각 주제별 문항의 수가 5~8 개 정도 밖에 되지 않아 적은 수의 문항으로 신뢰도를 추정하면 측정의 오차가 너무 크기 때문에 5개의 실험 주제를 하나로 묶어서 추정한 신뢰도 계수 Cronbach α는 0.86이다. 28 문항으로 구성된 본 실험 평가 문항의 총점이 70.64이고 변산은 174.24이다.

문항을 표준화시킨 후의 신뢰도(standardized item alpha)는 모든 문항 점수를 각 문항 점수의 표준 점수로 치환하여 얻은 Cronbach α값을 나타내나. 문항 점수들의 배점이 다를 경우, 즉 1번 문항은 2점 만점이고 2번 문항은 6점 만점으로 서로 평가 척도가 다를 때 신뢰도 지수를 추정하기 위하여 사용된다. 그러나 모든 문항 점수가 동일 척도일 때는 고려하지 않는다. 따라서 본 실험 문항의 신뢰도는 .85로 비교적 높은 값을 나타내고 있다.

3. 평가자간 신뢰도와 일치도

1) 평가자간 신뢰도

평가자간 신뢰도를 구하기 위하여 실험 집단에 투입한 5개의 실험 평가 문항을 5명의 평가자에게 채점을 의뢰하여 평가 준거에 따라 채점하도록 하였다.

평가자 간의 신뢰도를 구하기 위해 산출한 평가자간 상관계수가 .88이상으로 각 평가자간에 매우 높은 상관을 보이고 있다. 이는 평가자의 주관이 배제된 객관적인 평가 결과라고 볼 수 있으며 또한 평가의 준거가 명확하게 제시되어 채점상 문제가 극소화된 것이라고 볼 수 있다.

2) 평가자간 일치도

채점을 의뢰한 5명의 평가자들이 피험자들에 대해 평가 준거에 따라 공정하게 채점을 하였을 때, 이들 평가자들 간에 어느 정도 일치하고 있는지의 여부를 검증하기 위하여 Kendall의 일치도 계수(W)를 구하였다.

평가자간의 일치도가 어느 정도 이상 되어야 한다는 기준은 없지만 여러 전문가들의 공통적인 견해를 종합해 보면 .70 이상은 되어야 그 문항에 대한 평가 준거가 명확히 세워 졌다고 볼 수 있다. 본 연구에서 개발한 실험 평가 문항에 대한 평가자간 일치도는 〈표 4〉와 같다.

〈표 4〉 Kendall의 일치도 계수에 의한 평가자간 일치도

문항 번호 \ 실험	실험 1	실험 2	실험 3	실험 4	실험 5
1	.90	.45	.96	.79	.78
2	.93	.58	.85	.84	.95
3	.99	.82	.75	.83	.72
4	.85	.88	.85	.74	.92
5	1.00	.97	.96	.75	.85
6				.92	.91
7					.85
총계	.93	.74	.87	.81	.85

평가자간 일치도 계수(W)는 .74~.93으로 상당히 높은 일치도를 보이고 있다. 이는 평가자들이 문항을 잘 이해하고 제시된 준거대로 평가했다고 볼 수 있다. 다만, 실험 2의 1, 2번 문항의 평가자간 일치도 계수에 문제가 있어 이를 수정 · 보완하여 제시하였다.

3) 평가자간 신뢰도와 일치도와의 관계

동일 집단을 여러 명의 평가자가 나누어서 평가할 때는 일치도가 높아야 한다. 평가자간 신뢰도만 높으면 한 집단의 평가 점수에 따른 순위가 다른 결과를 가져올 수 있다. 결과 중심 평가에서 평가자에 따른 불안정성을 제거하기 위하여 동일 집단을 한 평가자가 모두 평가할 때는 신뢰도만 높은 문항을 사용해도 되지만 많은 수의 한 집단을 여러 명이 나누어서 평가할 때는 신뢰도와 일치도가 높아야 한다. 왜냐 하면 신뢰도가 높은 문항이 반드시 일치도가 높은 것이 아니기 때문이다.

실제로 본 연구에서 보면 평가자간 신뢰도는 각 실험 주제별로 평가자간의 상관 계수가 .80이상으로 상당히 높게 나왔다. 그러나, 평가자간 일치도를 보면 전술한 바와 같이 실험 2의 1, 2번 문항은 상당히 낮은 일치도를 보이고 있다. 따라서 평가자의 채점이 얼마나 객관적으로 평가되었는지를 알아보기 위해서는 평가자간 신뢰도와 일치도를 모두 검증하여야만 한다.

4. 문항의 난이도 및 변별도 분석

본 연구의 현장 검증 결과에서 나타난 문항의 난이도 및 변별도는 표집의 크기가 작고 표집 대상이 특수 선발 집단인 관계로 분석된 자료의 결과 해석에 무리가 있을 수 있기 때문에 문항의 개발 방향과 수정, 보완에 대한 지침 자료로만 활용하였다. 실험 주제별, 탐구 과정 요소별 난이도 및 변별도는 다음과 같다.

1) 실험 주제별 난이도 및 변별도

현장 검증 결과 개발한 실험 탐구 능력 평가 문항의 평균 난이도는 69.0이고, 주제별로 보면 실험 3이 60.4로 가장 낮고 실험 2가 77.9로 가장 높게 나타났다.

변별도는 실험 2가 .17로 가장 낮고 실험 5가 .37로 가장 높게 나타났고, 변별도 평균 지수는 .30이다.

2) 탐구 과정 요소별 난이도 및 변별도 지수

탐구 요소별 난이도, 변별도 지수를 비교한 결과를 나타내면 〈표 5〉와 같다. 〈표 5〉에 나타난 바와 같이 변별도 모두 동일한 탐구 과정 요소에 관한 평가 문항일지라도 문항에 따라 많은 차이가 있음을 알 수 있다. 평균값을 보면 탐구요소 중 '예상' 문항은 난이도가 81.3으로 쉬운 것으로 나타났고, 변별도가 .09로 변별력이 아주 낮은 문항으로 나타났다. '자료 변환' 이나 '적용' 문항은 비교적 어렵고 변별력도 높은 것으로 나타났다.

〈표 5〉 탐구 요소별 난이도 및 변별도

탐구 요소	문항 번호	난이도	변별도
가설 설정	A1, D1, E1	75.6	.25
실험 설계	C1	57.1	.29
측 정	B1, A2, C2, D2	76.8	.20
추 리	D3, D4, E2	65.5	.36
예 상	B2, E3	81.3	.09
자료 변환	A3, B3, C3, E4, E5, E6	62.3	.40
결론 도출	A4, B4, C4, D5, E7	71.1	.24
적 용	A5, B5, C5, D6	63.9	.44

5. TESIS와의 상관관계

본 연구에서 개발한 실험 탐구 능력 성취도와 지필 평가에서 나타난 탐구 능력 성취도간의 상관 관계를 알아보기 위하여 실험 집단에 5개의 실험을 모두 실시한 후에 이항로(1991)가 고등학교 지구과학 소재를 중심으로 개발한 탐구 능력 평가 도구를 투입하여 구한 상관 관계는 〈표 6〉과 같다.

본 실험 주제를 모두 합한 총점(실험 평가)과 TESIS(지필 평가)간의 상관 계수는 $r = .45$ 으로서 탐구적 실험을 통한 탐구 능력과 지필 평가를 통한 탐구 능력과는 상관도가 낮은 것으로 나타났다. 결정 계수 γ^2은 .174, 즉 17.4%이다.

따라서 실험 탐구 능력 성취도는 지필 평가에 의한 탐구 능력 성취도의 17.4%만을 예언 또는 설명할 수 있는 것으로 해석할 수 있다.

이와 같은 낮은 상관도는 김동찬(1991)의 '실험 과정에서의 탐구 능력과 지필 평가에서 나타나는 탐구 능력과의 상관 관계'에서 상관이 있기는 하지만 낮은 상관을 나타낸다는 연구 결과와 거의 일치하는 결과를 보여 주고 있다. 따라서 보다 타당한 탐구 능력 평가를 위해서는 실험을 통한 과정 평가가 동시에 이루어져야 함을 알 수 있다.

Ⅳ. 결론 및 제언

과학 교육에서 실험을 통한 탐구 능력의 평가는 "어떻게 객관성 있게 평가하는가?"의 문제가 큰 과제라고 볼 수 있다. 본 연구에서는 지구과학 교과를 중심으로 일반계 고등학

〈표 6〉 실험평가와 지필평가를 통한 탐구능력과의 상관관계(N=25)

상관관계	실험 1	실험 2	실험 3	실험 4	실험 5	총점	지필평가
실험 1	1.00						
실험 2	.64**	1.00					
실험 3	.65**	.57*	1.00				
실험 4	.47*	.45	.57*	1.00			
실험 5	.39	.41	.45	.62**	1.00		
총 점	.82**	.77**	.83**	.77**	.74**	1.00	
지필평가	.25	.23	.60**	.28	.38	.45	1.00

* $p<0.01$ ** $p<0.001$

교에서 실험을 통하여 탐구 능력을 측정하기 위한 평가 도구를 개발하였다. 개발한 실험 탐구 능력 평가 도구를 현장 적용과 검증을 통하여 신뢰도와 타당도를 개선한 결과와 이를 통한 탐구 능력 평가의 타당화 방안에 대한 결과는 다음과 같다.

실험 탐구 능력 평가 문항은 '토양의 공극률과 투수율의 측정', '태양의 방위각과 고도 측정', '수온 약층', '단열 팽창과 응결', '해양 지각의 이동 속도'의 5개의 실험 주제에서 28개의 실험 탐구 능력 평가 문항을 개발하여 현장에 투입하였다. 개발한 평가 문항의 내용 타당도 지수는 82.7%이고, 신뢰도(Cronbach α) 지수는 .86으로 비교적 높게 나와, 지구과학 교과의 실험을 통한 고등학생들의 탐구 능력을 측정하기 위한 검사지로 사용하는 데 적절한 것으로 판단할 수 있다. 개발한 실험 탐구 능력 평가 문항의 주제별 평균 난이도를 분석한 결과 60.4~77.9이고, 변별도 평균 지수는 .17~.37로 나타났다. 실험 주제별 문항에 따라 난이도 및 변별도가 허용 범위를 벗어난 문항도 있으나, 실험 집단이 선발 집단이고 실험 평가 문항이라는 것을 고려하면 난이도 및 변별도는 대체로 양호한 것으로 판단할 수 있다.

5명의 평가자가 평가한 평가자간 신뢰도는 상관 계수(r)가 .80이상으로, 평가자간 일치도(W)는 실험 2의 1, 2번 문항이 0.45, 0.58인 것을 제외하고는 모두 .76이상으로 상당히 높게 나왔다. 이는 논술형 주관식 문항의 평가에서 문제가 되고 있는 평가자의 주관성이 배제된 객관성 있는 평가가 되게 하기 위해서는 평가의 준거가 명확하게 기술되고 상세화된 채점표가 필요함을 의미하며, 또한 여러 명의 평가자가 평가할 때는 신뢰도가 높다고 해서 반드시 일치도가 높은 것은 아니기 때문에 평가자 간의 신뢰도 및 일치도를 모두 검증하여야 함을 반증하는 것으로 볼 수 있다. 개발한 본 실험 탐구 문항과 TESIS와의 상관 관계를 조사한 결과 상관 계수(r)는 .45로 비교적 낮은 상관을 보이고 있다. 이는 선행 연구의 연구와도 일치하는 결과로서, 실제 실험에 의한 탐구 능력의 평가는 문제 상황이 복잡하고 종합적이기 때문에 앞으로 좀 더 많은 연구가 수행되어 보다 정확하고 타당성 있는 평가가 되도록 해야 할 것이다. 본 연구를 통해 개발된 평가 도구는 현장 검증 결과 평가자의 주관성이 배제된 객관도가 높은 실험 평가 도구로 검증되었기 때문에 현재의 우리나라 여건상 중간, 기말 고사 등의 각종 실험 평가에 활용될 수 있을 것이며, 이와 유사한 연구의 방법이나 지침으로 이용될 수 있을 것이다.

본 연구에서는 연구의 시간과 여건상 결과 중심의 평가만을 중심으로 했지만 연구의 진행 과정에서 나타난 문제점과 결과를 바탕으로 좀더 바람직하고 타당도 높은 과학 탐구 능력 평가를 위한 향후의 연구 과제로는 실험 과정과 기능에 대한 과정 중심의 평가 방법에 대한 이론적, 실천적 후속 연구가 있어야 할 것이다.

참 고 문 헌

교육부(1992). 고등학교 교육과정, 교육부.

국립교육평가원(1992). 대학수학능력시험 실험 평가 문제집.

권재술, 김범기(1994). 초.중등학생들의 과학 탐구 능력 측정 도구의 개발. 한국과학교육학회지, 14(3), 251-264.

권치순(1988). 바람직한 자연과 학습 평가 방안의 탐색, 서울교육대학 과학교육 연구, 제 5집.

김동찬(1991). 지필 평가에서 나타난 학생들의 탐구 능력과 실험 과정에서 보여주는 탐구 능력과의 관계. 한국교원대학교 대학원 석사학위논문.

김성훈(1992). 문항 제작과 문항 분석 방법, 교육평가연구원, 37-65.

성태제(1995). 타당도와 신뢰도, 서울 : 양서원.

우종옥, 이항로, 이경훈(1992). 대학 수학 능력 시험의 자연과학 탐구 능력 평가를 위한 행동 요소의 추출과 평가 목표의 상세화 연구Ⅱ, 한국과학교육학회지, 12(2), 81-95.

이항로(1991). 고등학생의 과학 탐구 능력 측정을 위한 평가 도구 개발-지구과학 소재를 중심으로-, 한국교원대학교 대학원 석사학위논문.

임인재(1991). 교육, 심리, 사회 연구를 위한 통계 방법, 서울 : 박영사.

최병순, 고재중(1994). 화학 실험 과정에서의 탐구 능력 평가 도구의 개발, 한국교원대학교 부설 교과교육 공동연구소, 211-230.

허 명(1991). 중등학교의 과학 탐구능력 신장을 위한 학습 지도 및 평가 방법의 개선 방안, 한국과학교육학회지, 10(2) : 1-9.

Doran, R.L., Boorman, J., Chan, F., and Hejaily, N. (1993). Alternative Assessment of High School Laboratory Skills, *Journal of Research in Science Teaching*, 30(9), 1121-1131.

Doran, R. L., Kanis, I. B., and Jacobson, W. J.(1991). Assessing Science Laboratory Process Skills at The Elementary and Middle/Junior High Levels, Second IEA Science Study, Teachers College, Colombia University.

Kanis, I.B.(1991). Ninth grade Laboratory Skills-an Assessment, *The Science Teacher*, 29-33.

Klopfer, L.E.(1971). Evaluation of Learning in Science, In. B. S. Bloom, J. T. Hasting & G.F.Madaus(Ee.), *Handbook of Formative and Summative Evaluation of Student Learning*, New York : McGraw-Hill.

Kyle, William C. Jr.(1983) The Distinction between Inquiry and Scientific Inquiry and Why School Students Should Be Cognizant of the Distinction, *Journal of Research in Science Teaching*, 17: 123-130.

Leonard, W.H.(1983). An Experimental Study of a BSCS-Style Laboratory Approach for University General Biology, *Journal of Research in Science Teaching*, 20(9), 807-813.

Shymansky, J.A., Kyle, W.C., and Alport, J.M.(1983). The effects of new science curricula on student performance, *Journal of Research in Science Teaching*, 20 (5), 387-404.

Tamir, P.(1990). Practical Examination, In H.J. Walberg and G.D. Haertel(Ed.), *The International Encyclopedia of Educational Evaluation* Pergamon Press : Oxford, 476-481.

Tamir, P., Nussionvitz, R., and Friedler, Y.(1982). The Design and Use of Practical Tests Assessment Inventory, *Journal of Biological Education*, 16(1), 42-50.

Tamir, P., and Lunetta, V.N.(1978). An Analysis of Laboratory Activities in the BSCS Yellow Version, *The American Biology Teacher*, 40 : 426-428.

Tamir, P. and Amir, R.(1987). Inter-relationship among laboratory process skills in biology. *Journal of Research in Science Teaching*, 24(2), 137-143.

Tobin, K.G., and Capie, W.(1982). Development and validation of a group test of integrated science processes. *Journal of Research in Science Teaching*, 19, 133-141.

(ABSTRACT)

A Study on Validation by the Development of a Science Process Skills Test with Science Experiments

Woo, Jong-Ok · Lee, Hang-Ro · Kim, Seung-Hun*
(Korea National University of Education) · (Sim-Won High School)*

The purpose of this study is to develop a valid and reliable instrument, applicable to high school Earth Science class experiment.

In advance of developing items, I was selected 14 inquiry process skills and specified evaluative objectives for each of them to develop scales and criteria for them.

I developed 28 evaluation items for 5 experiment subjects among those of high school Earth Science class. The first field trial was performed a sample of 5 high school students, and the second one using a sample of 25 high school students. The results are as follows.

(1) The content validity and reliability(Cronbach α) of the developed items were 82.7% and .86, respectively, the developed instrument in this study is considered valid and reliable.

(2) The average difficulty index was .69 and the discrimination index was .30.

(3) Answer sheets based on the reported results were rated 5 teachers and Inter-rater Reliabilitiy and Inter-rater Consistency were analyzed, its indices were .80 and .76, respectively.

(4) The developed items show a low coefficient of .45 with TESIS, a set of paper-and-pencil test items developed by Lee, Hang-Ro(1991).

That the experiment assessment is solely subject to the rater's viewpoint has been one of the major problems raised concerning the matter. This research, however, shows that a set of more specified scales and criteria for the evaluation will make it more valid, reliable and efficient.