과학 탐구능력 측정을 위한 표준화 검사지 개발

-중학교 2학년의 자료 분석과 해석 능력을 중심으로-

이연우·우종옥
(한국교원대학교 지구과학교육과)

〈1991년 5월 1일 받음〉

I. 서 론

1. 연구의 배경과 필요성

매일 폭발적으로 증가하는 과학 지식(산물)을 학생들에게 모두 가르칠 수는 없으며, 또 가르친다 하더라도 불과 몇 년 후가 되면, 새로운 이론에 의해 그 지식은 별로 쓸모없는 지식이 되어 버린다. 따라서 과학 교육에서는 과학 활동의 산물인 지식자체보다는 그 지식을 창출해내는 과학적 활동과정을 중시하게 되었다. 그러므로 과학교육은 학생들에게 과학적 개념이나 원리를 발견하는 과정을 스스로 터득하게 하는 것이 중요한 목표가 된다.

최근 여러 가지 연구에 의하면, 단순한 주입식 교육보다 지식을 창출하는 과정(탐구과정)을 스스로 습득하도록 하는 교육이 과학적 사고력을 훨씬 더 신장시켜 주는 것으로 밝혀졌다. 교육개발원(이범홍, 김영민, 1983)에서 연구된 바에 의하면 탐구학습이 전통적인 학습에 비해 탐구능력 신장뿐아니라, 지식, 이해영역의 발달에도 효과적인 것으로 나타났다. 또 탐구 중심 학습과 강의 중심 학습, 그리고 양쪽의 절충식 학습을 비교 연구한 결과에서 문제 해결력을 증진시키는 데에는 실험중심의 탐구 학습형태가 절충식이나 강의 중심 학습보다 효과적인 것으로 나타났다. 그리고 탐구과정은 과학적 태도도 기를수 있다고 한다. 즉 미신에 의한 신념이 아닌 실제적인 경험을 통해서 얻은 결과를 믿는 태도를 함양할 수 있고(Davis et al., 1961), 또한 태도를 갖게 되는 세 가지 중요한 변인(Insko, 1967)을 주의 집중(attention), 이해(comprehension), 수용(acceptance)이라고 볼 때, 탐구과정을 통한 과학적 지식의 이해가 과학에 대한 태도 및 과학적 태도 변화에 중요한 역활을 한다고 볼 수 있다.

그러면 탐구 학습이 왜 일선 학교에서 제대로 잘 이루어지지 않을까?

그 원인은 입시제도, 다인수 학급 등 여러 가지 요인이 있겠지만, 허명(1987)은 그 원인을 다음과 같이 지적하고 있다. (1) 탐구지도에는 시간이 많이 소요된다. (2) 단순한 개념을 많이 전달하는 데에는 비효율적이다. (3) 교사에게 많은 부담을 준다. (4) 타당도와 신뢰도가 높은 탐구능력 평가 방법의 개발이 어렵다. 본 연구에서는 네번째의 평가도구의 개발에 관심을 가지고자 한다.

우리나라 건국 이래, 국정지표에서 과학·기술의 중요성을 논하지 않은 적이 없으며, 해마다 시도 교육위원회의 교육시책에 과학·기술 교육의 중요성을 밝히지 않은 적이 없었다. 하지만, 아직까지 과학교육에 대한 뚜렷한 성과를 제시한 자료는 없으며, 1983년부터 기초과학 육성과 과학 인구의 저변확대를 위하여, 초·중·고등학교에 새로운 과학교육 투자가 시행되고 있지만, 여기에 따르는 교육 성과 조사는 교육 외적인 요소(실험기구의 확보율, 조교의 유무, 실험실 확보율 등)에서 부분적으로 실시될 뿐, 탐구능력과 같은 학생들의

사고력의 신장에 대한 조사는 거의 실시되고 있지 못하다. 그러면 일선학교에서 과연 이러한 평가가 제대로 이루어질 수 있는 여건이 성립될까? 서울 대림여중(1985)에서 조사한 바에 의하면 중학교 과학교사 1명이 1번의 실험수업을 실시한 후, 읽어야할 실험보고서의 숫자는 평균 약 800명 분 정도이며, 1반의 보고서(70명분)를 읽는데 소요되는 시간은 약 30분 정도로서 1명의 실험보고서를 읽는데 30초도 소요하지 않는 것으로 나타났다. 이러한 여건하에서 어떻게 과학 교육평가가 제대로 이루어질 것이며, 학생들의 탐구능력의 신장정도를 추적할 수 있겠는가?

교육평가의 목적에는 여러 가지가 있지만, 그 중 가장 중요한 것의 하나가 평가에 의해 학습목표의 적절성을 확인하고, 교수 방법을 개선하며, 학생들의 학습에 있어서 어려움의 원인을 진단하고, 이러한 정보를 종합하여 미진한 부분을 재투입하는 데 있다(Gronlund, 1985). 그런데 1983년부터 기초과학 육성과 과학인구 저변 확대를 위하여 투자한 과학교육에 있어서는 탐구력 신장을 위한 학습목표의 적절성 여부나 탐구능력 중에서 학습이 잘 되지 않는 부분의 진단, 탐구학습 방법에 대한 개선 등이 제대로 이루어지지 않고 있다고 보아야할 것이다. 따라서 보다 구체적인 탐구능력 측정 방안이 모색되고, 표준화된 탐구능력 측정 평가문항이 많이 개발되고 보급되어야 할 것이다.

본 연구에서는 탐구요소들 중에서 학생들이 학습하기 어려운 요소를 진단할수 있고, 학생들의 탐구력 신장 정도를 부분적이나마 추적할 수 있는 표준화된 검사지를 개발하고자 한다.

2 연구의 제한점

본 연구에서 개발된 검사지는 다음과 같은 몇 가지 제한사항을 지닌다.

(1) 본 연구에서는 연구대상을 표집(sampling) 하는 데 있어, 연구기간과 자료처리 시간을 고려하여 중학교 2학년 학생만을 대상으로 하였다.

(2) 본 연구에서 연구대상을 표집(sampling) 하는 데 있어, 무작위 군표집(randomcluster－sampling) 방법을 사용하였다.

(3) 연구대상 표집에 있어, 군표집(cluster－sampling)의 방법은 1학급 단위로 하였으며, 이때 1학급의 크기가 시 단위 지역과 읍·면 단위 지역에서 서로 차이가 나는데, 이것은 시 단위급과 읍·면 단위급의 학생 비율과도 같다고 보았다.

(4) 본 연구에서는 탐구과정 요소 중에서 자료 해석 및 분석에 관한 요소만으로 한정하여 평가 목표와 평가 문항을 개발하였다.

(5) 표준화된 검사지(standardized test)는 동형 검사지(equivalent forms)와 비교검사지(comparable forms)가 함께 개발되어야 하지만, 여러 가지 제약조건(연구기간, 문항 개발의 어려움 등) 때문에 후속 연구에서 이루어지도록 하고, 본 연구에서는 1개의 검사지만 개발하도록 한다.

II. 과학교육 평가의 방향

평가는 의도된 학습목표(learning outcome)를 분명히 하여주고, 어떤 기간 동안 해야 할 목표(goal)를 정확히 제시하여 주며, 학습과정과 관련된 보상(feedback) 효과를 얻을 수 있고, 학생들이 특히 습득하기 어려운 학습목표가 무엇인지 또, 다음에는 무엇을 가르쳐야 될 것인지에 대한 정보를 얻을 수 있는 수단이 되므로, 평가가 올바르게만 운영된다면 교사와 학생, 그리고 학교 경영자 모두에게 이로운 것이다(Gronlund, 1985). 그러나 요즈음 우리 주변에서 인식되고 있는 평가는 단순히 개인차에 대한 등급을 매기는 것이 주요 목표로 되고 있으며, 그러한 평가도구의 성질도 일반화된 공동목표－인지적 영역, 정의적 영역, 심체적 영역이 잘 조화된 목표－를 위해 만들어진 것이라기 보다는 어떤 특수한 목적을 위해 만들어진 평가도구만을 고집하고 있다. 이러한 현상을 NSTA (National Science Teachers Association)에서도 얘기되고 있다. NSTA는 "70년대의 학교 과학교육" (School Science Education for the 70s, 1971)에서 70년대의 과학교육 목표와 평가에 대하여 몇 가지 문제점을 제시하고 있는데, 그 내용은 종전에 주로 실시되던 지필평가를 학생 자신의 평가, 절대적 표준치(criterion performance)를 기준으로 한 측정, 보다 높은 사고력을 요구하는 평가, 정의적 영역의 목표를 측정할 수 있는 평가 등으로 보완 되어야 하며, 이것에 부합하는 평가도구도 개발되어야 한다고 하였다 (Doran, 1980).

따라서 본 연구에서는 보다 신뢰성 있게 학생들의 능력을 측정할 수 있는 표준화된 검사지의 개발에 그 촛점을 맞추고자 한다.

Ⅲ. 연구의 방법

1. 개발하고자 하는 검사지의 성격과 준거 설정

탐구능력 검사는 지필검사, 실험실기 검사, 보고서 검사, 개인연구 검사 등 여러 가지 방법이 있으나, 연구자는 현실적인 여건(1학급 당 인원수의 과다, 연구기간, 문항 개발의 어려움 등)을 감안하여 지필검사로 하고자 한다. 그리고, 문항의 형식은 4지 선택형이며, 문항수는 20문항으로 하고, 중학교 2학년 학생을 대상으로 하며, 평가 시간은 30분으로 한다.

탐구능력은 ASE에서 정의한 바와같이 단순한 지시문이나 짧은 기간 동안의 학습에 의해 습득되는 것이 아니라 오랜 기간 동안의 반복된 훈련에 의해서만이 습득되는 인지능력이다. 따라서 탐구능력의 평가는 단순한 성취도 평가(achievement test)의 성격에서 벗어나 소양 평가(aptitude test)의 성격을 지녀야 한다고 생각한다.

Cronbach(1984)는 학교에서 이루어지는 모든 평가는 성취도 평가의 성격과 소양 평가의 성격을 모두 지니고 있으나, 평가 내용이 어느 쪽에 더 많은 비중을 두느냐에 따라 5개의 단계로 구분할 수 있는데(〈표 1〉 참조), 이 중에서 A단계와 D단계를 성취도 평가에, C단계와 D단계, E단계를 소양 평가에 속한다고 하였다. 그리고 성취도 평가는 주로 학교에서 이루어지는 수업내용에 크게 영향을 받는 평가이며, 소양 평가는 학교 안과 밖, 즉 다양한 교육환경에 의하여 영향을 받으며, 학교 수업 내용과는 거의 무관하여 과거에는 선천적인 것으로 생각되기도 하였다.

이와같은 이론에 의한다면 탐구능력 평가는 단순

〈표 1〉 Cronbach의 성취도 평가와 소양 평가의 분류표

성취도 평가

A	교과내용에 치중한 평가
B	일반적인 교육발달 사항에 관한 평가
C	학교생활 내용에 치중한 소양평가
D	일반생물 문화권에 치중한 언어로 된 소양평가
E	일반생활 문화권에 치중한 비언어로 된 소양평가

소 양 평 가

한 성취도 평가일 수가 없으며, Cronbach 분류표의 C단계 정도의 소양 평가는 되어야 한다고 생각한다. 따라서 탐구능력 평가의 내용은 교실에서 이루어진 수업내용에 치중한 평가(contentoriented)에서 탈피하여 우리 생활주변의 다양한 소재에서 평가 내용을 선정하여야 한다.

또, Gronlund(1985)는 다음과 같은 목적을 위해서는 Norm－Reference Test를 이용하는 것이 바람직하다고 하였다.

a) 여러 개의 학습과정에서 공통적으로 사용되는 학습목표나 기본 능력에 관한 학생들의 발달 정도를 평가하는 경우

b) 1년 주기 이상의 학생들의 학습진행을 평가하는 경우

c) 교수 목적을 위하여 학생들을 그룹화하는 경우

d) 학생들이 상대적으로 강한 부분과 약한 부분를 알아보고자 하는 경우

e) 학생들의 성취수준을 일반적인 수준과 비교하고자 하는 경우

DES(1982)에서 보고된 바에 의하면 탐구능력은 단순한 과학교과에서만 활용되는 것이 아니라, 여타 다른 과목(일반사회, 수학 등)에서도 공통으로 필요한 능력이라고 하였다. 그렇다면 이것은 위의 a)번 내용과 잘 일치한다. 또 탐구능력은 장기간의 도제에 의해서만 습득이 가능한 인지능력이므로 b)번 내용과도 잘 일치한다. 그리고 본 연구자는 개발하고자 하는 평가도구가 부분적이나마 진단평가의 성격을 지니고자 하므로 d)번과 e)번 내용과도 일치된다. 따라서 연구자가 개발하고자 하는 평가도구는 소양 평가(aptitude test)인 동시에 Norm－Reference Test이고자 한다.

2. 탐구능력 요소 선정과 평가목표 설정

(가) 탐구능력 요소의 선정

탐구능력 요소의 선정은 검사지의 준거와 성격에서 밝힌 몇 가지 제약조건을 만족하는 범위에서 선정하였다. 즉 지필검사로 가능한 요소라야 하고, 중학교 2학년 학생이 30분 이내에 완성하기 위해서는 평가 문항이 20문항 정도이어야 하며, 부분적이나마 각 요소별로 학생들의 능력을 진단할 수 있어야 한다. 따라서 선정되는 요소의 수는 4－5개 정도가 필요하다. 또 이러한 제약조건을 만족하면서, 어느 한 영역을 모두 평가할 수 있는 부분은 허명(1984)의 SIEI에서 두번째 영

역인 자료해석 및 분석이었다. 허명의 자료 해석 및 분석에는 5개의 요소가 제시되어 있지만, 외삽(extrapolation)과 내삽(interpolation)은 예상(prediction)에 포함시켰으며, 허명의 분류표에는 나와 있지 않은 증거제시(verification)는 원인설명(causal explanation)에 포함시켰다.

(나) 탐구능력 요소의 정의와 평가목표 설정

위에서 선정한 4개의 요소에 대해 정의를 내리기 위하여 허명(1984), Molitor와 George(1976), Abraham과 Nelson(1973), Anderson(1976)을 참조하고(〈부록 1〉 참조), 설문지, 개인 면담, 세미나 등을 통하여 전문가의 자문을 얻어 〈부록 2〉과 같은 정의를 내렸다. 여기서 〈부록 2〉의 내용은 〈부록 1〉의 내용을 대치하는 것이 아니라 보완하는 것이다.

이와같은 방법으로 얻은 요소의 정의에 의하여 평가목표를 설정하고, 이것을 과학교육 전문가에게 돌려 자문을 구한 후, 수정 보완하여 〈부록 2〉과 같은 평가목표를 설정하였다.

3. 검사지의 개발

중학교 학생들의 인지수준, 설정된 평가목표, 그리고 Gronlund(1985)가 제시한 선택형 문항을 개발할 때 고려해야 할 14가지 사항 등을 참조하여 문항을 개발하였으며, 이것을 2차에 걸쳐 과학교육 전문가 6명에게 의뢰하여 내용의 타당도, 문항의 명료성, 정답의 객관도, 평가목표와의 일치정도 등을 점검하였다.

이렇게 하여 개발된 검사지는 각 탐구요소 별로 5문항씩 총 20문항이며, 내용 타당도 지수(CVI)는 85%이고, 정답의 객관도 지수는 91.7%이다.

4. 검사지의 현장검증과 수정보완

(가) 1차 현장검증

1차 현장검증은 학생들이 검사지를 완성하는데 걸리는 시간의 적절성 여부와 문항의 수준이 중학생들에게 맞는지, 문항 개발이 이상적으로 잘 되었는지 등을 알아보기 위하여 실시하였으며, 검사결과에 의해 결함이 있다고 판단되는 문항은 수정 보완하였다.

1) 검사대상의 표집

검사대상의 표집방법은 학급단위로 군표집(cluster sampling) 방법을 사용하였으며, 표집대상은 중학교 2학년 학생이고, 표집규모는 시 단위급에서 남·녀 각 1학급씩과 읍면 단위급에서 남·녀 각 1학급씩으로 총 199명으로 하였다.

2) 문항 분석

문항분석은 개발된 검사지의 문항이 측정대상의 평균적인 능력과 어느 정도 일치하는지, 문항자체의 결함은 없는지 등을 알아보기 위한 것이다.

정답율은 Doran(1980)에 의하면 대략 20−80% 정도가 무난하며, 전체 정답율의 평균은 4지 선다형의 경우 63%가 가장 이상적이라고 하였다. 1차 현장검증의 결과 15번 문항의 정답율이 91.46%나 되므로, 버리고 새로운 문항으로 대치하기로 한다.

변별도 지수는 +1.0에서 −1.0까지 나올 수 있는데 +쪽으로 큰 값이 나올수록 그 문항은 좋은 문항이며, −값이 나오면 그 문항은 삭제해야 할 문항이다. 그리고 보통 +0.2 이상이면 무난한 것으로 해석되고 있다(Gronlund, 1985).

본 검사지에서는 16번 문항이 +0.05가 나왔으므로 삭제하기로 하였다.

변별도가 정답에 표기한 학생들 중에서 상위 그룹에 속하는 학생들이 하위 그룹에 속하는 학생들보다 더 많아야 한다는 전제하에서 문항의 적절성을 평가하는 것이라면, 오답의 효율성은 정답을 제외한 나머지 답지(오답)에서는 하위 그룹에 속하는 학생들이 상위 그룹에 속하는 학생들보다 더 많이 표기되어야 한다는 전제하에서 문항의 적절성을 평가하는 것이다. 그리고 아무도 선택하지 않은 답지(오답)가 있다면 이것 역시 답지로써 역활을 못하므로 수정되어야 한다.

1차 검증의 결과, 6번문항의 (A)답지와 16번 문항의 (A)답지, 20번 문항의 (C) 답지가 상위 그룹의 학생이 하위 그룹의 학생보다 더 많이 선택하였으므로 오답으로써의 결함이 있다고 생각되어 수정하기로 하였다.

이상과 같은 문항분석에 의해 15번 문항과 16번 문항은 삭제하고 새로운 문항을 개발하여 대체하였으며, 6번 문항의 (A)답지와 20번 문항의 (C)답지는 수정하여 새로운 검사지를 만들어 2차 현장검증을 하였다.

(가) 2차 현장검증

1) 검사대상의 표집

1차 현장검증에서 어느 정도 문항의 효율성이 검증되었기 때문에, 2차 현장검증에서는 단지 수정된 문항

에 대한 효율성과 적절성만을 알아보기 위한 것이므로, 시 단위급에서 남·녀 각 1학급씩 총 2개교에서 108명을 표집하여 검증을 실시하였다.

2) 문항 분석

2차 현장검증의 문항분석은 1차 현장검증의 문항분석과 같은 방법을 사용하였다.

각 문항의 정답율은 13번 문항이 다소 높게 나왔으나 대체로 받아 들일 수 있을만큼 양호하게 나왔다고 판단하였으며, 전체 정답율의 평균은 68% 정도로써 이상적인 값 63% 보다는 다소 못하지만 그런대로 괜찮은 값이라고 생각하였다.

각 문항의 변별도는 모두 +0.2 이상이므로 수정할 사항이 없었으며, 오답의 효율성은 4번 문항의 (C)답지와 10번 문항의 (A)답지에서 상위 그룹 학생수가 하위 그룹 학생수보다 1명 더 많게 나왔으나, 1차 현장검증 때에 결함이 없는 것으로 나타난 답지들이므로 수정하지 않고 그냥 사용하기로 하였다.

5. 검사지의 표준화 작업

표준화 작업은 완성된 검사지를 가지고 전국 단위의 모집단(universe)에게 적용하여, 검사결과 해석의 기준이 되는 표준점수를 구해내는 작업이며, 검사 실시단계에서 부터 표준조건(standard condition)을 유지하면서 검사를 실시하여야 한다.

1) 검사 대상의 표집

원래 표준화 작업은 모집단(universe) 전체에게 검사를 실시하여야 하는 것이 원칙이나, 통상 모집단에 가까운 표집집단(sample group)을 선별하여 적용한다. 여기서 표집집단의 선별방법은 무작위 표집(random sampling)을 적용하지만, 연구자는 무작위 군표집(random cluster-sampling)으로 하였다.

먼저 표집대상이 특정 지역에 편중되는 것을 방지하기 위하여 제주도를 제외한 전국을 5개의 지역군으로 나누어, 1개의 지역군에 2개 도(道)가 포함되도록 하였다. 그리고 2개 도(道) 중에서 무작위로 1개 도를 선별하여 도시지역 표집대상으로 하고, 나머지 1개 도는 시골지역(읍·면 단위급) 표집대상으로 삼았다.

시골지역 표집대상이 된 도(道)의 경우에는 그 도(道) 내에 있는 군(郡)들 중에서 무작위로 1개의 군을 표집하고, 다시 그 군(郡) 내에 있는 읍과 면들 중에서 무작위로 선별하였다. 이때 읍의 경우에는 보통 남자 중학교와 여자 중학교가 각각 1개교 씩 있으므로, 이 2학교가 모두 표집대상이 되지만, 면단위의 경우에는 남·녀 공학인 중학교가 1개교 뿐이므로, 면단위가 선별되는 경우에는 2개의 면을 선정하여 1개의 면은 남학생을, 다른 1개의 면은 여학생을 1학급씩 표집하도록 하였으며, 도시지역의 선별방법도 시골지역과 비슷하게 실시하였다. 이렇게 하여 선정된 표집대상 학교와 표집규모는 〈표 2〉와 같다.

〈표 2〉 표준화 작업의 표집대상과 규모

( ) 안은 학생수

	도 시	지 역	시 골 지 역		합계
	남학생	여학생	남학생	여학생	
Ⅰ 지역 (서 울)	강남구 신사동 신사중 (67)	강남구 신사동 신사중 (47)			2개교 (226)
	영등포구 신길동 대영중 (56)	영등포구 신길동 대영중 (55)			
Ⅱ 지역 (경 기 강 원)	강원도 원주시 진광중 (55)	강원도 원주시 치악중 (54)	여주군 여주읍 여주중 (51)	여주군 여주읍 여주여중 (52)	4개교 (212)
Ⅲ 지역 (충청도)	충북 청주시 서원중 (52)	충북 청주시 중앙여중 (56)	공주군 우성면 우성중 (47)	공주군 이인면 이인중 (51)	4개교 (206)
Ⅳ 지역 (전라도)	전주 직할시 북성중 (56)	전주 직할시 정신여중 (54)	정읍군 신태인읍 신태인중 (54)	정읍군 신태인읍 왕신여중 (30)	4개교 (214)
Ⅴ 지역 (경상도)	경남 마산시 중앙중 (55)	경남 마산시 양덕여중 (55)	달성군 현풍면 현풍중 (43)	달성군 화원면 달성중 (49)	4개교 (202)
합 계	6 개교 (314명)	6 개교 (322명)	4 개교 (195명)	4개교 (202명)	18개교 (1033)

2) 검사 실시 방법

표준화 검사지는 검사시행 단계에서부터 훈련된 감독관의 감독하에서 표준조건(standard condition)을 유지하면서 검사가 실시되어져야 한다. 따라서 본 연구에서는 연구자가 직접 모든 표집대상 학교에 방문하여, 연구자가 직접 감독하면서 표준조건을 유지하였다.

부정행위는 성적과 관계없이 습관적으로 일어날 수 있기 때문에, 먼저 학생들 사이의 간격을 띄우도록 하였으며, 답안지를 나누어 주고, 학교명과 이름을 쓰게 한 뒤, 검사의 목적, 검사결과의 처리방법과 용도, 답안지 작성요령, 모르는 것이 있을 때의 처리방법 등을 자세히 설명한 뒤, 충분한 동기유발이 되어 최대한 성의껏 답안지를 작성할 수 있도록 하였으며, 검사지를 나누어 주고 다시 한번 더 확인하였다. 그리고 검사시작 시각과 검사종료 시각을 종이에 기록하면서 정확히 체

크하였으며, 검사가 진행되는 동안에 주변의 방해사항이 없도록 최대한의 정숙을 유지하도록 힘썼다.

3) 문항 분석

문항 분석은 2차 현장검증시에 약간의 결함이 발견된 문항이 있었으나 별로 큰 결함이 아니기에 무시하고 넘어 갔었다. 그러나 그 문항이 전국 표집단계에서도 역시 결함이 발견된다면, 그 문항은 수정되어야 하므로 다시 문항 분석을 하여 보았다.

〈부록 3〉에서 보는 바와 같이 정답율은 대체로 양호한 편이며, 변별도도 8번 문항이 약간 낮은 듯 하지만 전부 양호한 편이다. 그리고 오답의 효율성도 2차 현장검증시에는 4번 문항과 10번 문항에 약간의 결함이 있었지만, 전국 표집단계에서는 아주 양호하게 나왔다. 신뢰도(K-R 20)는 0.69로 나왔다.

4) 표준점수(standard score) 설정

표준점수는 모집단의 평균치를 기준으로 하여 설정하는 것으로 검사결과를 해석 하는 데 아주 용이한 방법으로 알려져 있다. 표준점수에는 Z-score, T-score, IQ지수 Stanine 등이 있으나, 본 연구에서는 평균이 100이고, 표준편차가 15인 IQ지수를 표준점수로 사용하기로 하였다.

설정된 표준점수는 〈표 3〉에 제시된 바와 같으며, 여기서 도시지역에 있는 학생들의 검사결과를 도시지역내의 학생들간에 비교하고자 할 경우 혹은 남학생들끼리 비교하고자 하는 경우 등에는 지역단위 표준점수(local Score)를 사용할 수도 있다. 따라서 지역단위 표준점수도 함께 산출하여 제시하였다.

〈표 3〉을 보는 방법은, 예를 들어 본 검사지로 측정한 결과가 16개를 맞은 경우, 이 학생의 표준점수(standard score)는 110점이 된다. 그런데 본 검사지 자체의 오차가 8.3이므로 이 학생의 실제 점수는 101.7에서 118.3 사이에 있다고 보면 된다.

〈표 3〉 맞는 갯수에 따른 표준 점수 산출표

맞은 갯수	Standard Score	Local Score			
		도 시	시 골	남학생	여학생
0	38	28	45	36	38
1	42	33	49	40	43
2	47	38	54	45	48
3	51	43	58	49	52
4	56	48	62	54	57
5	60	53	67	58	62
6	65	58	71	63	66
7	69	63	76	67	71
8	74	68	80	72	75
9	78	73	85	76	80
10	83	78	89	81	85
11	87	83	93	85	89
12	92	87	98	90	94
13	97	92	102	95	99
14	101	97	107	99	103
15	106	102	111	104	108
16	110	107	116	108	112
17	115	112	120	113	117
18	119	117	124	117	122
19	124	122	129	122	126
20	128	127	133	126	131

5) 탐구능력 요소별 진단 기준 설정

실제로 진단평가의 성격을 지닌 검사지는 1개의 요소별 문항수가 적어도 10문항 이상이 되어야 한다(Gronlund, 1985). 그런데 본 검사지는 1개의 요소별 문항수가 5문항이므로 대략적인 진단밖에 할 수가 없다. 따라서 본 연구에서는 탐구능력 요소별로 단순히 상, 중, 하 정도의 진단만 할 수 있도록 진단기준을 설정하였다. 그 기준은 대략 평균을 중심으로 하여 68% 이내에 들면 "중"으로 평가하고, 그 이상은 "상", 그 이하는 "하"가 되도록 하였다(〈표 4〉 참조).

〈표 4〉 탐구 능력 요소별 진단기준

탐구능력	문항 번호	평균 (맞은 갯수)	표준 편차	진단 기준(맞은 갯수)		
				상	중	하
추 리	1, 8, 9, 15, 18	3.80	1.06	4.87 이상	2.74-4.86	2.73 이하
관계설정	5, 11, 12, 13, 16	3.23	1.18	4.42 이상	2.05-4.41	2.04 이하
원인설명	2, 4, 10, 19, 20	3.17	1.11	4.29 이상	2.06-4.28	2.05 이하
예 상	3, 6, 7, 14, 17	3.56	1.22	4.79 이상	2.34-4.78	2.33 이하

Ⅳ. 개발된 검사지에 대한 기술

1. 검사지의 개관

본 연구에서 개발된 검사지의 특성은 〈표 5〉에서 제시된 바와같이 탐구능력 중에서 데이타 해석 및 분석능력(추리, 관계설정, 원인설명, 예상)에 대하여 각 요소별 5문항씩, 총 20문항으로 되어 있으며, 중학교 2학년 학생을 대상으로 하며, 검사 시간은 30분이다.

〈표 5〉 데이타 해석 및 분석능력 검사지의 특성

문 항 수	20 문 항
검사지의 형태	객관식 4지 선택형
적용 대상	중학교 2학년 학생
검사 시간	30분
변 별 도	0.00 – 0.20 (1 문항) 0.21 – 0.40 (10 문항) 0.41 – 0.60 (9 문항) 평균 : 0.39
정 답 율 (%)	21.0 – 40.0 (2 문항) 41.0 – 60.0 (4 문항) 61.0 – 0.60 (9 문항) 81.0 – 100 (5 문항) 평균 : 68.8%
신뢰도(K-R 20)	0.69
표준 편차 (SD)	3.31
표준 오차 (SE)	1.84

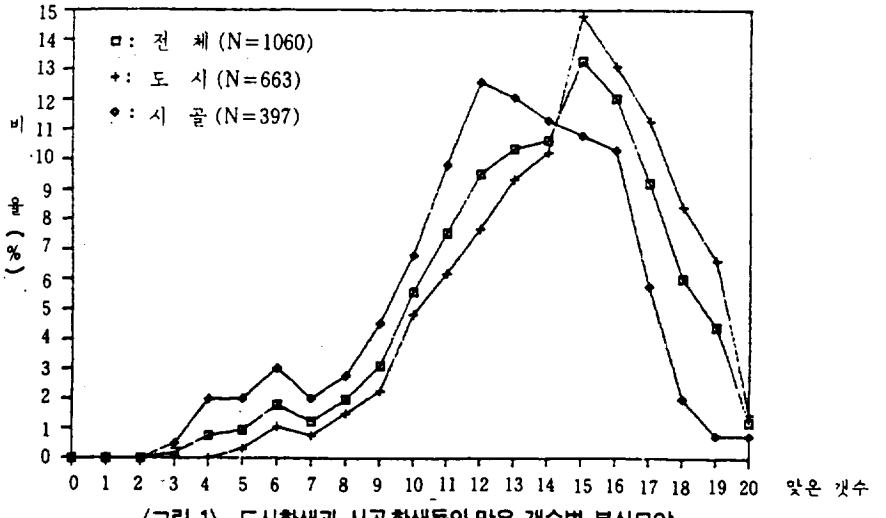
내용타당도 지수는 85%, 정답의 객관도 지수는 91.7%, 평균 정답률은 68.8%, 평균 변별도 지수는 0.39, 표준편차(SD)는 3.31, 표준오차(SE)는 1.84, 신뢰도(K-R 20)는 0.69이다.

2 맞은 갯수별 분산 모양

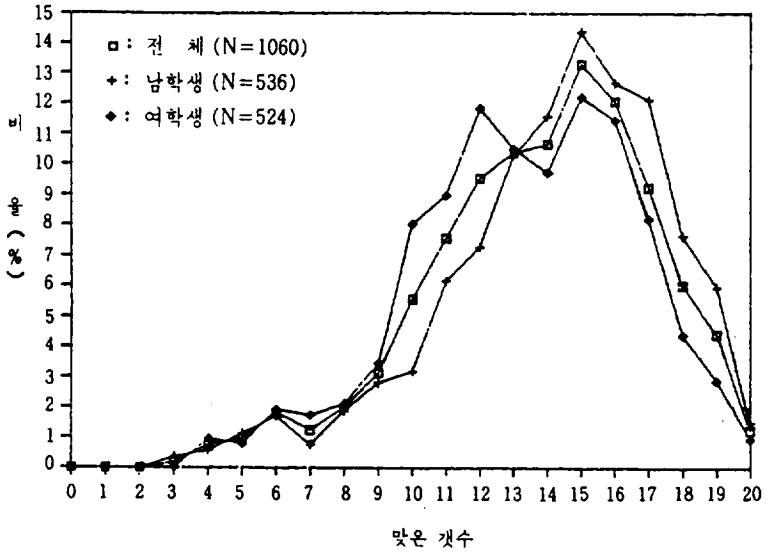
가장 이상적인 검사지의 분산모양은 50%(10개)를 중심으로 하는 정상분포 곡선이 되는 것이다. 하지만 본 검사지는 4지 선택형이므로 63%(12.5개)를 중심으로 하는 정상분포가 되어야 한다. 그러나 실제 곡선은 최빈값이 15개(75%)이며, 우측으로 기울어진 분포곡선을 보이고 있다. 도시지역과 시골지역의 분포곡선을 비교하여 보면, 도시지역은 최빈값이 15개에서 나타나지만, 시골지역은 12개에서 나타나고 있으며, 시골지역의 학생들이 도시지역 학생들보다 전체적으로 낮은 쪽으로 편향된 분포곡선을 보이고 있다(〈표 6〉와 [그림 1] 참조).

또 남학생과 여학생의 분포곡선을 비교하여 보면, 최빈값은 남학생이나 여학생이나 모두 15개(75%)에서 나타나지만, 전반적인 분포곡선의 모양은 남학생보다 여학생이 낮은 쪽으로 편향되고 있음을 알 수 있다(〈표 6〉와 [그림 2] 참조).

그리고, 여기서 도시학생과 시골학생의 분포곡선의 편향정도가 남학생과 여학생의 분포곡선의 편향정도보다 더 심하게 나타나고 있다. 이것은 남학생과 여학생의 능력 차는 거의 없지만, 도시학생과 시골학생의 능력 차는 조금 있는 것으로 해석된다.



〈그림 1〉 도시학생과 시골 학생들의 맞은 갯수별 분산모양



〈그림 2〉 남학생과 여학생들의 맞은 갯수별 분산모양

〈표 6〉 맞은 갯수별 돗수 분포표

맞은갯수	전 국	도 시	시 골	남학생	여학생
0	0	0	0	0	0
1	0	0	0	0	0
2	0	0	0	0	0
3	2	0	2	2	0
4	8	0	8	3	5
5	10	2	8	6	4
6	19	7	12	9	10
7	13	5	8	4	9
8	21	10	11	10	11
9	33	15	18	15	18
10	59	32	27	17	42
11	80	41	39	33	47
12	101	51	50	39	62
13	110	62	58	55	55
14	113	68	45	62	51
15	141	98	43	77	64
16	128	87	41	68	60
17	98	75	23	55	43
18	64	56	8	41	23
19	47	44	3	32	15
20	13	10	3	8	5
합 계	11060	663	397	536	524

V. 요약 및 결론

최근의 여러 가지 연구결과에 의하면 탐구학습 방법이 탐구능력뿐만 아니라, 지식 및 이해영역의 발달은 물론이고, 과학적 태도까지도 기를 수 있다고 한다. 그럼에도 불구하고 일선학교에서 탐구학습이 잘 이루어지지 않고 있다. 그 이유에는 여러 가지가 있겠으나, 그 중의 하나가 바람직한 평가도구가 부족하다는 것이다. 그래서 본 연구에서는 데이타 해석 및 분석 능력에 해당되는 탐구능력에 대하여 표준화된 검사지를 개발하고자 하였다.

탐구능력은 단 시일에 형성되는 인지능력이 아니라, 오랜 기간동안의 반복된 훈련에 의해서만이 획득이 가능하기 때문에 탐구능력 평가는 성취도 평가의 성격을 지닌다기 보다는 소양 평가의 성격을 지닌다고 생각하였다. 따라서 평가문항은 우리생활 주변의 다양한 소재로부터 구하려고 하였으며, 각 탐구능력 요소별로 그 의미를 분명히 하기 위하여, 탐구능력 요소의 정의를 내리고, 그 정의에 의하여 평가목표를 설정하였으며, 설정된 평가목표에 의해 개발된 문항을 2차에 걸쳐 타당도와 객관도, 문항의 명료성 등을 점검한 뒤, 다시 2차에 걸친 현장검증에 의해 수정·보완 하여 검사지를 완성하였다. 완성된 검사지를 전국 18개교에서 중학교 2학년 학생 1060명을 무작위 군표집하여 표준화 작업을 하였다. 개발된 검사지의 내용타당도 지수는 85%, 정답의 객관도 지수는 91.7%, 평균 정답율은 68.8%, 평균 변별도 지수는 0.39, 표준편차는 3.31, 신뢰도(K-R 20)는 0.69이다.

평가의 가장 주요한 목적은 학생들의 능력을 등급(grade)으로 나누는 데 있는 것이 아니라, 학습의 효과를 측정하거나, 미진한 부분을 점검하고, 그 원인을 알아내어 재투입하는 데 있다. 따라서 본 검사지 개발의 주안점을 학생들의 능력에 등급을 매기는 데 두지 않고, 보다 심층적인 학습의 부진한 원인이나 학생 개별의 능력정도를 식별하는데 두었다. 그리고 본 검사지는 표준화된 검사지이기 때문에 학생들의 탐구능력 신장정도를 추적할 수 있을뿐 아니라, 탐구능력 중에서 학생들이 잘 하는 부분과 잘 못하는 부분을 진단할 수도 있다. 따라서 이렇게 개발된 검사지들이 일선학교에 많이 보급된다면, 학생들의 탐구능력을 신장시키는 데 많은 도움을 줄 수 있으리라 생각된다.

VI. 제 언

본 연구를 진행하는 동안에 다음과 같은 필요성을 발견하였기에 여기에 그것들을 진술하고자 한다.

첫째, 본 연구에서는 지필고사에 의한 표준화 작업을 실시하였지만, 실제로 탐구능력 요소들 중에는 관찰방법에 의한 과정평가(procedure evaluation)나 보고서 평가(product evaluation) 등에 의해서 더 큰 효과를 얻을 수 있는 영역이 있다. 이 분야에 대한 표준화 작업도 매우 어렵긴 하지만, 개발되어져야 한다고 생각한다.

둘째, 각 탐구능력 요소별로 경계구분이 애매모호하여 어떤 탐구능력 요소들은 서로 포함관계에 있는 경우도 많이 있다. 이러한 포함관계나 독립관계를 규명하여 탐구능력 요소들의 정의를 분명히 내려야 할 필요가 있다.

셋째, 표준화된 검사지에는 동형검사지(equivalent forms)와 비교검사지(com－parable forms)가 함께 개발되어야 하지만, 연구기간과 문항개발의 어려움 때문에 본 연구에서는 개발하지 않았다. 따라서 후속 연구가 계속 이루어져 비슷한 검사지가 많이 개발되어져야 한다고 생각한다.

넷째, 그동안 국내에서 개발된 평가도구들에 대한 표준화 작업은 거의 이루어지지 않은 상태이다. 따라서 이 분야에 대한 계속적인 연구가 있어야 하겠다. 특히 표집집단을 구성하는 데 있어, 지역경계와 사회·경제지표(socio-economic level)의 경계 등을 어떻게 구분할 것인지에 대한 연구가 있어야 하겠다.

다섯째, 본 연구와 같이 개발된 평가도구들을 일선학교에 많이 홍보하여, 실제로 활용될 수 있도록, 대학의 연구소내에 그러한 홍보기관이 있었으면 한다.

참 고 문 헌

대림여중(1985), 실험목표 상세화에 의한 탐구능력 측정 방안, 서울시 과학과 시범학교 운영보고서, pp.21－48.

이범홍·김영민(1983), 과학과 수업과정 모형과 평가방법 개선 연구－국민학교 자연과를 중심으로－, 한국교육개발원 연구보고, RR.83－7, pp.21－53.

허 명(1987), 탐구학습의 이론과 실제, 과학교육, 24(4) : 22－29.

Cronbach, L. J.(1984), Essential of Psychological Testing 4th ed., New York : Harper & Row, pp. 75－135.

DES(1978), Science Progress Report 1977－78, Assessment of Performance Unit, pp. 2－22.

DES(1982). Science in Schools, Age 15 : Report No. 1, Assessment of Performance Unit, pp. 1－5, pp. 75－167.

Doran, R. L.(1980). Dasic Measurement and Evaluation of Science Instruction, National Science Teachers Association, Washington D. C., pp. 13－18.

Gronlund, N. E.(1985). Measurement and Evaluation in Teaching 5th ed., Macmillan Publishing Co., New York, pp. 5－21, pp. 169－212, pp. 263－319, pp. 346－378.

Insko, C. A.(1967). Theories of Attitude Change, Prentice Hall Inc., Englewood Cliffs, New Jersey, pp. 35－40.

Molitor, L. L., George, K. D.(1976). Development of a Test of Science Process Skills, Journal of Research in Science Teaching, 13(5) : 405－412.

Myung Hur(1984), The Analysis of Inquiry Learning among High School Diology Students and its Application to the Development of an Instrument for Evaluating Inquiry Activity in Science Curricula, Ph. D. Thesis, Columbia University.

Nellist, J., Nicholl, D., (ed.) (1986). ASE Science Teacher's Handbook, London : Hutchinson & Co., pp. 1－39.

Nelson, M. A., Abraham, E. C.(1973). Inquiry Skill Measures, Journal of Research in Science Teaching, 10(4) : 291－297.

ABSTRACT

The Development of A Standardized Test of Science Inquiry Skills : Interpreting and Analyzing Data for Eighth Grade Students

Youne-Woo Lee, Jong-ok Woo
(Korea National University of Education)

This study has formed a clear definition of the elements of inquiry skills : inference, determining relationship, causal explanation, prediction, and created the goals of assessment and the items of assessment. They have been checked the validity and the objectivity and the clarity of the items by six professors of science education. At the same time, the two times of the field trial has been executed, and checked the discriminating power and the difficulty index and the effectiveness of distracters, and modified the items.

The test developed in this way was administered to 1060 students of the eighth grade, randomly cluster-sampled from the universe, and standardized.

The test is the aptitude test as well as the norm-reference test, and has twenty items. The testing-time is thirty minutes. And the cotent validity is 85%, the objectivity of the answer keys 91.7%, the mean of items difficulty 68.8%, the mean of discriminating power 0.39, the standard deviation 3.31, the reliability(K-R 20) 0. 69.

Because it is the standardized test, it can diagnose the well-developed skills and the ill-developed skills of the students, and monitor the development of skills.

〈부록 1〉 데이타 해석에 관한 탐구능력 요소의 정의

탐구능력 요소	탐 구 능 력 요 소 의 정 의
추 리 (inference)	· 관찰한 데이터로부터 새로운 사실을 찾아내는 과정이며, 추리된 사실은 직접 관찰 가능한 것이 아니라, 관찰된 증거와 과거의 경험에 의해 강하게 지지받는 것이다(허 명, 1984). · 물체나 사건의 관찰된 성질을 근거로 하여, 물체나 사건을 관찰할 수 없는 성질을 판단할 수 있는 능력(Molitor & George, 1976). · 가설의 연장선상에서 탐험할 수 있는 영역에 있는 관찰로부터 탐험할 수 없는 영역으로 투사하는 능력(Abraham & Nelson, 1973).
관계 설정 (determining relationship)	· 2개 이상 변인 사이의 관계를 알아낼 수 있는 능력(허 명, 1984)
원인 설명 (causla explanation) [증거제시(verification) 포함]	· 원인설명 : 주어진 효과의 원인을 알아낼 수 있는 능력(허 명, 1984) · 증거제시 : ① 자기가 선택한 추리에 타당도를 부여하기 위해 충분하고 필요한 관찰된 사실을 선택할 수 있는 능력(Molitor & George, 1976) ② 추리를 증명할 수 있는 능력(Abraham & Nelson, 1973)
예 상 (prediction) [외연(extrapolation) 및 내삽(interpolation) 포함]	· 예상 : 특별한 실험상황에서 일어날 수 있는 것을 예언하는 것으로 완전하지 않은 현상이나 정보에 대해 실험적인 설명을 가할 수 있는 능력(허 명, 1984) · 외연 및 내삽 : 얻어진 데이터를 가지고, 그 범위밖의 영역까지 확장시킬 수 있는 능력(Anderson, 1976)

〈부록 2〉 연구자가 내린 탐구능력 요소의 정의와 평가목표

탐구능력 요소	탐구능력 요소의 정의	평 가 목 표
추 리 (inference)	· 이미 일어난 사건에 대해 유추하는 것으로, 다소 즉흥적이며, 덜 체계적이고, 여기서 유추된 사실은 실험에 의해 증명될 가능성이 희박하다.	· 사건의 순간장면(snapshot)을 표, 그래프, 그림, 언어로 주어질 때, 이것에 근거한 새로운 사실이나 성질을 찾아낼 수 있다.
관계 설정 (determining relationship)	· 관찰된 사실 내에 들어 있는 변인들 사이의 관계를 알아내는 것으로 주어진 상황에 들어있지 않은 부분을 유추하는 것이 아니라, 단순히 주어진 상황에서 관계를 읽어내는 것이다.	· 과학적 사실을 표, 그래프, 그림, 언어로 주어지면, 그 사실내에 들어 있는 변인들 사이의 관계를 설정할 수 있다.
원인 설명(causal explannation) [증거제시(verification) 포함]	· 이미 관찰된 사실이나 선험적 지식을 근거로 하여 효과의 원인을 설명하거나, 증거를 제시하는 능력으로써 보다 논리적이고, 체계적이다.	· 실험 혹은 자연적 결과나 추리된 사실이 표, 그래프, 그림, 언어로 주어지면, 그 원인(증거)을 유추할 수 있다.
예 상 (predction) [외연(extrapolation) 및 내삽 (interpolation) 포함]	· 아직 일어나지 않은 미래의 상황을 유추하거나, 사건의 일련성(sequence)을 유추하는 것으로, 보다 체계적이며, 앞으로의 실험이나 다른 사건에 의해 증명될 가능성이 많다.	· 사건의 연속적인 장면이나 여러 가지 실험결과를 표, 그래프, 그림, 언어로 주어질 때, 데이터가 없는 부분이나 주어진 데이터 밖의 영역까지 유추할 수 있다.

(부록 3) 문항 분석표(표준화 작업)

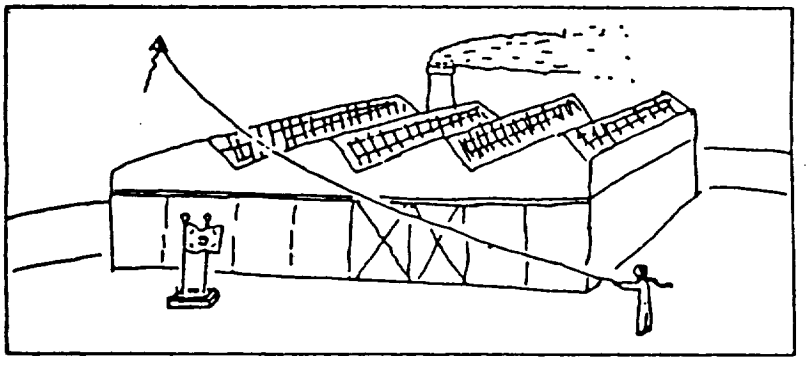
문항 번호		선택형 보기 (A)	(B)	(C)	(D)	무답	변별도 (T=600)	정답율 (N=1060)	문항 번호		선택형 보기 (A)	(B)	(C)	(D)	무답	변별도 (T=600)	정답율 (N=1060)
1	U	0	6	0	294•	0	0.35	83.35	1	U	13	247•	16	24	0	0.49	56.60
	L	17	53	42	188•	0				L	81	99•	68	52	0		
2	U	213•	37	50	0	1	0.47	46.42	12	U	6	1	16	277•	0	0.53	66.98
	L	73•	86	102	38	0				L	70	19	92	117•	2		
3	U	273•	8	16	3	0	0.41	72.36	13	U	0	0	2	298•	0	0.33	88.40
	L	150•	54	63	32	1				L	43	24	33	200•	0		
4	U	298•	0	1	1	0	0.28	87.45	14	U	0	1	293•	6	0	0.38	82.36
	L	213•	19	9	59	0				L	23	49	180•	48	0		
5	U	57	61	164•	18	0	0.34	31.32	15	U	20	5	3	272•	0	0.37	73.30
	L	78	106	61•	55	0				L	43	40	55	161•	1		
6	U	5	53	225•	17	0	0.47	49.43	16	U	4	290•	5	1	0	0.46	79.43
	L	34	86	85•	94	1				L	66	153•	63	18	0		
7	U	2	273•	25	0	0	0.45	72.83	17	U	0	8	289•	3	0	0.45	79.15
	L	27	139•	98	36	0				L	26	78	155•	40	1		
8	U	287•	11	1	0	1	0.20	86.32	18	U	2	0	273•	25	0	0.32	76.98
	L	228•	52	10	10	0				L	13	13	176•	98	0		
9	U	35	19	4	242•	0	0.45	59.43	19	U	28	269•	2	1	0	0.30	73.49
	L	86	83	24	107•	0				L	93	178•	12	16	0		
10	U	58	171•	23	47	1	0.34	37.83	20	U	0	11	18	271•	0	0.38	71.98
	L	69	69•	94	67	1				L	10	18	115	156•	1		

●: 정답 평균: 13.7604(68.80%) 표준편차: 3.3088 KR 20: 0.69

〈부록 4〉 문항의 예

■ 추리에 관한 문항의 예

아래의 그림은 한 학생이 연을 날리는 모습을 주위 환경과 함께 그린 장면이다.



위의 그림에서 모순되는 상황이 있다면 어느 것입니까?

(A) 국기의 펄럭임
(B) 굴뚝의 연기
(C) 공장의 지붕모양
(D) 연이 날르는 모습

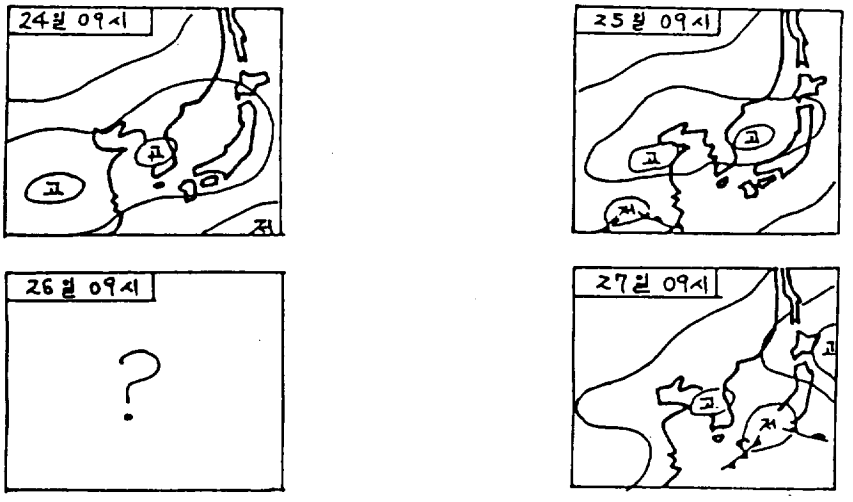
■ 원인 설명에 관한 문항의 예

영희는 학교로 오는 도중에 덧신을 잃어 버렸다. 영희의 친구들은 영희의 덧신을 찾아 주고자 한다. 영희의 친구들이 영희의 덧신을 찾는데 도움이 될만한 사실은 다음 중에서 어느 것입니까?

(A) 영희의 덧신을 빨간 색이었다.
(B) 영희의 덧신은 매우 따뜻했다.
(C) 영희의 덧신은 아주 아름다웠다.
(D) 영희의 덧신은 생일 선물로 받은 것이었다.

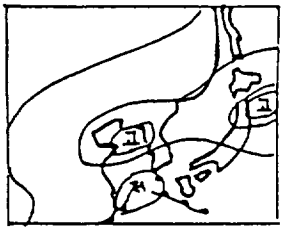
■ 예상에 관한 문항의 예

아래의 그림은 우리나라 주변의 3일간의 일기도를 간략하게 그린 그림이다. 일기도를 자세히 관찰한 뒤 물음에 답하시오.

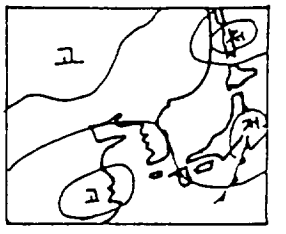


26일 09시 우리나라 주변의 일기도의 모습은 다음 중 어느 것이라 생각됩니까?

(A)

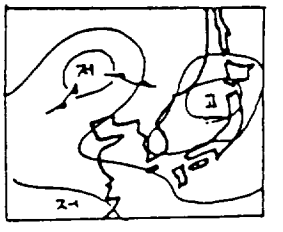


(B)



(C)

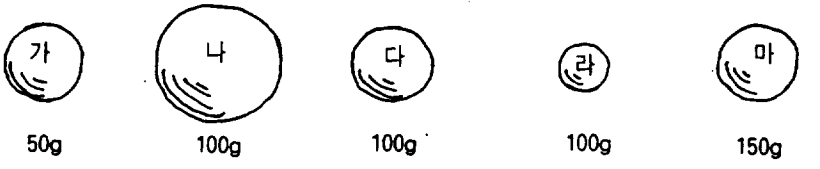
(D)



■ 관계설정에 관한 문항의 예

철수는 물체가 떨어지는 정도를 알아보기 위하여 다음과 같은 실험을 하였다.

아래 그림과 같은 공을 5개 준비하였는데, (가), (다), (마)는 공의 크기가 서로 같고, (나)는 (가)보다 크며, (라)는 (가)보다 작다. 그리고 각각의 질량은 아래에 표시된 것과 같다.



위의 공을 10m 높이에서 가만히 떨어뜨리고, 떨어지는 시간을 측정해보니 다음 표와 같았다.

공	떨어진 시간
(가)	1.03 초
(나)	1.21 초
(다)	1.02 초
(라)	9.85 초
(마)	1.02 초

다음 중에서 공의 떨어진 시간과의 관계를 가장 잘 설명한 것은 다음 중에서 어느 것입니까?

(A) 질량이 클수록 빨리 떨어진다.
(B) 부피가 클수록 천천히 떨어진다.
(C) 질량과 부피가 모두 클수록 빨리 떨어진다.
(D) 질량이나 부피는 떨어지는 시간과 관계가 없다.